Leveraging Computer Vision and Humanoid Robots to Detect Autism in Toddlers

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Marie Denise Manner

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Maria L. Gini, Adviser

December 2018

ACKNOWLEDGEMENTS

I would like to thank my advisor, Maria Gini; without you to support my research, guide me in my struggles, and challenge me, I would not be who I am today. I would also like to thank my principal investigator, Jed Elison, for your immediate and wholehearted welcome of me and my robot; your support for my research is immeasurable. Thanks are also due to many wonderful colleagues and friends who have impacted and shaped my life, of which I will only name a small percentage. First, my labmates, from the original underground Keller TA lab to the bright and sunny underground Walter lab, with whom many late night discussions, powerpoints, and papers have flown – especially Ernesto, Mark, Julio, Emery, Nick, and Josh. Second, my adopted Elabmates, without whom my research would have been quite impossible to conduct or analyze, especially Nadja, Kirsten, Carolyn, and Layne. Third, my family – Mom, Gary, Carrie, Nate, Maria – and friends – especially John, Matilda, Austin, and David – who have watched and cheered, even if a bit bemusedly. You keep me grounded, sane, and happy.

A huge thank you goes to the families who participated – their belief in research as well as their time and energy makes it possible to study child development. Lastly, I want to thank the National Science Foundation's grant, "I/UCRC: Robots and Sensors for the Human Well-being" (NSF IIP-1439728), the MnDRIVE RSAM initiative at the University of Minnesota, and the Graduate Assistants in Areas of National Need (GAANN) from the Department Of Education.

DEDICATION

To my mother, Aileen.

ABSTRACT

Autism Spectrum Disorder is a developmental disorder often characterized by limited social skills, repetitive behaviors, obsessions, and/or routines. Early intervention significantly improves long-term outcomes for toddlers identified in the second year of life and is the best approach for affecting lasting positive change for children with an ASD. Research shows that children with autism especially enjoy technology, including autonomous (or seemingly autonomous) robots. Tying these together, we hypothesize that observing play interactions between very young children (2 - 4 years old) and a humanoid robot can help us identify children with autism; this first requires us to generate a very large, thoroughly characterized dataset of typically developing children. We begin with an eye tracking experiment comparing four different robots and a young human peer; this shows us which type of robot may be of most interest to children in an in-person, real-life play scenario, and if that robot is as interesting as a peer. Using the robot found to be most interesting in the eye tracking experiment, we next detail a human-robot interaction experiment that engages 2 - 4 year old children in a series of social games with a small humanoid robot; we then analyze the social distances, or proxemics, of the child throughout the interaction. To generate the proxemics data, we use a highly automated person detector which utilizes two state-of-the-art convolutional neural networks; with the proxemics and other development assessment data, we compare and group participants and discuss the implications of those results. A subset of robot interaction participants also finished the eye tracking task, so we discuss the relationship between the human-robot interactions and eye tracking results. Lastly, to validate the generalizability of our automated tracker, we test the system on two other child development experiments, a multiple-participant in-group bias play scenario for 5 and 8 year old children, and an unsolvable box task for toddlers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Autism spectrum disorder (ASD, or autism) is a neurodevelopmental disorder defined by behavioral symptoms that include communication and social behavior deficits as well as restricted and repetitive behavior patterns. ASD becomes more prevalent every year; the Centers for Disease Control and Prevention began recording trend data in 2000, when about one of 150 children had an ASD [76], until the most recent 2014 data that estimates roughly one of 59 children aged 8 years old in the United States has an ASD [7]. Males represent 3 in 4 individuals diagnosed with ASD. The typical age of diagnosis in children is around four years old, but symptoms appear earlier in life. Characterizations of autism, which vary in severity and presence from individual to individual, can include impairments in social interactions, language and communication skills, executive control, and eye contact. Symptoms may also include rigid, repetitive stereotypic behaviors such as rocking, spinning, head banging, or hand flapping, heightened sensitivity to some senses (such as to touch or smells), and insistence on sameness in routines or environments. While no child with autism behaves exactly like another, a mix of these symptoms starting at a very young age can indicate that a child may have a developmental disorder and should get tested.

Once a child is suspected of having autism, getting the resources required for proper diagnosis and treatment can be a challenge. The waiting period for getting an affected child together with an experienced clinician able to produce a diagnosis with the Autism

Diagnostic Observation Schedule (ADOS, [64]) commonly takes up to six months. The ADOS is a semi-structured, time-intensive assessment that includes tests of imaginative play, social cues, and communication. Once diagnosed, children with autism may receive intensive intervention before atypical patterns of behavior and brain function become firmly established.

Robotics research in autism is over a decade old, yet does not currently meet standards of psychology and child development researchers [28, 79, 99]. The goal of this research is to address some of these concerns (detailed in Section 2.8), and ultimately to make diagnosis of ASD or other developmental disorders more easily accessible. Accessibility in this case means a portable, reliable, accurate diagnostic tool to assist trained clinicians or other, possibly less highly trained physicians. It may simply mean making tools for more quickly distinguishing atypical behavior 'in the wild,' so that parents and physicians have earlier warning to assess children. In this research, our main tools are commercially available cameras and robots, existing open-source software, trained nueral networks, and new programs produced and freely released.

This thesis develops a baseline behavior analysis for typically developing children from ages two to four years old and develops and provides the tools necessary for describing this baseline cohort. These children span from very low to high risk of developing autism (based on risk factors such as having a sibling with autism), which allows us to further characterize social and autistic behaviors in the context of a highly interesting technology, a humanoid robot. Provided participants with autism, we would be able to place children with autism on a scale or cluster in relation to our large group of typically developing children. Given this placement, we could then say if any particular undiagnosed child that later participates in the study is in an area of high risk or low risk to be diagnosed with autism compared to their typically or atypically developing peers.

In essence, this thesis helps address the visible differences between typically developing children and children with autism, by using and developing state-of-the-art technology.

This research advances the application of computer vision and deep learning to autism diagnosis by identifying and categorizing children's reactions to playing with a robot. We do this by working with typically developing children and children at high risk for autism in eye tracking experiments and robot interaction experiments. Specifically, we find eye gaze preferences on robots, explore joint attention via orientation detection, and study and classify proxemics via person tracking during social interactions with robots.

This work was introduced in [70] and presented in [71], in which we detail the experiments and presented initial results of clustering children based on various developmental assessments and reactions to the robot. At the same time, our experiments address many issues brought up in Diehl's research into basic categories of robotics use in autism therapy [28], which will be detailed in the next chapter. Our results include software to semi-automate person tracking in overhead video footage, which results in raw output of proxemics, or social distance, data. With the proxemics data, we quantify children's comfort with interacting with the robot, which we also use to differentiate child behaviors. We also explored orientation estimation through two methods that had a classification accuracy of around 60% and 80% accuracy, respectively, on head direction on a small dataset; however, in practice the methods did not perform well enough to use in classification of looking behaviors.

1.1 Summary of contributions

This thesis contributes to several areas.

In computer science and engineering:

- A human-robot interaction protocol verified by child development psychologists, tested with 59 toddler-aged participants, that is suitable for use with boys or girls at least two years of age.

1.1. Summary of contributions

- The packaged, executable protocol deployable on any off-the-shelf NAO robot (with limbs), for NAO operating system $2.1.4$ and Choregraphe version $2.1.4$.
- TensorFlow neural network weights for detecting person and NAO-robot location from an overhead (e.g. surveillance) perspective.
- A validation of automated proxemics tracking over two different research experiments: one giving children an impossible-to-solve problem, and another testing if very young children demonstrate in-group bias from subtle cues like clothing color assignment.

In eye tracking literature in support of robotics:

- Evidence that children aged 18 to 45 months show strong gaze preferences for a humanoid robot over non-humanoid robots, especially for boys.
- Evidence that children aged 18 to 45 months show similar gaze preferences for a humanoid robot and a live human peer.
- A relationship analysis between eye tracking data and human-robot experiment interaction data, from children who participated in both experiments.

In child psychology, especially in phenotype characterizations that may relate to developmental disorders:

- A large number of typically developing participants ($N = 59$), ranging from low to high risk for autism spectrum disorder, characterized with 11 developmental assessments (besides the two experiments detailed in this work).
- Characterization of common movement patterns in child interactions with a humanoid robot.
- Evidence that a distance ratio of child to robot and caregiver, together with assessment scores, clusters clinically different groups of participants.

1.2 Organization

Chapter 2 contains a broad background in current theories in psychology as pertains to children with autism, beginning with the definition of autism and diagnosing autism in very young children. We then move to important skills in child development including gaze and joint attention, commonly studied with eye tracking technology, and also social distances, or proxemics. We then discuss existing work in behavior recognition and people tracking in developmental disorder research. Lastly, we give a brief background in state-of-the-art neural networks for object detection that we use first, for tracking people throughout human-robot interaction experiments, and second, for differentiating the orientation of people.

In Chapter 3 we discuss two experiments in eye tracking with toddlers to discover their preferences between different types of robots and a fellow social being. We conclude each eye tracking task with its results, and end by discussing the conclusions that can be drawn from both experiments. In Chapter 4 we detail an extensive human-robot interaction experiment performed with 59 children, aged 2 - 4 years. We discuss results of clustering and comparing children based solely on proxemics data, as well as proxemics data in combination with specific autism-adjacent scores of developmental assessments. We show work on differentiating child orientation, and end with lessons learned from working with young children and robots. In Chapter 5 we discuss a subset of children that participated in both the eye tracking experiments and the human-robot interaction experiments. We explore the relationship between experiment results from those children.

In Chapter 6 we apply tracking software and proxemics analysis to two other research projects. The first project examines how children respond to an unsolvable task; the second examines the age at which unacquainted children demonstrate implicit in-group out-group bias during free-play scenarios. Both projects can ultimately be helped with the addition of automated proxemics analysis as detailed in the previous chapters.

Lastly, Chapter 7 describes the conclusions, limitations, open issues, and future work in these thesis topics. We discuss the implications for future research in computer science and psychology of eye tracking for robot preference detection, differentiating autism phenotypes with automated proxemics detection, and automated proxemics analysis.

1.3 Terminology

The latest version of the Diagnostic and Statistical Manual of Mental Disorders, Edition 5, was published in 2013. Edition 5 changed the definition of autism to encompass several disorders under one term, Autism Spectrum Disorder, or Autism, or ASD. The current diagnostic criteria is given in Section 2.2, and the new DSM edition explains that: “Autism spectrum disorder encompasses disorders previously referred to as early infantile autism, childhood autism, Kanner’s autism, high-functioning autism, atypical autism, pervasive developmental disorder not otherwise specified, childhood disintegrative disorder, and Asperger’s disorder.” [6].

Because research prior to 2013 used earlier diagnostic labels, research participants diagnosed prior to 2013 may have had one of these subdiagnoses. Computer science and robotics researchers working with atypically developing children frequently did/do not list (and likely, do not have access to) the complete diagnostic descriptions of participants, simply that they have autism or one of its subdiagnoses. Because the exact subdiagnosis, if any, is unknown, this thesis will frequently refer to a participant of previous research as ‘child with an autism spectrum disorder,’ ‘child with an ASD,’ or ’child with autism,’ depending on the language used in the original research. Generally speaking, this thesis discusses two general populations: (1) typically developing (TD) children, who are in the majority of developmental milestones, drive the means and standard deviations of assessments, and provide baselines for the majority of the Western ideas and concepts discusses in this thesis, and (2) children with or at risk of autism, or (an) autism spectrum disorder, who are atypically developing. Some background work also references atypically developing children that do not have ASD, which will be explicitly stated.

Throughout the thesis, we use abbreviations for autism spectrum disorder (ASD), typically developing (TD), human-robot interaction (HRI), socially assistive robotics (SAR), convolutional neural network (CNN), and standard deviation (st. dev.).

Chapter 2

Background and Preliminaries

2.1 Problem overview

Socially Assistive Robotics (SAR) is a relatively recent area of robotics research, having started within the last two decades, and is aimed at helping populations with special needs. Socially assistive robotics spans research from robots as therapy tools for children with pervasive developmental disorders, to robots as tools for adults as companions or helpers. Robotics research with children with autism stems from the fact that afflicted children tend to especially enjoy autonomous (or seemingly autonomous) robots [26], and researchers have used a wide variety of robot appearances and abilities in this area [99]. While the reason for this high level of interest is unknown, researchers clearly have the potential to leverage robotics for autism diagnosis or treatment [97]. Diehl and colleagues [28] review a slew of research in four areas: responses to humans vs. responses to robots, using robots to elicit behaviors, using robots to teach or practice skills, and using robots for giving performance feedback. They concluded that while the research was promising, it was mostly exploratory, and the authors gave feedback for the computer science community at large using robots in autism therapy or diagnosis. Their criticisms included lack of robot integration to established treatments, lack of study participant follow-up, small sample sizes, little scrutiny on the actual therapeutic protocol, and little detailed characterization of participants [28]. Diehl's review recommends addressing these problems as well as focusing on what child characterizations indicate the most individual benefit from robot interaction.

We leverage autonomous robots, which are frequently of great interest to children with autism, to help further characterize an autism phenotype. We do this by collecting a thoroughly characterized dataset of typically developing children, with the future goal of describing a cohort with autism in the exact same way, to see which differences may be indicative of autism. The dataset includes quantitative data on how children interact with a social robot, those children's scores on assessments that gauge particular developmental skills and behaviors that are strong indicators of autism, how those scores and interaction data relate to each other. This research contributes both to computer science and psychology, and we now review relevant background in both fields: autism spectrum disorder, diagnosing autism, joint attention, eye tracking, proxemics, behavior tracking, human-robot interactions, and deep learning tools.

2.2 Autism spectrum disorder

Autism spectrum disorder, or ASD, or autism, is a developmental disorder characterized by deficits in social skills and repetitive and restricted behaviors. The Diagnostic and Statistical Manual of Mental Disorders, Edition 5, [6] establishes the following criteria for autism:

> A. Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history (examples are illustrative, not exhaustive; see text):
>
> 1. Deficits in social-emotional reciprocity, ranging, for example, from abnormal social approach and failure of normal back-and-forth conversation; to reduced sharing of interests, emotions, or affect; to failure to initiate or respond to social interactions.

2. Deficits in nonverbal communicative behaviors used for social interaction, ranging, for example, from poorly integrated verbal and nonverbal communication; to abnormalities in eye contact and body language or deficits in understanding and use of gestures; to a total lack of facial expressions and nonverbal communication.

3. Deficits in developing, maintaining, and understanding relationships, ranging, for example, from difficulties adjusting behavior to suit various social contexts; to difficulties in sharing imaginative play or in making friends; to absence of interest in peers.

Specify current severity: Severity is based on social communication impairments and restricted, repetitive patterns of behavior.

B. Restricted, repetitive patterns of behavior, interests, or activities, as manifested by at least two of the following, currently or by history (examples are illustrative, not exhaustive; see text):

1. Stereotyped or repetitive motor movements, use of objects, or speech (for example, simple motor stereotypes, lining up toys or flipping objects, echolalia, idiosyncratic phrases).

2. Insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior (e.g., extreme distress at small changes, difficulties with transitions, rigid thinking patterns, greeting rituals, need to take same route or eat same food every day).

3. Highly restricted, fixated interests that are abnormal in intensity or focus (e.g., strong attachment to or preoccupation with unusual objects, excessively circumscribed or perseverative interests).

4. Hyper- or hyporeactivity to sensory input or unusual interest in sensory

aspects of the environment (e.g. apparent indifference to pain / temperature, adverse response to specific sounds or textures, excessive smelling or touching of objects, visual fascination with lights or movement).

Specify current severity: Severity is based on social communication impairments and restricted, repetitive patterns of behavior.

C. Symptoms must be present in the early developmental period (but may not become fully manifest until social demands exceed limited capacities, or may be masked by learned strategies in later life).

D. Symptoms cause clinically significant impairment in social, occupational, or other important areas of current functioning.

E. These disturbances are not better explained by intellectual disability (intellectual developmental disorder) or global developmental delay. Intellectual disability and autism spectrum disorder frequently co-occur; to make comorbid diagnoses of autism spectrum disorder and intellectual disability, social communication should be below that expected for general developmental level.

Note: Individuals with a well-established DSM-IV diagnosis of autistic disorder, Asperger's disorder, or pervasive developmental disorder not otherwise specified should be given the diagnosis of autism spectrum disorder. Individuals who have marked deficits in social communication, but whose symptoms do not otherwise meet criteria for autism spectrum disorder, should be evaluated for social (pragmatic) communication disorder.

Specify if:

- With or without accompanying intellectual impairment
- With or without accompanying language impairment

- Associated with a known medical or genetic condition or environmental factor (Coding note: Use additional code to identify the associated medical or genetic condition.)
- Associated with another neurodevelopmental, mental, or behavioral disorder (Coding note: Use additional code[s] to identify the associated neurodevelopmental, mental, or behavioral disorder[s].)
- With catatonia (refer to the criteria for catatonia associated with another mental disorder) (Coding note: Use additional code 293.89 catatonia associated with autism spectrum disorder to indicate the presence of the comorbid catatonia.)

Note the DSM-5 guide requires specifying whether autistic behaviors occur with other disorders. Autism has no clear cause and no genetic test; tests for autism are more thoroughly discussed in Section 2.3. Some genetic disorders can appear like autism, but have a testable, distinct cause. For example, Fragile X Syndrome [38] is caused by changes in the fragile X mental retardation 1 (FRM1) gene. This gene makes a protein required for standard brain development, and people with the syndrome do not make the protein. The developmental delays, learning disabilities, and social and behavior problems produced by Fragile X Syndrome may look like autism, but they are distinct disorders, and not all people with autism have any clear genetic mutation.

Autism has lifelong repercussions; in 2011, the CDC estimated the total costs per year for children in the US in the $11.5 - 60.9$ billion dollar range [76], and other estimates agree, putting a lifetime cost of care for one individual in the $1.5 - 3$ million dollar range [41, 106]. Early intervention significantly improves long-term outcomes for toddlers identified in the second year of life [27] and is the best approach for affecting lasting positive change for children with autism. ASD results in significant family distress and financial burden, and early diagnosis and therapy is the best way to get children the help they need.

2.3 Early autism diagnosis

Diagnosing autism is a time intensive process performed by a highly trained clinician, conducted with an observation session with a child, caregiver interviews, or a mix of both. There are several diagnostic tools, with some variety by age, and popular tests go on to become revised and updated with time. For example, the Autism Diagnostic Observation Schedule (ADOS, [65]) was first published in 1989 and is still in use today, though it has been updated to the revised Autism Diagnostic Observation Schedule-Generic (ADOS-G, [64]). Other tests, for example the Modified Checklist for Autism in Toddlers (M-CHAT, [94]), Screening Tool for Autism in Two-Year-Olds (STAT, [108]), and Communication and Symbolic Behavior Scales (CSBS, [120]), have varying degrees of popularity and accuracy. There is no real consensus on which diagnostic tool is best, and researchers have started to track the over-all utility of different diagnostic tests, with important implications.

Øien and colleagues' recent longitudinal study found that in $69,668$ children assessed with the M-CHAT at 18 months old, $1,471$ screened positive, $68,197$ screened negative, and 228 of those negatives were later diagnosed with ASD (false negatives) [77]. The authors found clinical differences at 18 months old in the false negative group in social, communication, fine, and gross motor skills, so the test itself was missing clinically important markers. Falkmer and colleagues examined evidence-based tests reported from years 2000 to 2012 and found viable literature regarding 3 different observation tools, 6 parent or caregiver interview tools, 3 combination observation / interview tools, 2 screening tools with potential, and 4 Asperger syndrome-specific tools [31]. They found that the Autism Diagnostic Interview - Revised (ADI-R, [66] and ADOS had the most evidence and the highest sensitivity and specificity, and between the two have a clinical accuracy comparable to diagnosis from a multi-disciplinary team of experts. In 2018, Randall and colleagues reviewed six tests (all of which were reviewed earlier in Falkmer's literature review, with limited to substantial evidence of utility) to see which was the most effective at diagnosing as compared to a team of clinical experts [83]. They reviewed tests that interview caregivers – Autism Diagnostic Interview - Revised (ADI-R, [66]), Gilliam Autism Rating Scale (GARS, [43]), Diagnostic Interview for Social and Communication Disorder (DISCO, [121]), and Developmental, Dimensional, and Diagnostic Interview (3di, [105]) – tests that require a trained clinician to perform an observation – the ADOS-G – and a test that uses both a clinician observation and a caregiver interview, the Childhood Autism Rating Scale (CARS, [100]). They found that the ADOS is best for not producing false negatives, is similar to CARS and ADI-R in not producing false positives, but like the other tests is likely to over-diagnose autism in populations with a lower prevalence of the disorder.

Regardless of the tool used, the child is assessed for various interaction responses including imaginative play, social cues, and communication. The ADOS, for example, has four modules, each of which allows for people of varying levels of expressive language and behavior to participate in the test. Each module gives participants a chance to respond to a ‘press.’ The response of the child to each press in the ADOS gives the clinician opportunities to see how the child reacts to social cues. Reactions like eye contact, responding to a name-call, reacting to a multi-layered request, or verbally responding with a complex sentence are all important. A sample cue is pointing and asking the child to follow the clinician’s gaze and point. A typically developing child will respond to a gaze and point by looking in the direction of the gaze and point; a consistent lack or delay of gaze following or point following is often a symptom of autism. This bid for gaze- or pointing-following and subsequent focus by both parties is called joint attention, and this incredibly important social skill is discussed in the next section.

While a slew of tests to diagnose autism exist, there is still the question of the earliest age at which autism can be detected, prompting research into just how early symptoms appear. Zwaigenbaum and colleagues discuss a host of symptoms of high risk infants they followed longitudinally, some of which were later diagnosed with an ASD [126]. They found that infants later diagnosed with autism had visible atypicalities from controls by 12 months of age. The researchers made an observation scale, a computerized visual orienting task, and standardized measures of temperament, cognitive, and language development. Among other symptoms, they found atypical responses in eye contact, disengagement and delay of visual attention, orienting to name, imitation, social interests, early temperament, and language delays. The authors did not directly relate behavioral findings with physiological brain development, though they posited a few brain regions that may interact with particular developments. Zwaigenbaum's research recruited participants who were already at high-risk for autism, but some children will have no known risk factors; the child's caregivers must suspect that the depth of delay or atypicalities in their child warrants a doctor visit.

Fortunately, research shows that many parents do realize something is atypical in their child's development well before an official test is conducted. A study by Young, Brewer, and Pattison [125] asked parents when they first noticed 33 signs of autism in their children, when the child was tested, and when the child was diagnosed. This study found the mean age of abnormal developmental signs of 15.1 months (SD 11.2 months), with 95% of parents noticing social development anomalies when the child was less than two years old and many noticing problems before the child's first birthday. By contrast, the mean age of seeking professional advice was 26.8 months (SD 10.4 months), and the mean age of autistic disorder diagnosis was 41.82 months (SD 13.6 months). While there were limitations in the study, notably that parents were being asked to recall past events and therefore may have accidentally biased memories given that they know now what autism symptoms include, the study findings still match with other results and provide evidence for autism symptoms prior to two years of age.

While there is no blood test to take or even a genetic marker for autism susceptibility, researchers have been able to find some measurable biological differences between typically developed children and children with autism. In 1998, Szatmari, Jones, Zwaigenbaum, and Maclean reviewed research on autism recurrence rates in twins and families, and the evidence for genetics factors was compelling, though the particular genetics responsible were unknown at the time [109]. A decade later, with the advent of robust gene sequencing techniques and research, Losh, Sullivan, Trembath, and Piven reviewed studies searching for genetic links in autism [67]. These studies attempt to map chromosome abnormalities in people with autism compared to controls. The literature included research in genome-wide linkage analysis studies (which look at candidate regions of chromosomes), genome-wide association studies (which look at markers throughout the chromosome rather than at target locations), six specific genes that seemed most plausible to the authors based on replication studies, molecules that play important roles in synaptic development and function, and DNA structural variants such as microdeletions, duplications, and copy number variations. Overall, the research supports the idea that some particular rare mutations are responsible for some cases of autism and others are more likely based on a variety of subsets of genetic variants. The authors also match common autistic phenotypes with implicated chromosomes; for example, the most promising links between genetics and behaviors is the language deficit phenotype and variants on 20 chromosomes. They propose that future research focus on crafting these endophenotypes, or links between smaller gene constellations and subclinical disease markers. At this time, unfortunately, genetic mapping is still not systematic or predictive enough to be able to test for autism.

Research has also searched for more easily seen differences between children with autism and controls; for example, Elison and colleagues found that particular neural brain circuits in six month olds predicted response to joint attention in nine and ten month old children [29]. The researchers took diffusion weight images of sleeping infants, mapping three fiber tracts on both sides of the brain, including optic nerves, for a total of six regions of interest. Increased microstructural organization in one of the regions in the younger children was associated with better performance on joint attention tasks in older children. This may mean a symptom of autism, poor performance on joint attention tasks, can be found by mapping a fiber tract in the brain of a six month old child, giving us a much more identifiable, early biological marker of autism. However, MRIs and similar large, expensive tools cannot be deployed on a large scale for systematic autism detection.

We have two problems, then, in current autism diagnosis; the first is the variety of exams, of which the best performing still need highly trained clinicians to perform; those experts must devote a large block of time to the interaction, hand-code the results, and ultimately use their expert judgement. The second is the huge delay between noticing, seeking help, and receiving a diagnosis for children with autism, which can delay therapy by years. We therefore suggest that finding any additional warning signs, especially by using automated tools that are reproducible, portable, and evidence based, can take some diagnostic burden from parents or clinicians and help reduce the time needed to detect and diagnose autism. Ideally, automated diagnostic tools would be deployed in clinic waiting rooms; but before such a step, we need to provide metrics from quantitative data on typically developing children to see what a baseline of typically developing children looks like and what such a diagnostic tool might use to compare children viewed in the wild. Therefore, this thesis targets psychology theories that can be captured easily, automatically, and without adding much extra burden of data collection. Two such psychology topics are joint attention and social distances, or proxemics, and we next discuss what these are and what tools are available to automatically track them.

2.4 Gaze and joint attention

Joint attention is the shared focus between people on some object; typically developing infants show this ability by 12 months, and the time infants spend in joint engagement with their mother and how much the mother follows the infant's focus predicts early gestural and linguistic communication skills [11]. The lack of ability to hold joint attention can be an indicator of autism [53], and is part of the test for autism for young children. A bid for joint attention can be a simple conversational cue ('do you see that cat over there in the window?') or a survival tool ('do you see that lion in the tree?'). Joint attention can be initiated by verbal or physical cues varying in subtlety, from looking, to pointing, to speaking, or a combination thereof. A successful bid for joint attention results in the recipient of the request also looking at or paying attention to the object in question. Developing joint attention is an important milestone in an infant's first year, and Charman and colleagues found that joint attention and the ability to switch gaze in 20 month olds was associated with less severe difficulties in communication or language and social difficulties in 42 month olds [15].

In eye contact, a more direct social skill than joint attention, children with and without autism differ in some contexts but not others. Jones et al. investigated eye contact with an unknown adult (the examiner) during free play sessions and during conversations with no toys present between the experimenter and children with and without an ASD [54]. They found that both controls and children on the spectrum had more frequent and more lengthy eye contact with the examiner during conversations rather than free play, with no interaction between diagnosis and context. There was an effect of ASD severity on social affect to gaze duration, but not frequency of eye contact or on restricted and repetitive behaviors. They also noted that the toy in use during the scenario impacted the amount of eye contact, for example a board game versus a wind-up toy. From this research we know that context matters, then, and the authors encourage research to engage children in eye contact using different approaches, such as conversing with and without toys, and different toys.

A sensible line of research, then, is to use objects of special interest as a focus of joint attention. Work from Kryzak and Jones did just that, using circumscribed interests as attention getters while prompting the children to initiate joint attention in a small intervention study to see if the items helped participants learn to initiate joint attention [60]. A circumscribed interest is a subject in which a person with an ASD becomes intensely interested in and may study to the point of obsession or deep knowledge, such as cars or trains or particular games. Two of Kryzak and Jones' subjects learned to initiate joint attention with the use of these circumscribed interests as tools in the intervention, though the intervention with the third child had to be done with other preferred items. While not distinctly a circumscribed interest, using technology in general to orient children with autism has had support since early research in the 1970s, when Colby showed technology was a useful tool in teaching children with autism [18]. Research in the decades since has shown support for using technology to attract interest of children with ASD, even in infancy. Klin, Lin, Gorrindo, Ramsay, and Jones first showed in 2009 that two-year-old children with autism did not prefer to gaze at biological motion, as produced by point-light displays of humans playing games versus non-biological motion of the same displays but inverted, whereas typically developing controls and developmentally delayed controls did prefer the biological motion [59]. Three years later, Annaz and colleagues showed distinct preferences in 3–7 year olds with ASD for gazing at non-biological motion (a spinning top) over biological motion (a walking human), with no preference between non-biological motion or scrambled motion [4]. Typically developing controls preferred biological motion in both cases.

Technology, even if not a circumscribed interest of a child, could prove to be an interesting and motivating object of joint attention. Dautenhahn and Werry investigated interactive environments with older children (8 – 12 years old) specifically in trials with a mobile, autonomous robot, and found children with autism to have high levels of interest in robots [26]; other literature reviews supporting use of robots with children with autism concur [98, 47, 80]. Anzalone et al. used a humanoid robot to initiate joint attention to pictures of a dog or a cat on either side of a participant (a typically developing child control, or child with an ASD), and found that while the robot was able to initiate joint attention, compared to joint attention initiated by a therapist, the robot initiation resulted in significantly lower joint attention scores in both groups. These children, however, only had one three minute trial with the robot, and the average group ages differed. The group with ASD's (N = 16) was 9.25 years (st. dev. 1.87 years), but lower development age mean, at 7.47 years (st. dev. 2.9 years); the control group (N = 16) was 8.06 years (st. dev. 2.49 years). Their findings show that joint attention interaction depends on the partner, and the number of cues used (gazing, pointing, vocalizing, or a combination). Warren and colleagues compared groups of TD children and children with an ASD (N = 6), using a human and a robot to prompt for joint attention, over a period of four interactions [118]. They found that children with an ASD focused more on the robot than the human prompter, and directed their gaze to the correct item at the rate of the TD children. We understand from this research that context matters, and familiarity or number of interactions matters.

We have shown how variations in gaze and joint attention can reflect autism symptoms. Hardware and software tools, specifically eye tracking tools, have made gaze tracking much more computationally and financially feasible in large scale studies over the last two decades. Researchers have been able to look at other gaze patters, not simply joint attention, with computerized eye tracking. We now discuss what eye tracking research has contributed to the differences between typically and atypically developing people, especially those with autism.

2.5 Eye tracking

Eye tracking became popular after a 2002 review from Klin and colleagues that suggested eye tracking technology as a novel tool for looking at the gaze pattern differences of individuals with ASD or typically developing (TD) individuals [57]. That work examined the gaze differences as the individuals looked at movie clips containing social situations, and detailed how the TD viewer looked at the actor's faces whereas the ASD individual had a different gaze pattern. While the TD individual looked back and forth at the actors while they spoke, the individual with ASD's gaze was on non-essential parts of the social interaction and made different paths on the screen. Klin went on in [58] to show individuals with autism show abnormal gaze patterns, and some gaze patterns are strong predictors of social competence in ASD individuals.

Eye tracking studies are now commonly used on infants because they require no explicit communication or understanding of the task – anyone with sight, from a three month old to an adult, can look at a screen at the things they find interesting. Karatekin reviews at length common eye tracking measures and how each is used to study various clinical or typical populations [55], restricting the literature review to participants of four or five years of age and up. His review shows that eye tracking has been used for measuring visual-spatial attention, engagement and disengagement of that attention on targets, breadth of attention, and processing speed and ability. Eye tracking has been used as a way to track differences between typical and atypical development in disorders like Attention-Deficit/Hyperactivity Disorder (ADHD), Pervasive Developmental Disorders (PDDs), schizophrenia, and Psychotic Disorder Not Otherwise Specified to find difference in gaze patterns.

Constantino and colleagues used eye tracking studies to find significant genetic effects in 18–24 month old monozygotic twins, dizygotic twins, and paired sibling infants [21]. They found significant evidence for genetic factors in looking at eyes and mouth – 0.91 correlation for eye looking, 0.86 correlation for mouth looking in monozygotic twins, 0.35 correlation for eye looking, 0.44 correlation for mouth looking in dizygotic twins, and no significant correlation for siblings or random non-related pairs of infants at the time of the study, as well as high correlations at a follow-up 15 months later. They also found higher (but clinically non-significant) probability of monozygotic twins moving their eyes at the same time.

Overall, eye tracking is a useful tool in detecting differences in typical and atypical populations, though there are some specific concerns when applying eye tracking methods to children with autism. Sasson and Elison discuss best practices in eye tracking this population, including guidance on the eye tracking system, testing environments, procedures, analysis, and representing results in [96]. The added effort to generate accurate results is certainly worthwhile; Xu and colleagues showed that human gaze can actually be predicted based on the contents of an image when that image is broken down in terms of pixel, object, and semantic level attributes [123]. Object level attributes include things like size and solidity, whereas semantic level attributes are more abstract and include things like how much a thing directly relates to humans (like faces or touch), implied motion, and how much a thing relates to non-visual human senses (like taste or smell). Wang and colleagues demonstrated atypical visual saliency in people with an ASD versus typically developing people [117]. They showed that people with an ASD had a strong bias towards gazing at the center of an image, rather than at the semantically important parts of an image.

As referenced in Section 2.4, early eye tracking work found differences in gaze patterns in typically developing individuals versus those with autism [57, 58]. Shic, Jones, Klin, and Scassellati used a subset of Klin's 2002 eye tracking data and determined typically developed (control) individuals have a particular, learnable model of gaze as applied to viewing social scenes, and that those models transfer to other controls [102]. The strategies of individuals with ASD, however, did not transfer well to controls or to others with ASD, implying that people with ASD have ordered gaze patterns that are likely unique cascading specializations in attention, and these patterns replace the more standard gazing patterns of typically developed people.

Guillona, Hadjikhanib, Baduela, and Rogé reviewed over a decade of eye tracking literature around spontaneous gaze behaviors to see how much evidence exists for two important assumptions of the past: people with ASD have decreased social orienting and attention to relevant stimuli like faces compared to TD individuals, and people with ASD look at mouths more and eyes less than TD individuals [48]. Their review of 25 published studies on the first topic suggest people with ASD do vary in their attention to social orienting, but the differences do not generalize, and the (lack of) ability varies by context, and may be driven partly by typically developing individuals' quicker or extra attention to specific parts of faces during context dependent (e.g. game playing) or multi-person social scenes. Instead, ASD impacted individuals' orienting to a face in a social scene may be hindered, not lacking, e.g. by differences in ability to ignore competing objects in complex situations such as static versus dynamic content, variable number of faces or people, or mixed number of people and objects in a scene. The authors' review of 31 published studies on the second topic found no solid support for excess mouth and reduced eye gaze, and suggested again that typically developing participants drive results in specific ways depending on events, and stimulus-specific reactions might belong to a subgroup of people on the spectrum.

In short, automatic gaze detection has contributed a wealth of knowledge around how people respond to visual cues. The variety of stimuli, result metrics, and participant developmental age all matter, but eye tracking studies continue to be an important tool in researching developmental disorders. We now move to a more physical dimension of human interactions – social distances, or proxemics – in typically and atypically developing people.

2.6 Proxemics

In 1966, Edward Hall published some of the earliest work on physical and psychological distances between individuals, especially as they differed in different cultures; this is called proxemics, or the study of social and personal space and people's perceptions of it [49]. Hall showed that the distance between people and objects and people and other people varies by comfort level and society, and even by the sense which people use to judge the appropriate space (such as smell or touch). In 1973, Evans and Howard reviewed personal space research up to that time and concluded that personal space is influenced by sex (e.g. male / female pairs use less space than two females, who use less than two males), age (with standard personal space starting around 12 years) ethnicity (e.g. North Americans and Northern Europeans want more space than Mediterraneans), and familiarity (friendly people, friends, or people wanting to look positive use less space) [30]. Here we touch on more recent studies that show proxemics vary based on individuals' development, and while most proxemics research investigates person to person space, there is also research around person to object and person to robot proxemics.

In children, distances influence social ties to peers, especially distances to caregivers.

Legendre and Munchenbach studied 2-3 year old children and found that the presence of a caregiver moderates children's overtures to their peers, and further distance from an adult increases the chance of a positive interaction between peers during free play [62]. Gessaroli, Santelli, di Pellegrino, and Frassinetti investigated the personal space differences between TD and children with ASD both before and after spending time reading a well-liked book with an experimenter [42]. They found that children with ASD chose greater distances than TD children (by over 4 feet) at the beginning and end of the interaction when testing personal space; they also found that after reading a book together, TD children were comfortable with a smaller personal space (reduced by nearly a foot) while children with ASD were not. There was also an effect of where the participant started and who moved towards who during the personal space test: if the participant with ASD moved away from the experimenter before choosing when to stop, distances were greater than if the participant with ASD moved toward the experimenter before choosing when to stop. Participants with ASD did not show a difference in distance when the experimenter was the one moving towards or away from the participants; TD participants had no significant distance differences in either condition.

Asada and colleagues compared distances between typically developed people and people with an ASD [5] against another person and against an object. By asking people to stop when they felt comfortable when walking towards a male experimenter or an inanimate object (a coat rack), they found that individuals with an ASD required less personal space in both cases than typically developing individuals. In the same study, TD and individuals with an ASD preferred greater distances when being approached by an individual making eye contact than when being approached by an individual that was not. Mead and colleagues automatically annotated dyadic (two person) interactions, using proxemic features of two people with a non-responsive robot as a social cue [73]. Their work used individual, physical, and psychophysical features (e.g. torso pose, distance, and 'social' distance) to recognize behaviors. They found that training Hidden Markov Models on these features allowed them to recognize transitions in to (initiating) and out of (terminating) social interactions.

Mumm and Mutlu investigated different models of interpersonal distance using physical and psychological distance measures between adult participants and a robot, manipulating both the likeability / unlikeability of the robot and matching / avoiding eye contact [75]. They found evidence for effects of gaze behavior on physical distance (mutual gaze resulted in increased physical distance), likeability and gaze on physical distance (dislikeable robot with mutual gaze resulted in increased physical distance), gender on physical distance (males physically distanced themselves more) gaze and gender (males physically distanced themselves more when robot's gaze was increased), pet ownership (pet owners physically distanced themselves more), and whether the participant was in front of or behind the robot (participants got closer to the robot when they were behind it). The authors found a marginal effect for psychological distancing, in that participants disclosed more answers to personal questions to a likeable versus unlikeable robot, but the effect was not significant.

Working with children's proxemics and robotics, Feil-Seifer and Matarić looked at how to automatically determine if an interaction between a child and a robot was negative or positive, using, among other things, proxemics [35, 36]. Distance was used to classify locations such as being near the parent (within 1.25 meters), near the wall (within 0.3 meters), or behind the robot (any distance greater than $135°$ or less than $-135°$ to the robot's front). The goal of that work was to develop online systems capable of helping a robot autonomously alter its behavior given knowledge during free play about the child's avoiding or seeking behaviors.

Up to this point, we have discussed the psychology background relevant for this thesis in autism spectrum disorder, autism diagnosis, gaze, joint attention, eye tracking, and proxemics. We are now ready to discuss how state-of-the-art technology enables us not just to automatically find gaze patterns and social distances, but to identify patterns of physical motion. We now move to this difficult problem of detecting and classifying behaviors.

2.7 Behavior tracking

Because routine and repetitive behaviors are frequent symptoms of autism, research has tried to build tools to find these patterns of behavior. An ideal tool would recognize and classify classical symptoms of ASD, such as identifying a child with a movement tic and classifying the repeated motion as hand-flapping. Fulceri and colleagues showed that locomotion and grasping were visibly impaired in pre-schoolers with ASD [40], so it behooves researchers to automate the process and make detection methods more robust and able to find more complex behaviors. Hashemi et al. looked at classifying behavioral, or movement, symptoms of ASD including arm asymmetry, visual tracking, and attention disengagement [50, 51]. They studied clinical recordings of a clinician testing toddlers with the ADOS-T test (ADOS, toddler module, [68]) and the Mullen Scales of Early Learning [74], another standardized test. Their system had good agreement with expert ratings in gauging a disengagement of attention task based on detecting head turning, and successfully identified arm asymmetry in video clips of participants.

Rehg and colleagues introduce a database of video and audio recordings of a very structured, short interaction in a laboratory between child and experimenter to better enable the community to develop automated tools; they were able to use existing audio software to detect the initiation of phases of the interactions (with imperfect but reasonable performance, averaging 85% with a lowest performance of 57.14%), created a smile detector with balanced accuracy of 76.6%, and created a two-stage gaze detector that first looked to see if the child was facing the experimenter from a ceiling-mounted depth camera, and then at the pitch of the child's head from a child-facing camera [84]. They also created a system to predict if the child was engaged or not based on visual cues of the child touching particular objects, and audio cues of how much verbal communication was needed from the experimenter throughout the interaction; the system had good predictive power for a child engaged in playing with a ball (92.86%) and a child engaged in reading a book (73.33%).

2.7. Behavior tracking

A recent span of work tackles the very hard problem of identifying different children and their behaviors in a classroom setting. Sivalingam and colleagues instrumented a classroom with multiple cameras and depth sensors [104] to track children over sessions [116] and identified social groups and other risk markers [34]. Fasching, Walczak, Morellas, and Papanikolopoulos successfully classified repetitive body movements like hand flapping, shrugging, and ear-covering from video footage in school-age children [33]. Fasching and colleagues induced and labeled behaviors in children with obsessive compulsive disorder in a laboratory setup with a hand-washing scenario in a fake bathroom, and object organization on different carpet patterns [8]. They found that moving and ordering objects and longer hand washing were associated with obsessive-compulsive scale ratings. In related work, the authors then made tools to track the number of times participants applied and rinsed soap, and turned water on and off in the fake bathroom [32].

Note that much of this work is done in laboratory settings, whether the play is structured or unstructured, and ideally with as minimal invasiveness as possible, such that children undergoing assessment do not need to be instrumented with sensors. Even a very robust algorithm may not perform well on raw video data, so Rajagopalan, Dhall, and Goecke created such a database for testing against. They sorted public videos of children with ASD into common stereotypies (arm flapping, head banging and spinning) to give researchers difficult videos of children 'in the wild,' or in their home setting, for algorithm testing [82]. Generating free-play data in an easily accessible format, such as raw video, allows other researchers to reproduce results and test their algorithms on noisier, more unstructured data.

This thesis joins an important body of work that tries to robustly and automatically detect, track, and categorize motions and movements. At the same time, this research bears in mind the necessity of keeping children entertained and engaged in an interaction, generating different behaviors, and keeping the environment set-up both non-cumbersome and non-invasive. We now move on to the arguably most important part of this background work, human-robot interactions.

2.8 Human-robot interaction

Human-robot interaction (HRI) research encompasses any research on how humans and robots interact with one another, regardless of the physical (or digital) body of the robot, whether the robot is teaching or learning, whether the human(s) interacting with it are typically or atypically developing, and regardless of the human's age. HRI research includes helping clinicians develop or use tools, typically developing children learn skills, and robots as assistants, teachers, playmates, or companions. The latter type of research is also called Socially Assistive Robotics (SAR); SAR robots may be humanoid or not, depending on the goal of the research. We now discuss *why* robots, especially humanoid robots, have such a strong presence in diagnosis and therapy for developmental disorders.

The argument for using a humanoid robot in autism therapy comes from the idea that skills learned while interacting with a humanoid robot – which has parts analogous to a human such as limbs, a torso, a head, and sometimes even an expressive face – will transfer to interactions with other humans. Research has somewhat borne out this hope. Ricks and Colton reviewed trends in work with children with autism and robots, specifically robot-assisted therapy [88]. They reviewed a range of research examining if human-like or non-human-like robots are the best for use with children with autism, specifically in the areas of diagnosis, self-initiated interactions, turn-taking activities, imitation, emotion recognition, joint attention, and triadic interactions. Self-initiated interactions and turn-taking activities did not require humanoid robots, because the research focused on rewarding behaviors with desirable reactions and encouraging children to get used to waiting for responses before they do or say something. Such research encouraged children to communicate to get desired results and to develop the back-and-forth pattern of social interactions. Imitation, emotion recognition, and joint attention work used robots to demonstrate actions, from gross motor skills like raising arms to finer motor skills like smiling, to try to determine if a robot could teach imitation as well as a human. Robots participating in joint attention became excited when the child responded appropriately, looking at the object of the robot's gaze. These areas contrast with triadic interactions, in which the robot itself is an object of joint attention.

Ultimately, Ricks and Colton concluded that the use of humanoid versus non-humanoid robot depended on the goal of the research; very human-like robots should probably be used when the goal is to learn or simulate human-human interactions, and otherwise simpler humanoids may work best. Robins, Dautenhahn, and Dubowski compared four robot types, all of humanoid shape: (1) a pretty or (2) plain small humanoid doll, and (3) a plain and masked or (4) regularly dressed human pretending to be a robot [93]. They found that children preferred the plainer looking robots in each size case, which influenced their later studies and tells us that the simplicity in appearance of robots may be part of their attraction.

Research has looked into using robots as social mediators and has shown that skillful use of robots may facilitate human contact as well as increase engagement, attention, and social behaviors, and even reduce stereotypies. Research in this vein includes sharing interest (and joint attention) with adult TD humans in [91], playing video games and increasing cooperation with adult TD humans after playing with robots in [114], triadic game playing between children with autism in [115], and imitation games that segued into the robot as an object of joint attention and entertainment in [89]. In Robins and Dautenhahn's research, a human experimenter controlled a robot during a play session with a child with autism; once the child realized the adult was controlling the robot, the child interacted with the researcher, even smiling and giggling when the researcher 'messed up' the action the robot was supposed to do [89]. Shamsuddin and colleagues found that children show a reduced number of stereotypies and social deficits during interactions with a robot [101].

Research is also starting to explore how to best interact with children with different sensory preferences. Robins and Dautenhahn designed a robot to allow children with autism to play games while exploring the tactile sense (touch), and their findings suggested the robot game incidentally also encouraged understanding facial expressions and verbalizing [90]. Chevalier et. al split participants into groups based on differences in how much they relied on proprioceptive and visual cues to play an imitation game with a robot, finding that differing reliance levels impacted their ability to learn to imitate [16]. Perca and colleagues designed a game to find out how TD and children with ASD viewed social robots, asking them to sort six different robots (of anthropomorphic, zoomorphic, and caricatured types) into bins of humans, animals, machines and toys in [78]. In their study, both TD and participants with ASD associated the NAO robot (the robot used in our experiments, see Figure 3.1d) with a toy, but children with ASD also associated it with machines.

Some research is concerned with duplicating human patterns of interaction in order to teach robots to do the same – helping, in future, robots interact with and potentially teach skills to children or adults with an ASD better than they currently can. Xu, Zhang, and Yu looked at how changing a robot's joint attention and face looking patterns altered the way a TD human taught a robot about new objects [124]. The researchers tested three conditions that changed the amount of eye contact between a robot and a human trying to teach the robot about new objects – the robot simply followed the human's gaze, followed the human's gaze and performed extra face looking, or also initiated face looking if the human had not in the last three seconds. In the latter two conditions, the real-time responsiveness of the robot generated smoother speech-and-looking patterns similar to human teaching and learning.

Socially assistive robotics research spans more than tools for autism and developmental disorders. Departing from a humanoid appearance for their robot, Short and colleagues found that children remained engaged with a dragon robot that taught nutrition and healthy food choices over a six session, three week study with first-grade (six or so years old) children [103]. Children had a positive reaction, increased engagement over time, used slightly more complex speech over time, made healthier choices over time if not how to make healthier as well as faster choices, and learned slightly healthier (but not statistically significant) choices over time. The researchers also found that children with higher affect and self-regulatory ability were not able to interact with the robot significantly better than their peers, though that was one hypothesis of the work. Together, this shows support for using the robot as an interesting tool that sustains engagement with children. Though this example short-term study did not show much learning gains for the children, it provides support for the feasibility of socially assistive robotics.

Some work is looking at making using robots easier for clinicians to design or use. Villano and colleagues designed a wizard-of-Oz user interface for clinicians to control a NAO robot in a game of Simon Says (so called because the robot appears autonomous but is controlled behind the scenes by a human) [111]. Kim, Barakova, and Lourens described a prototyping framework to allow therapists to design interactions guided by an established treatment framework called Pivotal Response Training [56]. This type of research allows clinicians to rapidly design and implement therapeutic protocols with their patients, which makes robotic tools much easier to use in practice for researchers and clinicians, which can in turn guide the development of new robots and tools.

However, research also shows caution is warranted in mixing robots into therapy without careful consideration, and not every study agrees that robots should be used as tools in joint attention tasks. Robins, Dautenhahn, and Dubowski warn that using robots as a learning tool may generate isolation or encourage stereotypic behaviors if not used carefully, as when the researchers are trying to teach children to apply skills learned with robot interactions to interactions with other humans [92]. In another conflicting piece of research, Chiminade and Okka looked at the response times between using a robot or a human on a screen to cue the appearance of a target [13]. They found that participants took longer to orient to the target when a robot giving the cue rather than a human, suggesting that the robot was more distracting to use as a joint attention cue. This also warns us that novel stimuli may be more distracting than useful, and we must take this into account in any interactions with new robots.

Overall, however, when it comes to robots for use in therapy, there is still a lot of promise and evidence for robotic utility. In a 2016 literature review, Pennisi and colleagues reviewed 28 studies to empirically determine if social robots be a useful tool in autism therapy [80].

They found some generalizable observations, namely that participants with ASD performed better in robot versus human conditions, sometimes participants with ASD had behaviors with robots which are seen in typically developing people towards humans, and with robots in therapy, more stimulation was better than less. In hopes of showing similarly promising robotics research earlier in child development, at the earliest age at which a child may be diagnosed with ASD (2 years old), this thesis provides thorough characterization on a large number of varied but typically developing toddlers. With this broad baseline data, we hope to enable future researchers to understand behavioral development in robot interactions. Before we move to the experiments and results of this thesis research, we must discuss the tools used in processing video data.

2.9 Deep learning

This thesis requires offline video analysis, for which we use state-of-the-art deep learning methods. As will be discussed, we track people and robots over time from overhead video footage; for finding these objects automatically, we employ Faster R-CNN from Ren, He, Girshick, and Sun [85]. Every deep learning tool used in this thesis is implemented in the Python programming language and TensorFlow, an incredibly robust tool from researchers at Google [2]. We briefly discuss here the background neural network concept, specifically convolution neural networks, necessary to understand Faster R-CNN.

Deep learning uses various formulations of neural networks. Neural networks are based loosely on the concept of a human brain, in which a vast network (thousands or millions) of interconnected nodes take input and any particular node fires if its input exceeds some threshold. Neural networks can take as input any vector or matrix of numbers, or anything that can be cast in numeric form. Networks can be feed forward, in which input passes through different layers of nodes and eventually produces an output, or recurrent, in which nodes also pass information backwards into previous layers of nodes. Neural networks are given many examples of data, and learn over time how to differentiate the data. In networks used to identify images, for example, one might give the network a thousand picture examples of cats, a thousand picture examples of dogs, as well as the label of cat or dog. Eventually the network may learn to classify an image as a cat or a dog with the different features it has learned, such as that cats generally have rounder faces, smaller upright ears, and longer tails.

Convolutional neural networks, or CNN's, are networks specifically designed to take images as input, where an image is a rectangular matrix of pixels; the output could be simple, like a single number to represent categorization, or complex, like a new image. A convolution is a mathematical operation on two functions to produce a new function. In terms of image processing, this consists of taking a tiny region of an image, such as a 3x3 square of pixels, applying a function to it, such as multiplying by another small matrix (called a kernel), and storing the result in the corresponding location in an output matrix (another image). By sliding over the entire image by small increments and storing the results in a new output, we can produce a new image that has been changed in some way, such as smoothed or sharpened. In mathematical applications, the kernel is known; however, when training a convolutional neural network, the kernels themselves are what the network learns. These kernels are also called filters, and serve to detect different features in an image.

A convolution neural network includes an input (an image), several hidden layers of nodes, and an output. Layers commonly include convolutions, activations, pooling, and fully connected layers. We already discussed convolution, or applying some kernel to small areas of a matrix until the entire matrix has been convolved. An activation layer applies some non-linear function to the input; a very common choice is the Rectified Linear Units, or ReLU, which is a thresholding function. This layer adds a major non-linear component to a CNN, but ReLU is one of several viable options. A pooling layer takes the output of several nodes in the previous layer and combines them somehow into a single number; this effectly downsamples the input. Max pooling, in which case the single output of a node is

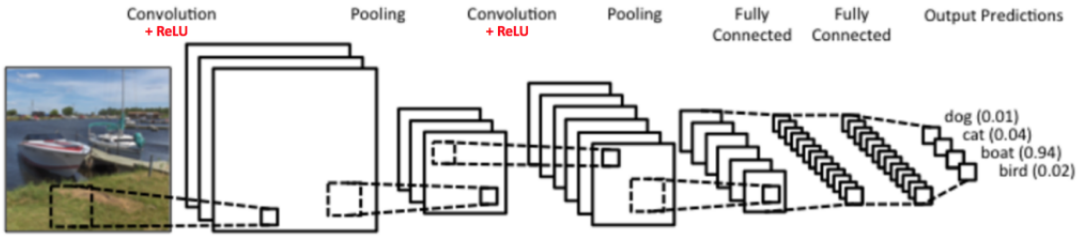


Figure 2.1: Sample convolutional neural network.

the maximum value of several nodes that feed into it, is a common choice. A fully connected layer is one in which every node in the layer takes as input the output of every single node in the previous layer. Some CNN's will also include drop-out layers, which randomly trim links between nodes, to help prevent over-training and make the trained network more robust. A very common CNN format is given in Eq. (2.1), where N, M, and K are some constants.

$$INPUT \rightarrow [[CONV \rightarrow RELU]{*}N \rightarrow POOL]{*}M \rightarrow [FC \rightarrow RELU]{*}K \rightarrow F \quad (2.1)$$

Region-based Convolutional Neural Networks, or R-CNN's, take in areas of interest, or regions, of a larger image, and classify the contents according to labels the network has learned. The success of an R-CNN depends largely on what method is used to find the regions that are submitted for classification. Very good region detectors, tied with a very good object classifier, will produce a network that is able to find objects in an image, classify them, and return the locations of the objects in the image, the labels of those objects, and the confidence the network has in each label.

The Fast Region-based Convolutional Neural Network method (Fast R-CNN) was introduced in 2015 by Ross Girshick for fast object detection in images [44]. Faster R-CNN combines Fast R-CNN with a new region proposal network, which is a fully convolutional network that predicts object bounds and objectness scores [85]. The new region proposal network, or RPN, is trained to find high quality regions, then uses Fast-RCNN for object detection. The RPN shares convolutional features with a detector network, making region proposals extremely cheap. This enables very high quality object predictions in an image at near real-time speeds. We will use the Faster-RCNN method to train a network to find and label our particular items of interest, a NAO robot and human beings, later in Section 4.4, and then just humans in Chapter 6.

2.10 Summary

This chapter summarized background in both psychology and computer science necessary for this thesis. We discussed the link between gaze and social patterns that differ in people with autism and typically developed people, namely in children. We have also discussed where eye trackers, robots, and computer vision technologies have a place in child developmental research, and why using socially assistive technology with children with autism is both feasible and desirable. We discussed current research in the autism therapy and diagnosis space that uses robots, and included some important criticisms from the psychology community. We are now ready to discuss the experiments designed and conducted for this thesis.

Chapter 3

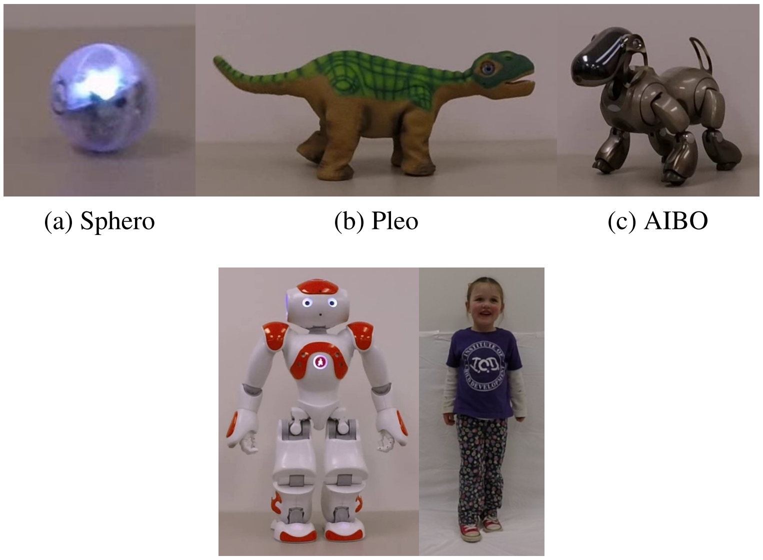
Eye Tracking Experiments

3.1 Problem definition and experimental design

We hypothesized that children would show statistically significant preferences between robots or a human peer in a paired visual comparison task and preferential orientation in a visual search task; if proven true, these preferences should influence the design of robots and experiments for our primary research interest, robotics in autism. We choose three robot types identified and discussed in [37]: anthropomorphic (having human qualities), zoomorphic (having animal qualities), and caricatured (having animated or cartoon qualities), leaving out one of the types, functional (designed purely for an intended operation).

To test our hypothesis, we use eye tracking, a technology commonly used to measure attention and gaze preferences [58], to compare and contrast five different items of four types: a social being (a 4 year old human girl), an anthropomorphic robot (the humanoid NAO), two zoomorphic robots (the dinosaur Pleo and the dog AIBO), and a caricatured robot (the ball Sphero). We designed two tasks: a dynamic video comparison that displayed two of these objects on screen at a time, and a static image search that displayed all five of the objects on screen at a time. These are shown in Figure 3.1.

We start by discussing the participants and validity criteria for inclusion in the study results. We refer to the four robots or the social being as animates (short for ‘animate



(a) Sphero (b) Pleo (c) AIBO

(d) NAO (e) Social

Figure 3.1: The five animates used in both eye tracking tasks.

object'). The Sphero robot (Figure 3.1a) is a mostly transparent sphere that can display different colors, sized 3 x 3 x 3 inches. The Pleo robot (Figure 3.1b) is a green and tan, Diplodocus-shaped dinosaur, about 4 inches wide, 8 inches tall, and 15 inches long. The AIBO robot (Figure 3.1c) is a smooth, brown, plastic dog, about 7 inches wide, 11 inches tall, and 12.5 inches long. The NAO robot (Figure 3.1d) is a white and orange humanoid robot, about 2 feet tall; Chapter 4 details the NAO robot more fully. The Social being (Figure 3.1e) is a four year old human girl, about 40 inches tall. The first and second tasks use the same animates, but differ in how many animates are on-screen and whether the display is dynamic or static.

The eye tracker used is a TOBII TX300 system from Tobii Technology AB [1]. The tasks were displayed on a 27-inch ASUS wide-screen monitor set to refresh at 120 Hz. The X300 is a desk-mounted dark pupil tracking system that has a temporal resolution of 3ms (sampling rate 300HZ) and a gaze accuracy of 0.4° according to the manufacturers; we compute our own gaze accuracy verification in the next section.

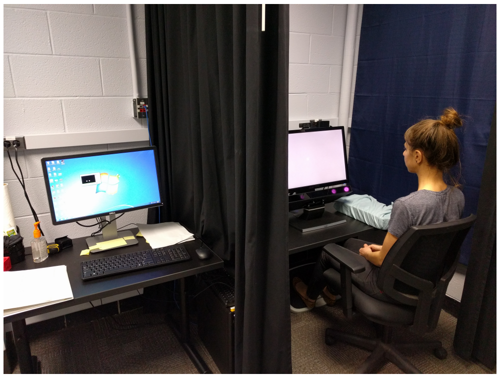


Figure 3.2: The TOBII eye tracking system. The researcher controls the experiment at the computer on the left, and the child sits on the parent's lap on the right.

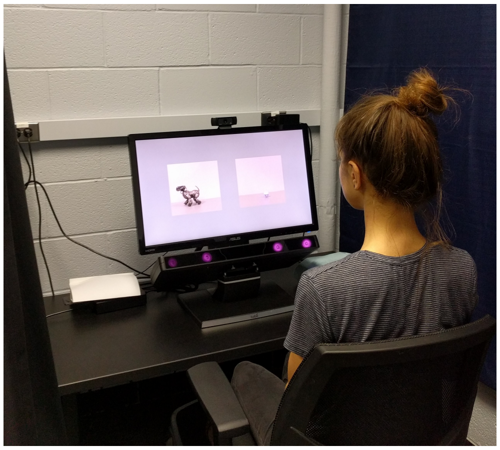


Figure 3.3: A closeup sample of what a participant sees during Task One.

The eye tracking hardware setup is shown in Figure 3.2 for both researcher and participant views; Figure 3.3 shows just the curtained partition where the child is located and what the child would see. During the eye tracking tasks, the curtain is closed and the lights are dimmed to the lowest setting for every participant, to ensure the screen is the only thing of interest during the task.

3.1.1 Data quality assessment

To ensure the quality of our data, we performed a calibration verification on the eye tracking accuracy before the eye tracking tasks. Actual eye tracking accuracy can vary from the manufacturer's reported accuracy, due characteristics of the room such as lighting, participant eye color and positioning [10], participant age [52], and even race [119], so we follow the calibration assessment protocol in [25]. We computed a calibration accuracy for each participant who took this calibration task and contributed at least 1 valid trial to each of the two tasks described in the following section.

Calibration trials were performed with 44 of the 50 participants who finished the eye tracking tasks described in Section 3.2 and Section 3.3. Of those 44, 33 participants contributed at least one calibration accuracy trial in the task. To compute calibration accuracy, we first computed the longest fixation within six degrees of visual angle on each of five calibration targets. We chose the longest fixation because it is most likely to be the fixation that was intended to land on the target. We then calculated the average accuracy and (x, y) coordinate error of those trials per participant (ignoring trials where there was no fixation, or no fixation closer than six degrees of visual angle). We removed trials where calibration accuracy was more than two standard deviations above the overall mean accuracy for the group (higher degrees indicates a higher error). This resulted in 32 of the 33 participants that contributed at least one valid trial to the calibration verification task. The average accuracy of these 32 participants is 1.026 degrees of visual angle, with an average RMS for the x, y coordinates of 0.25 and 0.27 (rounded to two places), respectively.

We then calculated the correlation coefficient between the calculated accuracy, the root mean squared (RMS) error of the recorded on-screen x and y coordinates of participant gaze, one dependent measure from the first task, and two dependent measures from the second task. The first task metric is the preference a participant had for looking at the NAO robot or the Social being; the second task metrics are the proportion of first looks to the NAO robot and the proportion of first looks to the Social being. There is a strong relationship between the X and Y coordinate RMS error, but no significant relationship between accuracy or RMS error and any measure from the two tasks; see Table 3.1 for details. This means that the data quality (i.e. accuracy of the eye tracking equipment) did not influence any of our dependent measures. Exactly one participant who passed the calibration verification task did not contribute enough trials for their results to be included in the Task Two analysis, resulting in 31 participants who passed calibration verification and enough trials below the threshold for every other task.

Table 3.1: Correlation coefficient matrix for accuracy, RMS error, and three dependent measures from the eye tracking tasks (rounded) for 31 participants.

	Accuracy	RMS X	RMS Y	Task 1	Task 2-NAO	Task 2-Social
Accuracy	1.	-0.06	0.00	-0.20	-0.15	0.00
RMS X	-0.06	1.	0.94	-0.12	0.21	-0.30
RMS Y	0.00	0.94	1.	-0.09	0.25	-0.30
Task 1	-0.20	-0.12	-0.09	1.	-0.02	-0.09
Task 2-NAO	-0.15	0.21	0.25	-0.02	1.	-0.37
Task 2-Social	0.00	-0.30	-0.30	-0.09	-0.37	1.

3.1.2 Data analysis approach

We are ultimately interested in children's preferences for a particular animate. Preferences are found in several ways: dwell time on a stimulus, reaction time to stimulus, and where a participant looked during the duration of a stimulus. Longer dwell times, faster reaction times, and more fixations all indicate a stronger preference for the stimulus. We used the R language in the RStudio software, which contains a multitude of tests for determining statistically significant differences between samples. When checking for statistical significance, we assumed the worst about our data, which is that the data did not come from a known or even normal distribution, and used non-parametric tests. We used the Friedman test for stimuli comparing multiple items to control for repeated tests, and the Wilcoxon rank-signed test when comparing two distinct items. We tested for effects of age and sex on our tasks with Spearman's rank correlation coefficient. Because not all participants underwent the calibration verification protocol, in the interest of full disclosure we generate and report three sets of results. These sets are: the minimum sample of high quality results (results only from children who passed the calibration verification task, 32 participants), the medium sample, of participants with known high quality results *and* those with the benefit of the doubt (those who passed the calibration verification task, and those who did not complete that task, 38 participants), and the maximum sample, of every child who took the eye tracking task in question (regardless of passing the calibration task, 50 for Task One and 49 for Task Two).

In each case, we hold the participants in each group to the valid trial criteria described next in Section 3.1.3, dropping trials as needed if they did not pass the moving threshold.

3.1.3 Participants and valid trial criteria

Participants were recruited from a laboratory-maintained database at the University of Minnesota's Institute of Child Development. Written informed parental consent was ensured in advance of all testing; all research was approved by the university's Institutional Review Board. Though 52 children (28 males, 24 females) were recruited, 2 children could not finish the eye tracking study, resulting in a final analysis of 50 total children, 28 males and 22 females, ages 18 to 45 months (mean 29.5, st. dev. 8.89 months). For each task, we require participants to contribute enough valid trials for inclusion in that task. Trials in which participants spent less than two standard deviations less than the mean time spent on animates (as opposed to total time on screen) were not included; this mean was generated by averaging all trials for a task, excluding participants who never looked at any animate at all in a trial. Under this validity criteria, if a child contributed fewer than half of all possible trials, the child was excluded from the task. We report the ratio of trials in which participants started centered on the screen to total valid trials after applying the inclusion criteria. Validity criteria and analysis criterion are given in Table 3.2.

Table 3.2: Valid trial criteria in Task One and Task Two for the maximum group (all children who participated in the tasks, regardless of passing calibration verification). Results rounded to one decimal place for ease of viewing; in practice, numbers were not rounded. One participant from Task One did not participate in Task Two due to researcher error.

	Task One	Task Two
Total Participants	50	49
Mean time on animates	3917.4 ms	2288.5 ms
Std. Dev.	1255.3 ms	676.3 ms
Validity Criteria	1406.7 ms	935.9 ms
Ratio Centered to Valid Trials	40.8%	68.4%
Analyzed Participants	48	45

3.2 Task one: dynamic video paired visual comparison

3.2.1 Experimental setup

The first eye tracking task we designed compared short videos of pairs of animates. Each possible pair (NAO-AIBO, NAO-Pleo, NAO-Sphero, NAO-Social, AIBO-Pleo, AIBO-Sphero, AIBO-Social, Pleo-Sphero, Pleo-Social, and Sphero-Social) was displayed in random order in 5-second videos. We created two sets of unique video pairs for each animate, and showed each video pair in both possible locations (e.g. NAO on left, Social on right, and the same animate videos but with Social on left, NAO on right) to guard against possible screen side bias. This resulted in 40 videos overall, with a brief break for the participant as needed between these two blocks. The first set of randomized 20 videos is called Block 1; the second set of 20 randomized videos is called Block 2. Each video is separated by a small, centered, attention-getting shape to re-center the participant's gaze. This attention-getter is meant to guide the participant's gaze to the neutral center of the screen before any stimulus is shown, so as not to bias any animate with an unintentional fixation. When the researcher overseeing the eye tracking tasks determined the child was looking at the screen during an attention-getter, she triggered the next video with the keyboard.

Along with controlling for the side of the screen the animate was on, we controlled

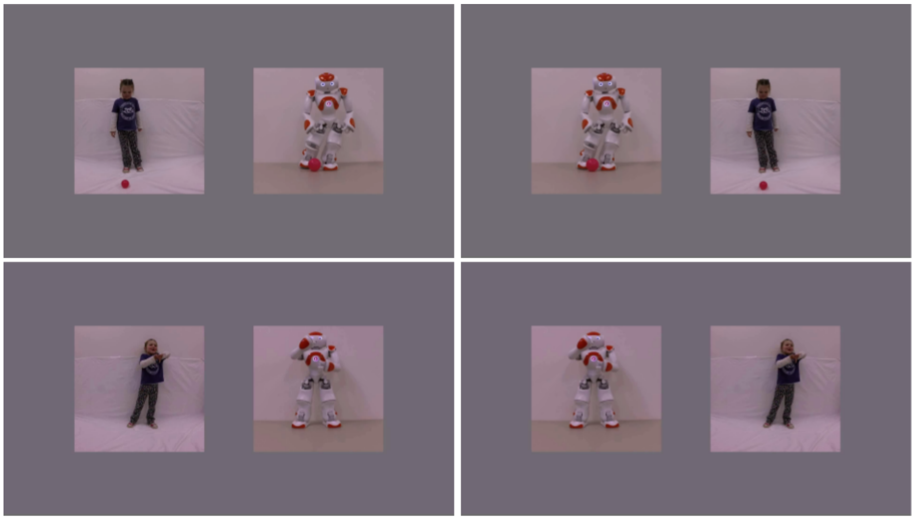


Figure 3.4: Sample frames from original and reversed images from both blocks of the Social-NAO pair from the dynamic video comparison task.

for similar movements by the on-screen animates. Each set of videos had the robots or Social being moving in a similar way, or with similar objects, to ensure the motions were not unfairly competitive. For example, one video pair of NAO and Social had each kick a pink ball towards the camera. Another pair, Social and AIBO, sat on the floor and moved slightly – the AIBO robot slowly rotated its head and shook its ears, and the Social being looked at the camera and rubbed her leg. A sample frame from each of the four NAO-Social videos is shown in Figure 3.4; the top two frames are the original and reversed videos in Block 1, and the bottom two frames are the original and reversed videos in Block 2.

In Block 1 and 2, each video is 561x536 pixels, in color, with a bland background (a white wall and tan floor). Videos in both blocks, attention getting shapes, and static picture displays are overlayed on a 1920x1080 pixel gray background. Each pair of videos is separated by 202 pixels of gray space. Participants sit roughly 60 cm away from the screen, at which distance one degree of visual angle is about 42 pixels. We included a 1 degree of visual angle buffer around each item, such that if the child looks within 42 pixels of a video, we consider that child to be looking at the video. This region that counts as ‘looking at an

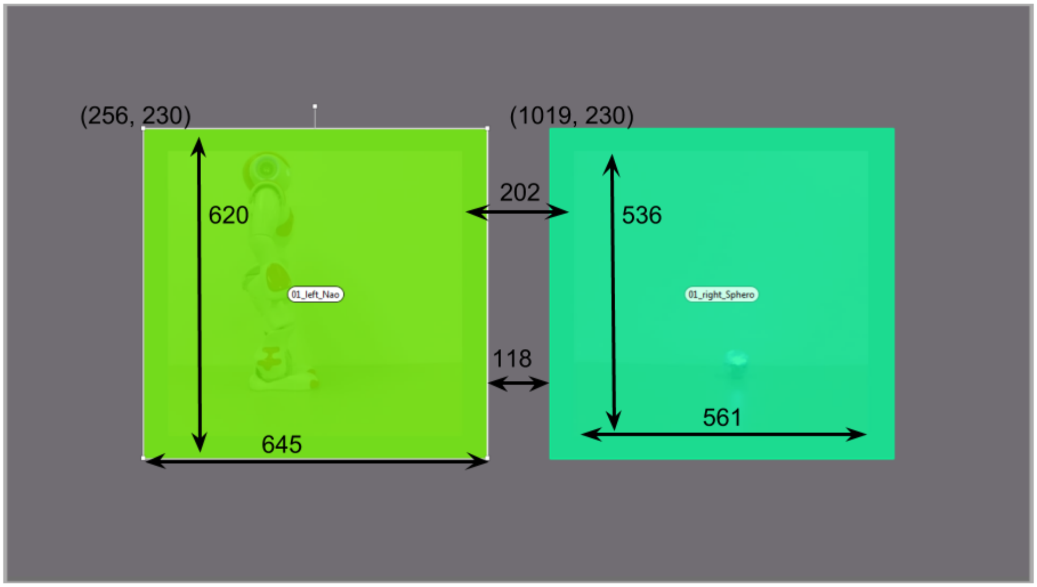


Figure 3.5: Sample frame of the dynamic video comparison task, including the area of interest buffers.

item' is called an area of interest, or AOI. The videos are 2 degrees of visual angle from each other (including the buffer zone), and when a child's gaze is focused on the center of the screen (from looking at the attention getting shape) at the beginning of the video, the child must move their gaze a full degree of visual angle to get to one video or the other. This gives us confidence that the child meant to look at the video, rather than recorded results being influenced by natural eye jitter or badly calibrated hardware. Figure 3.5 shows the location markup on a sample video. Each attention getting shape is 40x40 pixels, centered in the middle of the screen. Figure 3.6 shows a sample attention getter.

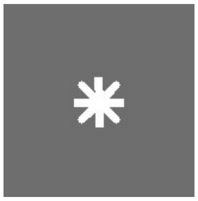


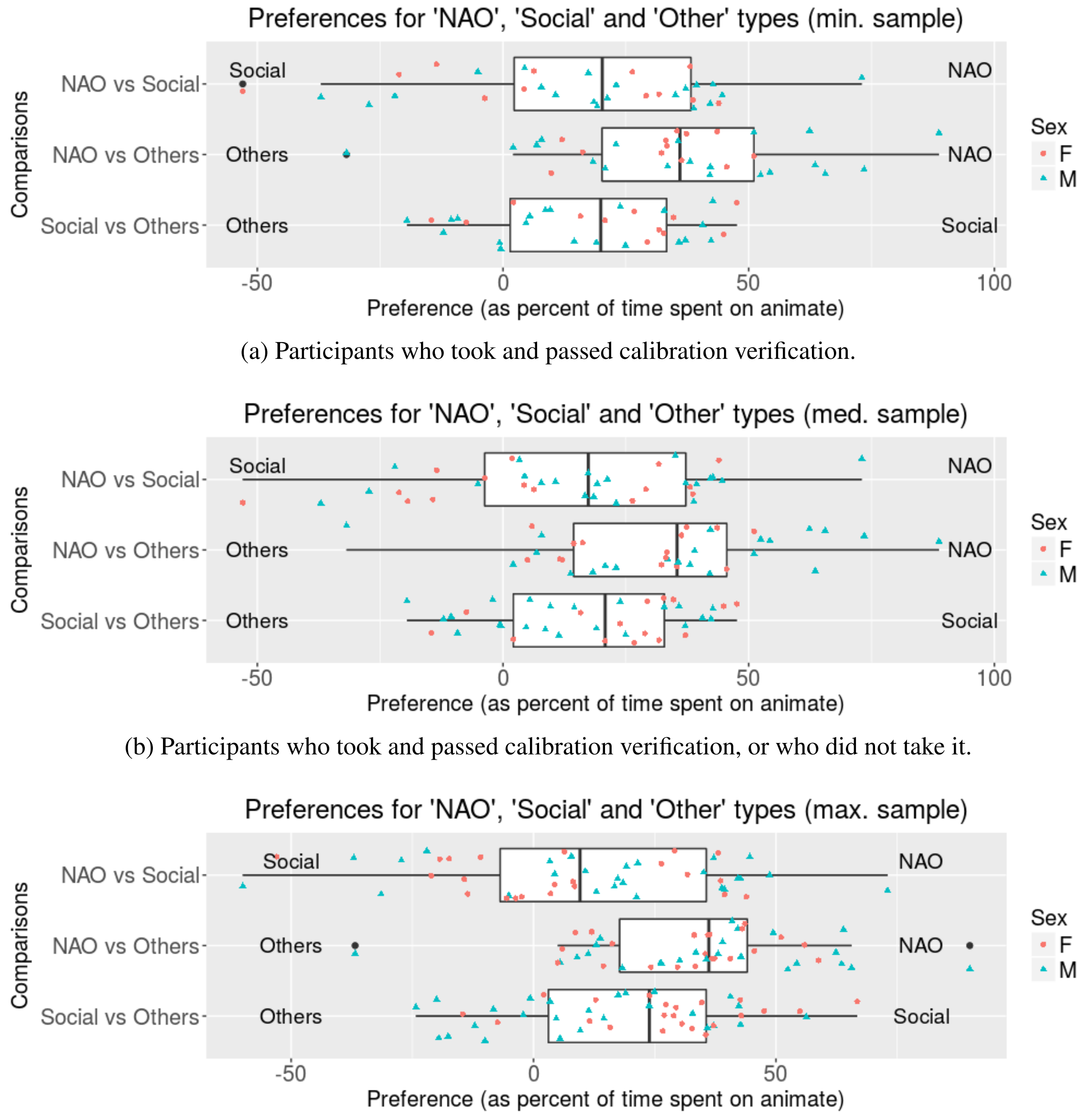
Figure 3.6: Sample attention getter. Not to scale.

Table 3.3: Ratio of time spent on animate in comparisons of Social vs other robots, NAO vs other robots, and NAO vs Social for the Maximum group.

	Social vs Others	NAO vs Others	NAO vs Social
Min.	-24.270	-36.80	-59.980
1st Qu.	3.086	17.78	-6.876
Median	23.890	36.15	9.589
Mean	19.080	34.07	10.190
3rd Qu.	35.600	44.08	35.610
Max.	66.770	89.98	73.030

3.2.2 Results

We first analyzed the preference of the participant in a particular pair, or trial. The preference is calculated by taking the time spent on each stimulus in a pair, dividing that number by time on screen, and subtracting the preferences. Preliminary analysis indicated a strong preference for Social and for NAO, so we collapsed the other robots (Sphero, Pleo, AIBO) into a single category we called “Other.” The subtractions are NAO ratio minus Social ratio (so a positive number indicates preference for NAO, and a negative number indicates preference for Social), NAO ratio minus Other ratio (positive number indicates preference for NAO), and Social ratio minus Other ratio (positive number indicates preference for Social). For this task, we added another validity criteria that a child had to contribute at least 50% valid trials for videos comparing the NAO and Social animates; this ensures that only participants who contributed to the Social vs. NAO comparisons are represented in the over-all preferences. No participants were dropped due to this added constraint. If the robot pair did not include NAO or Social, we ignored it. Figure 3.7 shows this preference ratio.



(a) Participants who took and passed calibration verification.

(b) Participants who took and passed calibration verification, or who did not take it.

(c) All participants (regardless of taking or passing calibration verification).

Figure 3.7: Ratio of time spent on animate in comparisons of Social vs Other robots, NAO vs Other robots, and NAO vs Social. Empty colored circles indicate a data point; solid black circles indicate outliers in the box-and-whisker plot. There was not a significant effect of which group participants were in on the over-all preference trends.

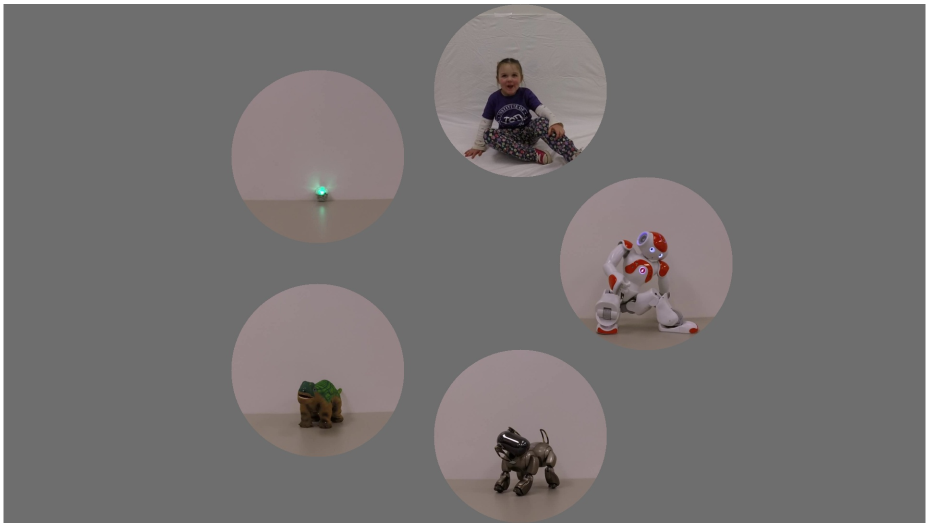


Figure 3.8: Sample image from the static search task.

3.3 Task two: static image visual search

3.3.1 Experimental setup

The second eye tracking task we designed measured which animate participants responded to first and in subsequent fixations, how fast participants responded to the first animate (excluding those who did not start the trial looking at the center of the screen), and which animate participants gazed at the most throughout the trial. We showed five images in random order, with all five animates ordered in different (specific) locations in each image; each image was shown on screen for 3 seconds. Images were separated by a brief attention-getting shape in the center of the screen. A sample image is shown in Figure 3.8.

Each picture is reduced to a circle 180 pixels in diameter, evenly spaced about the center of the gray background at $(960, 540)$, exactly 84 pixels away from their nearest neighbors, and 110 pixels away from the center of the screen, which is roughly 1.6 degrees of visual angle from the center. See Figure 3.9 for a sample location markup on the static pictures.

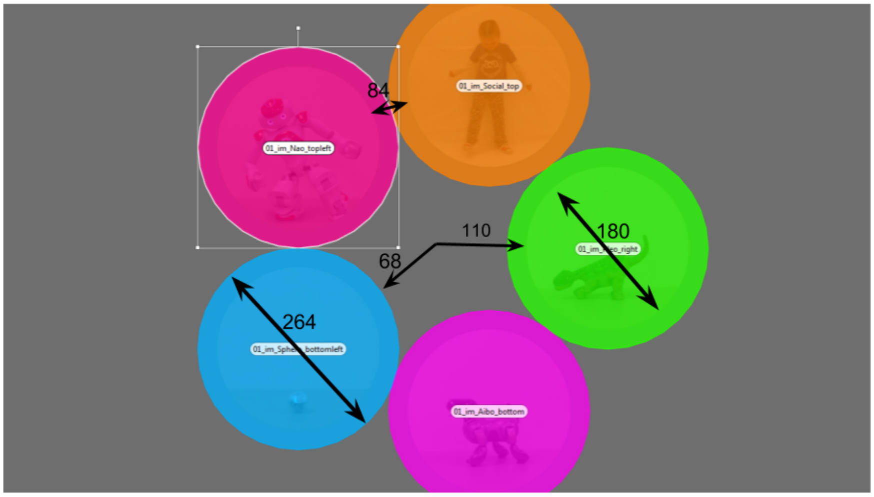


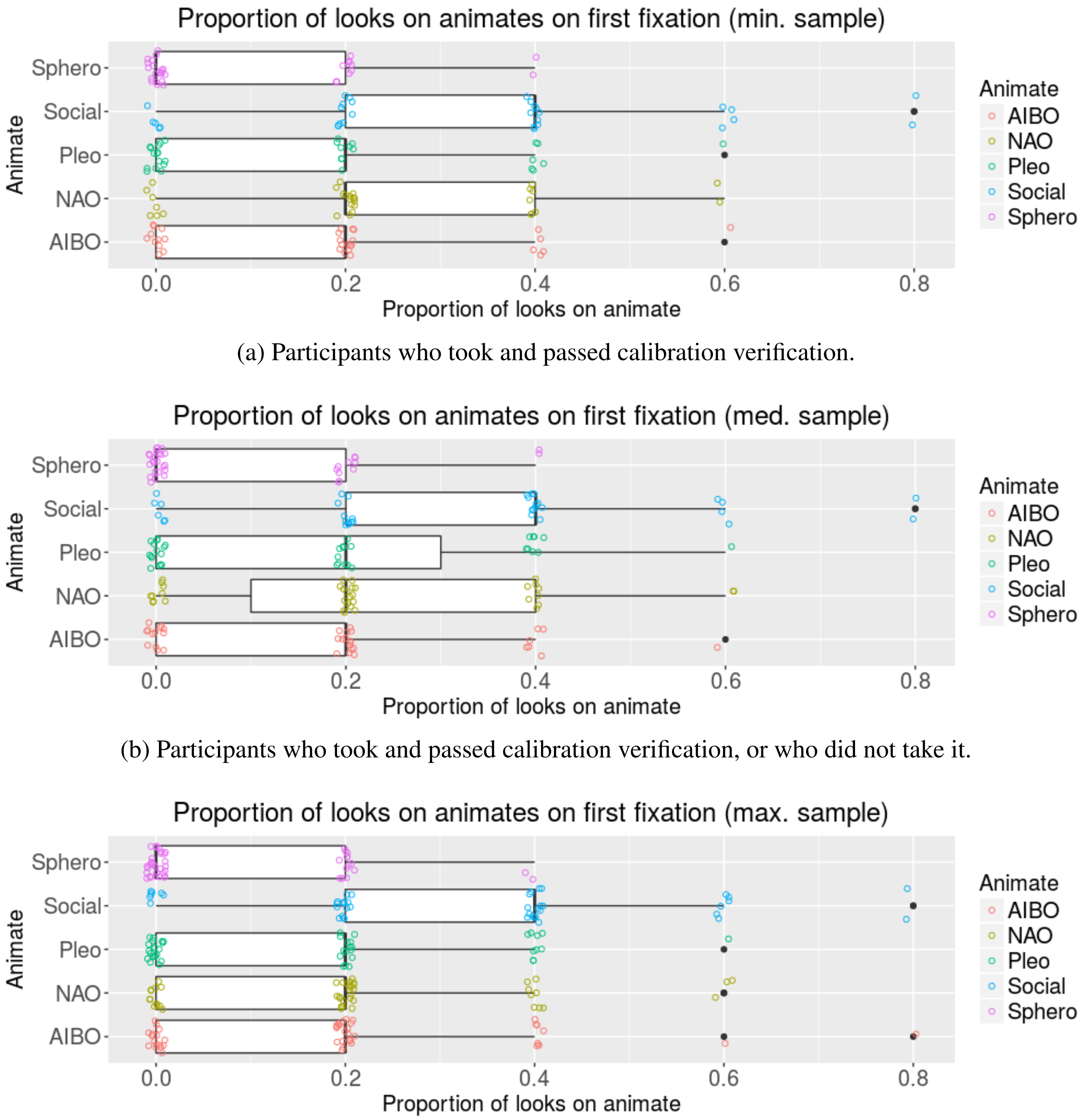
Figure 3.9: Picture dimensions including the area of interest buffers.

3.3.2 Results

The proportion of first looks to an animate is shown in Figure 3.10. These results indicate participants on average look more at the Social being on their first fixation than at any other animate. This logically leads us to ask if participants have any preference for the humanoid robot over any of the other robots, or if the different types of robots are equally interesting as each other, and equally less interesting than Social. Consequently, we calculated the proportion of looks on animates for the first ten fixations, shown in Figure 3.11. There is a statistically significant difference between Social and NAO on the first fixation ($p < 0.05$) in the maximum group, and none on subsequent fixations. Though the medium and minimum groups do not show statistically significant preferences between Social and NAO, the same trends are very visible.

Results also showed statistically significant ($p < 0.05$) differences between ratio of Social to the other non-NAO robots for all fixations until fixation 8 (not shown on graph for simplicity). While there appears to be a sex difference for the proportion of fixations per animate, tests were underpowered to detect the effect of sex; see Figure 3.12 for male participants and Figure 3.13 for female participants.

The average time on an animate is shown in Figure 3.14, and shows that participants spend the most time on the NAO and Social animates. The average reaction time to the first fixation (time to look from the center of the screen to the first animate) did not vary by animate; see Figure 3.15 for average times to react to an animate. This graph only includes children which began a particular trial with a centered gaze; the color bars indicate the error of the average point if a participant's first fixation was on the same animate in different trials.

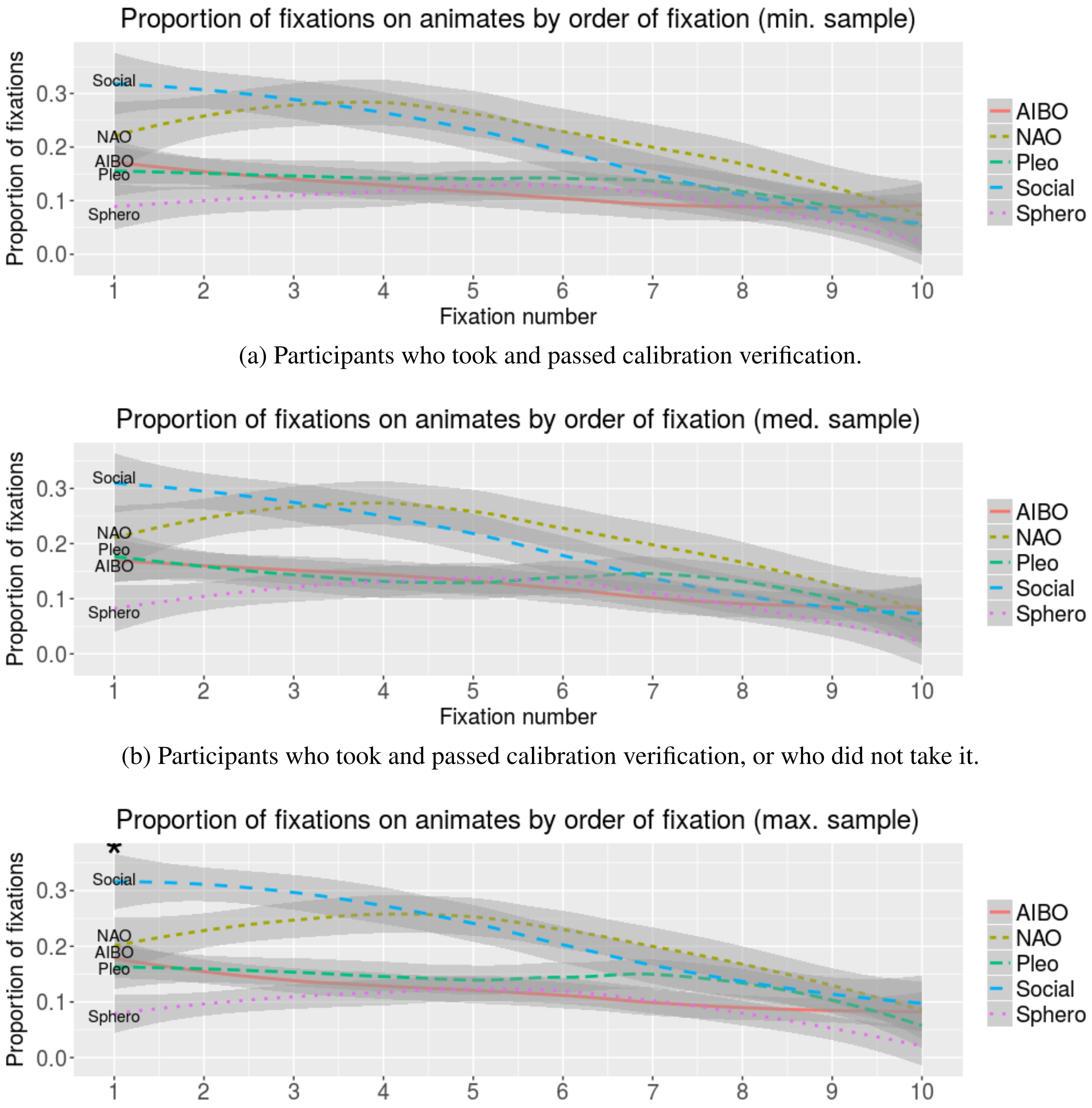


(a) Participants who took and passed calibration verification.

(b) Participants who took and passed calibration verification, or who did not take it.

(c) All participants (regardless of taking or passing calibration verification).

Figure 3.10: Proportion (out of 5) of first looks on different animates in the static search task. Empty colored circles indicate a data point; solid black circles indicate outliers in the box-and-whisker plot.

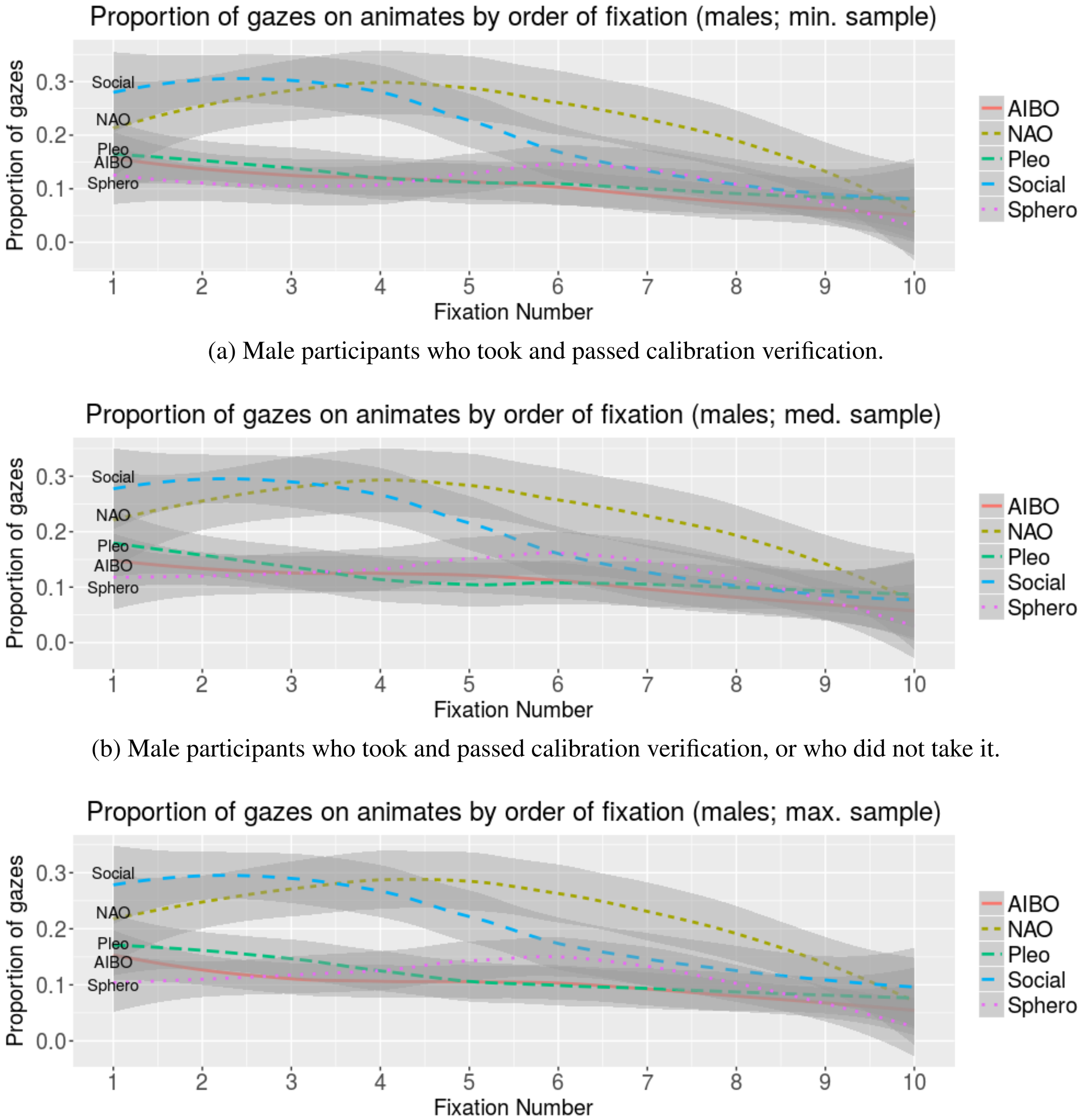


(a) Participants who took and passed calibration verification.

(b) Participants who took and passed calibration verification, or who did not take it.

(c) All participants (regardless of taking or passing calibration verification).

Figure 3.11: Ratio of gazes on animate by fixation number. Asterisks indicate statistically significant gaze differences between Social and NAO ($p < 0.05$).

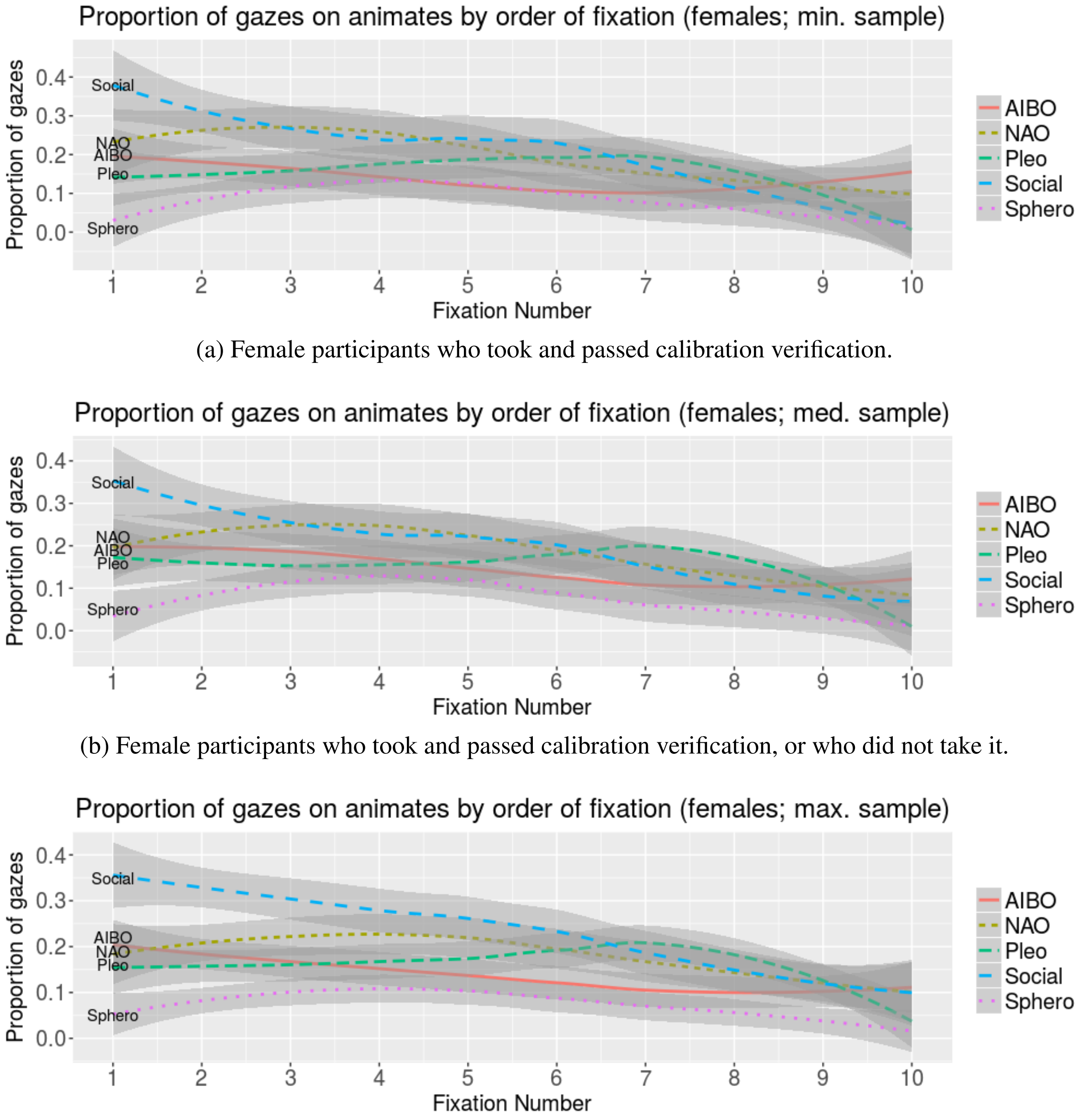


(a) Male participants who took and passed calibration verification.

(b) Male participants who took and passed calibration verification, or who did not take it.

(c) Male participants (regardless of taking or passing calibration verification).

Figure 3.12: Ratio of gazes on animate by fixation number for males.

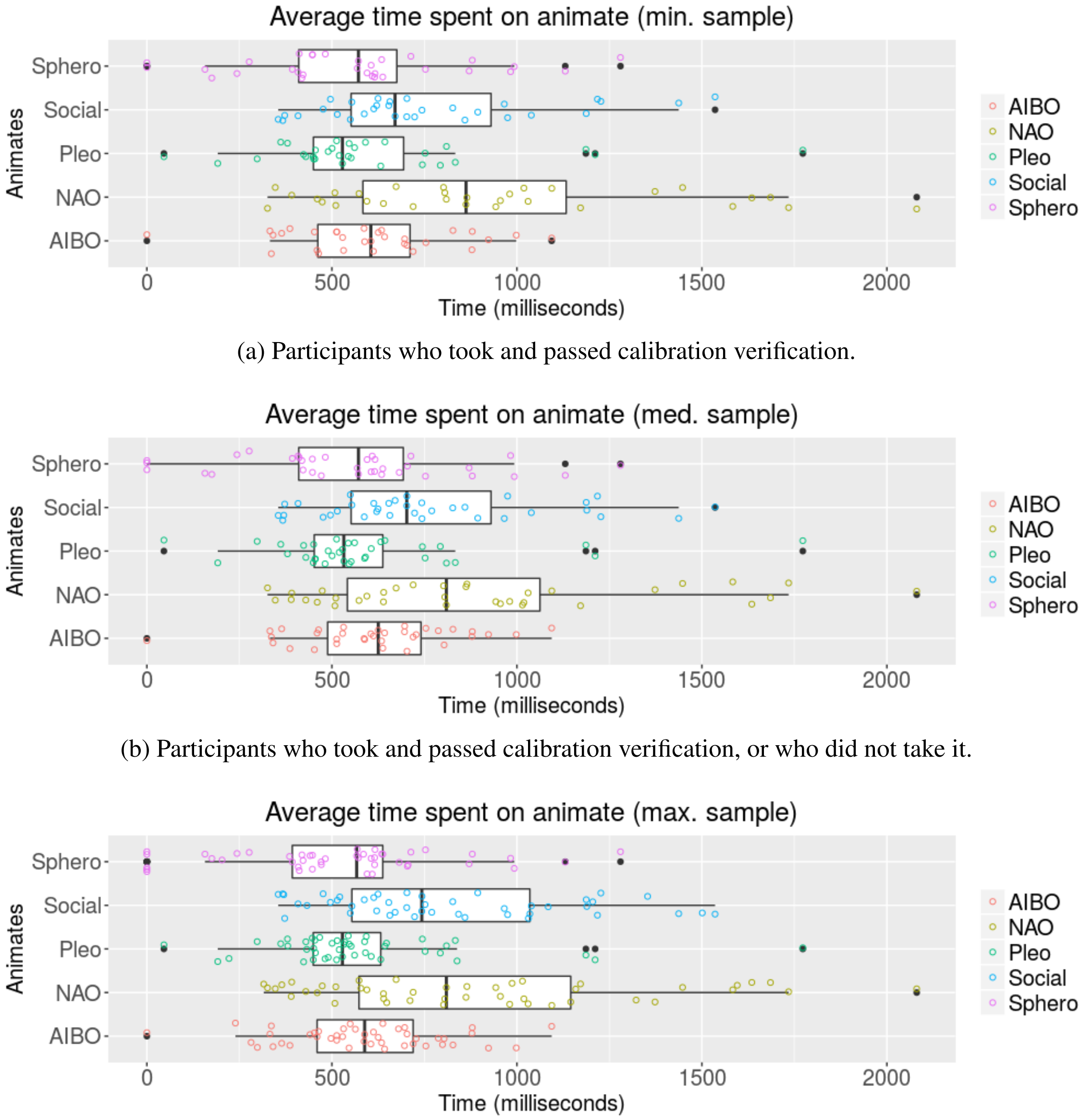


(a) Female participants who took and passed calibration verification.

(b) Female participants who took and passed calibration verification, or who did not take it.

(c) Female participants (regardless of taking or passing calibration verification).

Figure 3.13: Ratio of gazes on animate by fixation number for females.

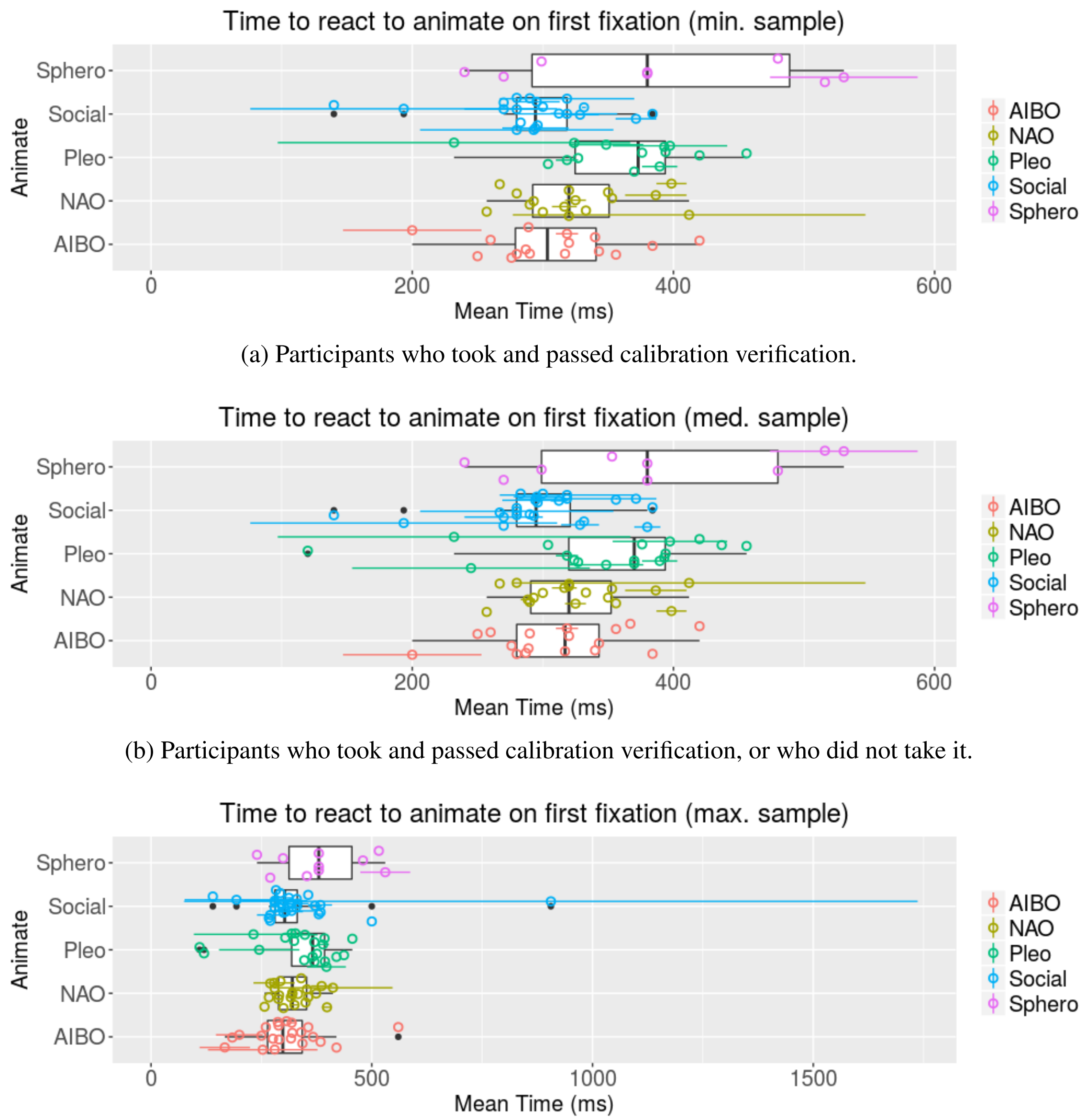


(a) Participants who took and passed calibration verification.

(b) Participants who took and passed calibration verification, or who did not take it.

(c) All participants (regardless of taking or passing calibration verification).

Figure 3.14: Average time spent on animate in any valid trial. Empty colored circles indicate a data point; solid black circles indicate outliers in the box-and-whisker plot.



(a) Participants who took and passed calibration verification.

(b) Participants who took and passed calibration verification, or who did not take it.

(c) All participants (regardless of taking or passing calibration verification).

Figure 3.15: Average time to react to an animate for participants who started the trial with centered gaze. Empty colored circles indicate a data point; solid black circles indicate outliers in the box-and-whisker plot.

3.4 Discussion

The first eye tracking task shows a preference for the NAO robot over the Social being, a strong preference for NAO over the other robots, and a strong preference for the Social being over the other robots. This indicates children are more interested in looking at the NAO and at the Social being over the other robot types displayed. The preference for NAO over Social is not as strong, indicating they may be nearly as interesting, but perhaps something about the novel technology aspect tips in NAO's favor. The second eye tracking task shows a clear initial preference for looking at the Social being, followed by a rise in interest in the NAO robot. Both animates dominate proportion of looks at animates, reaction times to animates, and total time spent looking at animates. We interpret these results as a strong interest in Social beings and humanoid robots more so than other robot types. Taken together, these eye tracking experimental results suggest children in this age range show a preference for the NAO that appears similar to the preference for live human peers, and the preference for NAO is clearly greater than those of other robots.

Future research should further explore if there is a preference for the gender of the displayed social beings given the gender of the participant. While we did not see statistically significant effects of gender on preference of NAO to social being, the rising interest in the NAO during the static image search was more visible in male participants than female participants. Researchers should also consider comparing a larger number of the robots currently on the market or used in research, and comparing those robots to each other in eye tracking experiments to quickly validate the preference of one humanoid over another for the particular research question.

Chapter 4

Human-Robot Interaction Experiments

4.1 Problem definition and experimental design

As part of our goal of helping diagnose Autism Spectrum Disorder earlier in life, we first need to further characterize the autism phenotype. To do so, we seek quantitative data about participants' development and how they react to a social robot.

We begin by collecting results of both well-established and newer developmental assessments. We have a large sample size of 59 children (detailed further in Section 4.2), all of a consistent age group (two to four years), and we use a documented, portable, and psychologist-reviewed robot interaction experiment. For each participant, we do the following assessments: Vineland Adaptive Behavior Scales [107], Mullen Scales of Early Learning [74], video-referenced rating of Reciprocal Social Behavior (vrRSB) [72], Social Responsiveness Scale- Second Edition (SRS-2) [20], Child Behavior Checklist (CBCL) [3], Infant Toddler Social Emotional Assessment (ITSEA) [12], Strengths and Difficulties Questionnaire (SDQ) [46], Repetitive Behavior Scale- Early Childhood (RBS-EC) [122], and Children's Social Understanding Scale (CSUS) [110]. Nearly half of these children also participated in the eye tracking experiment, detailed in Chapter 3.

We then designed a human-robot interaction (HRI) protocol for use with very young children. The HRI experiment is a 9-15 minute long interaction designed for children aged two to four years old with the NAO robot from SoftBank Robotics (previously Aldebaran Robotics). The robot plays interactive games with the child, including mimicry games, dances, and looking games. The entire interaction is recorded from several different perspectives for later offline analysis. The ultimate goal is to be able to automate interaction detection, including the distance from the child to the other actors (caregiver, experimenter, and robot) at any point in time, the object of the child's attention at any point in time, and child's reaction to the robot's bids for attention or reaction.

During the HRI experiments, we introduce a child to a new friend Robbie the Robot (a NAO). Robbie plays different games such as I Spy (a looking game that encourages the child to find objects in the room), Simon Says (a behavior imitation game that encourages the child to copy motions possible with gross motor skills like clapping and waving), and several dances. The set of games is in the same order for every child. The experimenter controlling the robot imitates some of the robot's movements, encourages the child to do the same, and plays along during some of the looking games, and encourages the child to do the same.

4.2 Participants

Children were recruited from a laboratory-maintained database at the University of Minnesota's Institute of Child Development. Written informed parental consent was ensured in advance of all testing; all research was approved by the university's Institutional Review Board. The experiment initially recruited 65 total children; of these, five of the experiments were not recorded due to equipment failure, or not performed because the child was judged not capable of interacting calmly with the robot. One participant laid on the floor while the caregiver and both researchers in the room tried to encourage the child to interact; as no other child had this highly unusual configuration (3 adults interacting directly with the child), this data was not included for analysis. Six of the 59 remaining experiments ended early due to equipment failure or because the child wanted to end the experiment early; scoring data for those 6 children were included in all analyses, and interaction proxemics and data were analyzed right up to the stopping point of the experiment. The final analysis involves 59 children – 31 boys, 28 girls, minimum age 25.3 months, maximum age 45.2 months, mean age of 33.1 months, st. dev. of 4.6 months. Notably, 29 of the 59 participants that were recorded for the HRI study discussed here also participated in the eye tracking study; these participants are discussed further in Chapter 5.

4.3 Experimental method

The NAO robot is a 2 foot tall humanoid robot; it has 25 degrees of freedom and a multitude of sensors, which include touch sensors in hands and feet, sonar, microphones, video cameras, and infrared LED receivers. NAO can also change the colored LEDs near what look like eyes, blink various lights around its body, and play music. NAO is capable of standing up, sitting down, walking, tracking movement, and tracking faces. NAO comes with a library of common behaviors (sitting, standing, blinking) as well as capabilities to design and record new movements, such as dances. Figure 4.1 shows the NAO robot in the middle of a dance.

Our NAO HRI program is composed of seven pieces designed to encourage social interaction, which we call a ‘press’ for social interaction. Details on programming the NAO robot, as well as the specific phrases used by the robot during the interaction, can be found in Appendix A. During all HRI experiments, there is always the child, the robot, and the experimenter controlling the NAO in the room; there is also usually at least one caregiver and another researcher. The HRI experimenter is seated on the floor next to the robot, controlling the robot via laptop if necessary, while the child is watching the robot (seated or standing on the floor, or seated on a chair or a parent). If the child is too shy to interact with the robot alone, they may sit next to or on the caregiver. If the child is comfortable, then the caregiver is seated away at a table talking with the other researcher, being interviewed for a development assessment regarding the child.

The interaction is recorded from up to four perspectives, which include from up to two sides of the room, from the robot's perspective, and from a GoPro mounted on the ceiling. The GoPro records video at a resolution of 1280 x 960 pixels at 30 fps. The GoPro is the only camera that is always located in the same place, and it is the only view from which we are able to see all participants in the room (unless, as in a few instances, the child plays under a table in the room, in which case their location is easily estimated). Figure 4.2 shows part of a frame from an overhead view; the child participant is facing the robot, close to the experimenter, and the child's parent is in discussion with another researcher. The overhead GoPro recordings produce all the results we discuss in this thesis.

In order, the presses are: (1) Simon Says, (2) Happy Dance, (3) I Spy, (4) ByeByeBye, (5) I Spy, (6) Simon Says, and (7) Tai chi. Presses 1, 3, 5, and 6 are games, and presses 2, 4, and 7 are dances. Each press is separated by verbal encouragement from the robot and/or expressing interest in playing more games. Encouragement includes phrases like "you're doing a great job" and expressing interest includes phrases like "now I want to get up and dance." The next press is always started manually by the experimenter, to give the child a chance to seat themselves more comfortably, get a snack, or otherwise take a short break. If the experimenter does not trigger the next press, it starts automatically after 1 minute passes.

The "Simon Says" presses are adaptations of the game Simon Says, a common American game with one leader and at least one follower, in which the leader says to do an action that the followers then imitate. In our version, the robot says an action (e.g. "I touch my nose"), performs the action, then asks the child to do the same thing (e.g. "can you touch your nose?"). The "I Spy" presses are adaptations of the game I Spy, a common American game in which one person picks a visible item, says what it is without saying where it is, after which the others have to find the same item. In our version, the robot looks around, says what they see (e.g. "I see a cat"), and asks the child if she can see it ("Do you see a cat?").

Presses 2, 4, and 7 are dances, during which the NAO plays music from its speakers and dances along with the song. The (2) “Happy” dance is the song “If you’re happy and you know it,” in which the singer encourages an action, for example “If you’re happy and you know it clap your hands. Clap, clap!” The robot stands, plays the music, and does the action to the singer’s prompting in each case, which are: clap your hands, stomp your feet, say OK, and do all three. The (4) ByeByeBye dance is a standing dance set to a popular 90’s song, with lyrics but no encouragement for action. The robot dances along to the song. The (7) Tai chi dance is a very slow, modified opening posture from the martial art Tai Chi followed by a stretch, all set to slow music; the robot stands, moves its arms, and balances once on each foot in turn.

The experimenter varies in reacting to the NAO. Presses 1, 3, 5, and 6 are made up of six different imitation or seeing requests. The experimenter reacts to the robot on every other request, starting at the first one, to demonstrate to the child what to do. For example, the first request in Press 1 is “I touch my nose. Can you touch your nose?” At this, the experimenter says “Yes, I can touch my nose!” and does so, and then encourages the child to touch their nose. The second request is to fist bump (close hands and gently tap knuckles together), but the experimenter obviously looks away at a nearby sheet instead of paying attention to the robot or following the request. The third request is to clap twice; the experimenter affirms she can do it, demonstrates it after the robot demonstrates it, and asks the child if she can also perform it. This is for later coding child behaviors to determine if the child performs the request at the behest of the robot, or the experimenter, or the caregiver, or none of the above.

Figure 4.1: NAO dancing.

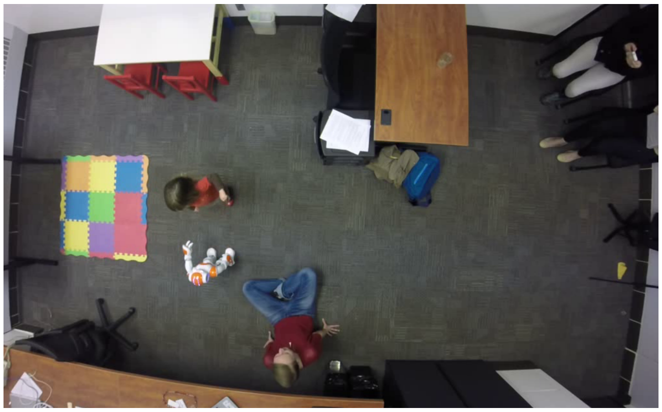
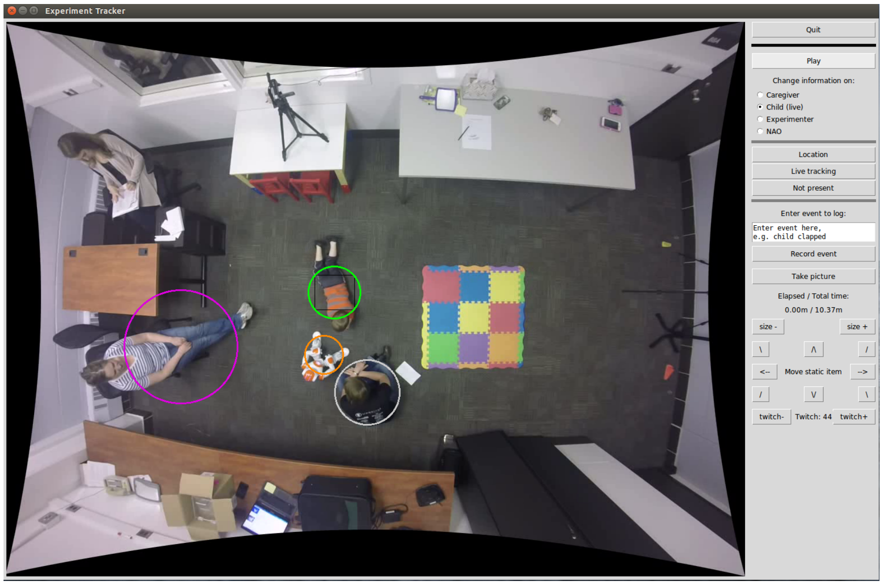


Figure 4.2: Example of the room layout as seen from the ceiling. The child stands in front of the NAO, which is also standing. The robot-controlling experimenter sits near the robot; the caregiver is seated in the top right of the screen shot, speaking with another researcher.

4.4 Video analysis: tracking people and robots

We are interested in the social distances between the child and the others in the room: the child's parent, the novel robot, and the novel human experimenter. To automatically determine these distances, we took two approaches. The first approach was to create software to track objects of interest in the overhead video footage. The software used the Python language and the open-source computer vision module, OpenCV. The program first reads a video and a configuration file that contains the names of the people or robot of interest; then it allows the user to choose the originating location of each item of interest and uses a tracking algorithm to track each area of interest over time. The tracking algorithm could be anything that exists in OpenCV (e.g. MedianFlow, TLD) or a programmer could write their own, but for our experiments the most reliable tracker available was the AdaBoost [39] algorithm implemented in OpenCV 3.1 and Python 3.4.

The first software version, seen in Figure 4.3, shows the experiment tracker interface. The user identifies the location of all the actors she wants to track (for example, NAO robot, child, parent, and experimenter). The software tracks each actor over time, and the user



Change information on:
○ Caregiver
● Child (live)
○ Experimenter
○ NAO

Enter event to log:

```
Enter event here,
e.g. child clapped
```

Elapsed / Total time:
0.00m / 10.37m

Twitch: 44

Figure 4.3: Sample run of object tracking in HRI experiment.

can adjust the size and location of the actor manually or simply let the tracker follow it. From the coordinates and size of each actor, we estimate the distance between each actor at every frame (given an off-line measure of the pixel-to-foot conversion rate in an undistorted video). This software version produced the distance data analyzed in this chapter.

The second version of the software is a more robust tracker that uses a neural network trained on over $16,000$ annotations of people and the NAO robot. Figure 4.4 shows an annotation from the HRI experiment described here. Other videos annotated for training the network included two other experiments in different rooms, both from ceiling perspectives. This software identifies all people and NAO robots in every frame over the length of a video. We wrote a script for processing the output box coordinates from the nueral network that first asks the user to label each actor (Figure 4.5) and then replays the entire video (Figure 4.6), allowing the user to re-key object labels as needed (because the tracker does not detect every object in every frame, occasionally generates false positives, or needs to track additional people entering the room or stop tracking people that have left the room). This CNN-based person and robot finder produced the results shown in Chapter 6.

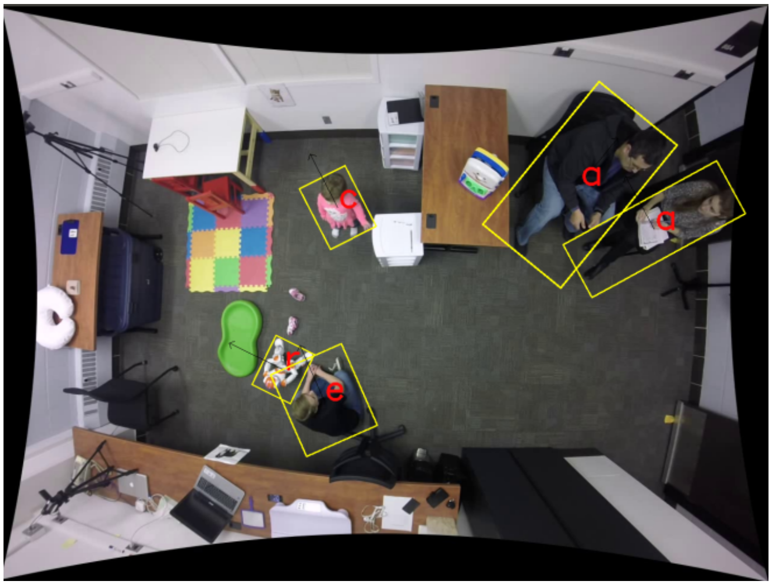


Figure 4.4: Sample annotation. Adults (caregivers and other researchers) are labeled ’a’, the experimenter controlling the robot is labeled ’e’, the child is marked with ’c’, and the robot is marked with ’r’. In practice, the network was trained on all humans as one class and the robot as another.

The neural network, an object detection network in the Region-based Convolutional Neural Network (R-CNN) family [45] known as Faster-RCNN, uses the TensorFlow Object Detection API [19]. Following the network retraining approach adopted in Rosebrock’s deep learning book [95], our network modifies a set of weights originally generated with the Common Objects in Context (COCO) dataset [63]. Using the publicly available COCO weights, we retrained the network on the human-robot annotations for $150,000$ epochs. We used computing resources from the Minnesota Supercomputing Institute, located in the University of Minnesota, to retrain the nueral networks in this thesis.

In either case, the software produces the (x, y) coordinates of each animate and the radius of the animate. We then convert the distance from raw pixel values to distance, using an empirically-found conversion rate of 52 pixels to 1 foot. Taking the center location of the animate, minus the radius of the animate, we calculated the distance in feet between objects of interest.

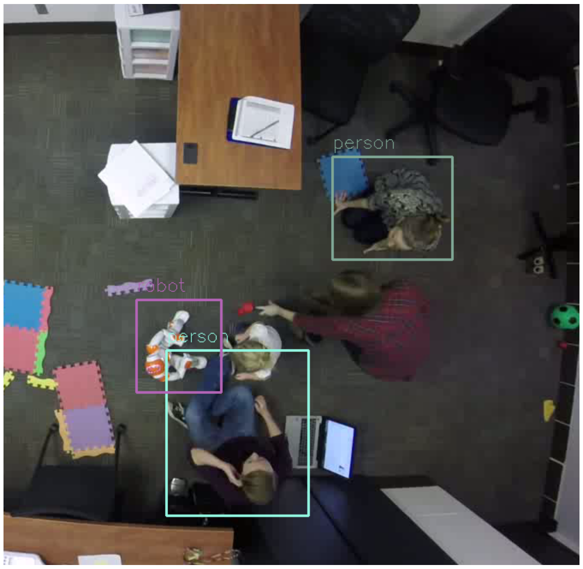


Figure 4.5: The user labels each actor detected in the first frame of the video.

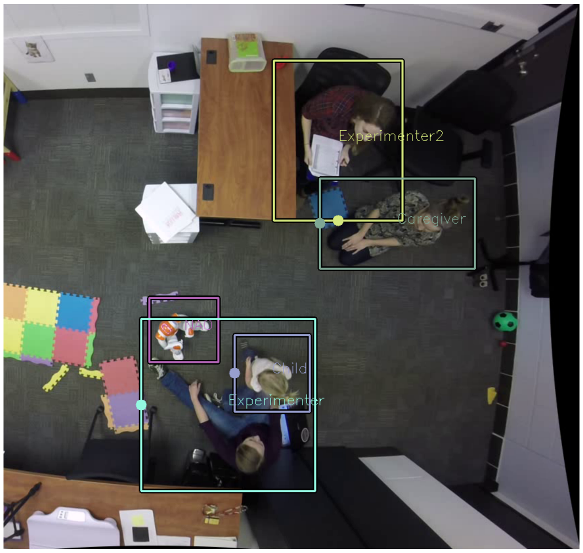


Figure 4.6: Sample run during replay of actors found throughout a video.

4.5 Video analysis: detecting person orientation

Along with knowing the distance from the child to other animates, we also want to know where she is looking. This requires calculating which direction the child is facing, or the orientation. The child's orientation allows us to calculate who or what the child is facing, which would tell us how interested the child is in gazing at a real-life robot, or if they prefer to look at the novel human experimenter, their parent, or something else in the room. To automatically determine the child's orientation, we needed to develop some new methods. Our work differs from other research in object and orientation detection by restricting ourselves to a single overhead camera; previous person orientation detection relies on more information. This data generally comes from additional cameras or from additional sensors, e.g. the commercially available depth sensor Kinect.

Much work estimates gaze orientation by tracking eyes or faces, which results in varying levels of accuracy and requires varying levels of cooperation. Highly accurate, commercially available eye tracking systems can require calibration for each participant and thus requires thorough cooperation in subjects, such as the TOBII system [1] used in Chapter 3, and also depends on the person being directly in front of the sensors. Other cooperation-free systems estimated gaze by first detecting facial features, such as work extending Active Appearance Models [22] and Constrained Local Models [23]. Both methods use facial features, meaning any video footage must show a lot of the subject's face.

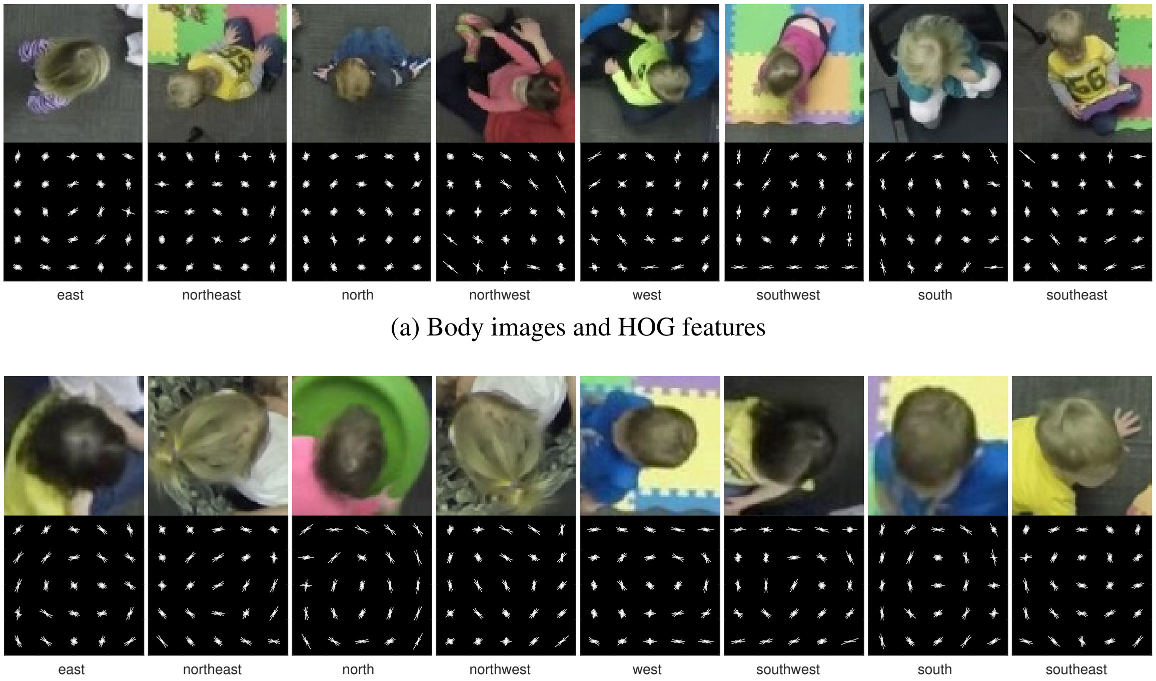
Multiple camera systems work well for tracking people and reconstructing the environment, but we cannot depend on multiple angles of our participants. Sivalingam studied a similar environment in [104], using multiple cameras and depth sensors to track children and adults in a classroom setting. That work was concerned with analyzing the motions of children and tracking movements and patterns across multiple sessions, whereas the children in our assessments will not repeat the assessments and give longitudinal data. Bidwell used an overhead camera on a child in a seated, known location to track gaze orientation from zero to 180 degrees left or right, but first found the orientation from another camera facing the participant in [9] and was able to keep the child directly under the overhead camera.

4.5.1 Multiple kernel learning

The work detailed here was first published in [69]. From the overhead video footage of our robot interaction experiments, we take frames from every video using the child's location to cut out a frame of the child's full body, and, manually, cropping the child's head in a closer frame. Every image is up- or down- sampled to size 80 by 80 pixels; samples of body and head images are shown in Figure 4.7. Each image was manually assigned an orientation: cardinal directions of north (straight ahead), east (to the right), south (towards the bottom of the photo) west (to the left), and intercardinal directions of northeast, southeast, southwest, and northwest. The first dataset we constructed contained over 300 samples each of both cardinal and intercardinal directions of overhead head photos, and over 200 samples each of cardinal and intercardinal directions of bodies. This set of $2400+$ head samples and $1600+$ body samples forms our training set. Some of the direction samples were copied and rotated and labeled as a new direction; to control for over learning on this augmented dataset, we sampled another two participant videos and labeled both body and head directions to make over 50 samples of every direction for testing. These data were not augmented by any additional rotations, so each of these $400+$ images are unique. This data forms the test set.

Dalal and Triggs used the histogram of oriented gradients (HOG) on a large dataset of photos of pedestrians in [24], and HOG features have been successfully used in other identification tasks, as in [61]. Our first approach to orientation detection was to take the HOG features of every photo in our dataset, using the testing and training images just detailed. Examples of training images and their HOG features are shown in Figure 4.7.

Multiple kernel learning (MKL) [14] is able to combine features at different levels in a well founded way that learns to incorporate a predefined set of SVM kernels automatically.



(a) Body images and HOG features

(b) Head images and HOG features

Figure 4.7: Examples of (a) full-body images and (b) head images of subjects facing each direction, and visualization of their HOG features.

It aims at removing assumptions of kernel functions and eliminating the burdensome manual parameter tuning in the kernel functions of support vector machines, or SVM's. Formally, it defines a convex combination of m kernels. The output function is formulated in Eq. (4.1):

$$s(\boldsymbol{x}) = \sum_{k=1}^{m} \left[\beta_k \left\langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}) \right\rangle + b_k\right] \tag{4.1}$$

where $\Phi_k(\boldsymbol{x})$ maps the feature data $\boldsymbol{x}$ using one of m predefined kernels, with an L1 sparsity constraint. The goal is to learn the mixing coefficients $\boldsymbol{\beta} = (\beta_k)$, along with $\boldsymbol{w} = (\boldsymbol{w}_k)$, $\boldsymbol{b} = (b_k)$, $k = 1, \dots, m$. The resulting optimization problem becomes:

$$\min_{\boldsymbol{\beta},\boldsymbol{w},\boldsymbol{b},\xi} \frac{1}{2}\Omega(\boldsymbol{\beta}) + C\sum_{i=1}^{N} \xi_i, \text{ s.t. } \forall i : \xi_i = l\left(s(\boldsymbol{x}^{(i)}), y^{(i)}\right) \tag{4.2}$$

where $(\boldsymbol{x}^{(i)}, y^{(i)}), i = 1, \dots, N$ are the training data and N is the size of the training set. Specifically, $\boldsymbol{x}^{(i)}$ is a HOG feature vector, with its corresponding training label $y^{(i)} = 1$ for a positive sample and $y^{(i)} = -1$ otherwise.

In Eq. (4.2), C is the regularization parameter and l is a convex loss function, and $\Omega(\boldsymbol{\beta})$ is an L1 regularization parameter to encourage a sparse $\boldsymbol{\beta}$, so that a small number of kernel functions are selected. This problem can be solved by iteratively optimizing $\boldsymbol{\beta}$ with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through linear programming, and optimizing $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed $\boldsymbol{\beta}$ through a generic SVM solver. These equations depict the standard binary classifier. In this work, they are extended to address the multiclass classification problem by one-against-all implementation of binary classifiers.

We used MKL to classify orientations on the body images and, separately, the head images. We trained three models on each image set – a linear kernel SVM, a radial basis function (RBF) kernel SVM, and a MKL model. For the single kernel approaches, the

Table 4.1: A quantitative comparison of the model performances.

Images	Kernel(s)	Accuracy	F1-score
body	linear	0.207	0.151
body	Gaussian	0.241	0.216
body	MKL	0.296	0.275
head	linear	0.453	0.453
head	Gaussian	0.485	0.483
head	MKL	**0.515**	**0.508**

regularization parameter C was optimized using a 3-fold cross validation. We selected a fixed Gaussian kernel (σ=0.5) for the RBF models. The MKL approach automatically selected kernels from a list of Gaussian RBF kernels (σ = 0.5, 1, 2, 5, 7, 10, 12, 15, 17, 20) and polynomial kernels (degree = 1, 2, 3). Its regularization parameter C was 1.

The classification performances were measured with accuracy and F_1 score. Accuracy is defined in terms of true positives, TP (positive examples labeled correctly), false positives, FP (negative examples incorrectly labeled as positives), true negatives, TN (negative examples labeled correctly) and false negatives, FN (positive examples incorrectly labeled as negatives). Accuracy is determined by Eq. (4.3):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

The F1 score combines $precision$ $(TP/(TP + FP)$ or the fraction of elements labeled with a label that is correct) and $recall$ $(TP/(TP + FN)$ or the fraction of elements with the label that were actually found) into one number, given in Eq. (4.4).

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.4}$$

Table 4.1 compares the performance of these models.

In general, the single overhead images are difficult to classify because of the lack of facial features. The MKL approach outperforms single kernel SVM's, by automatically selecting the best kernels. Particularly, head images are better than full body images for the task of orientation classification. The performance gap between these two image sets is over 20% accuracy. As shown in Figure 4.7, in most of the images, children are sitting on the ground, with a variety of body poses. Therefore, including the body region may introduce more noise than useful features.

Table 4.2 shows the confusion matrix of the classification. It can be seen that misclassification often happens between opposite directions or adjacent directions. Figure 4.8 presents success and failure examples of the test images, which include correct classifications (in blue boxes), misclassified adjacent directions (in yellow boxes), and other misclassified examples (in red boxes). Note that the test images are of different subjects from the training images. The model may fail when the test sample has a different hair color or style that is not seen in the training set.

Given that either version of the person tracker locates entire bodies of people and robots, using the best MKL model would further require an automated head detection on top of the person detection in video footage. Add to that, a 50% accuracy is simply not accurate enough to give us reliable, useful data regarding where children are looking. To that end, we took a second, more thorough approach to orientation detection, and turned again to convolutional neural networks.

4.5.2 Deep learning

Because HOG was easily confused by headshapes that are really $180°$ different, and sometimes had misclassifications for unknown reasons, we trained several convolutional nueral networks to identify orientation. We skipped trying to identify orientation simply on heads, training instead on full body images, so that identified persons from either version of the

Table 4.2: Confusion matrix of the best performing model (MKL on head images).

	E	NE	N	NW	W	SW	S	SE
E	24	13	0	0	7	4	0	3
NE	3	39	0	0	0	7	2	0
N	0	2	28	8	0	2	5	8
NW	4	1	2	28	7	0	2	10
W	5	7	3	4	19	24	1	1
SW	2	14	7	1	1	27	2	0
S	2	2	17	2	0	2	24	6
SE	3	0	0	10	2	0	5	37

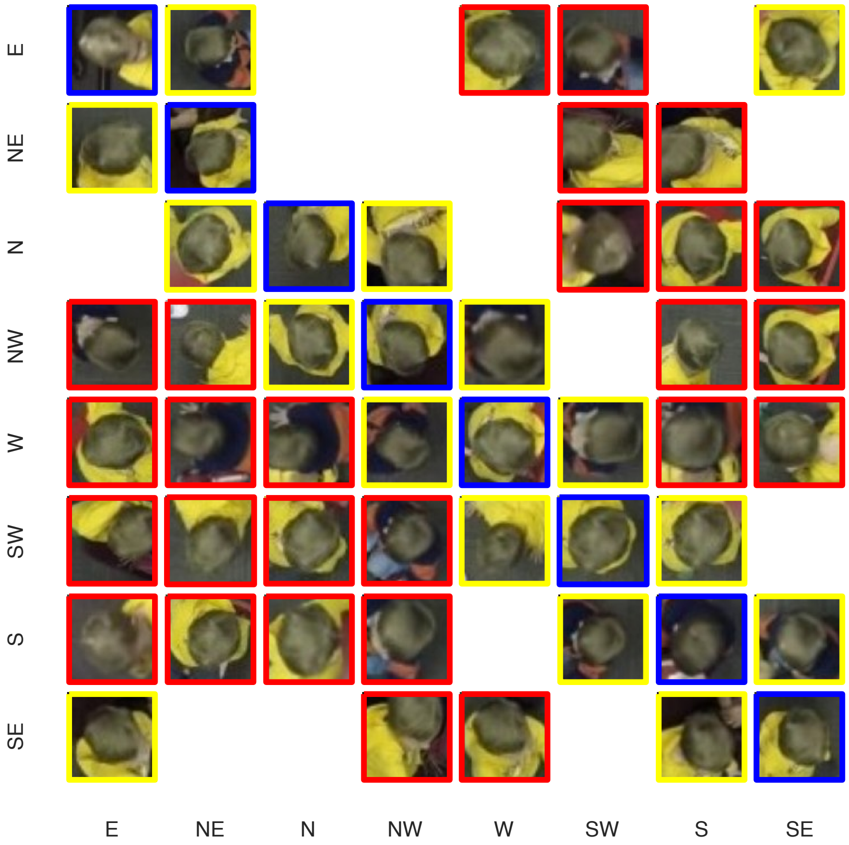


Figure 4.8: Qualitative visual representation of correctly and incorrectly labeled images. Blue boxes are correct classifications, yellow boxes are misclassified adjacent directions, and red boxes are the other misclassified examples.

Table 4.3: Best set of CNN's tested for orientation.

Dataset	Augmentations?	Epochs	Accuracy?
168	No	4,000	80.13
168	No	50,000	87.57
168	No	100,000	86.9
168	Yes	4,000	57.07
325	No	4,000	66.92
325	No	50,000	86.19
325	No	100,000	86.76
325	Yes	4,000	48.38
168 + 950	No	4,000	60.14
168 + 950	No	50,000	76.71
168 + 950	No	100,000	81.88
168 + 950	Yes	4,000	53.89
325 + 950	No	4,000	56.97
325 + 950	No	50,000	72.94
325 + 950	No	100,000	76.95
325 + 950	Yes	4,000	50.95

person trackers could be fed directly into the orientation classifier without attempting to identify heads within a person image and doubling the amount of detection actually necessary (though we also did that, with results no better than the networks described hereafter, so these details are omitted).

We first annotated a larger dataset, expanding the video sources. The first set of images sampled from the human-robot interaction videos discussed in Section 4.4; a sample is shown in Figure 4.9. The second and third set of images sampled from different sessions of an in-group bias experiment, detailed in Chapter 6, which both had very different backgrounds from the experiments detailed in this thesis and each other; these images are shown in Figure 4.11 and Figure 4.10. By sampling different overhead videos, we sought to make the network more robust. These images came directly from bounding boxes generated from the deep learning person / robot detector; the bounding boxes are very tight, and the overall images are quite good. This dataset contains $2,600$ samples total of the 8 cardinal and

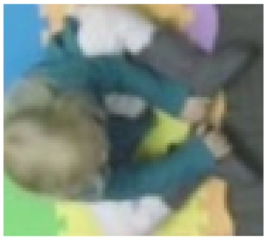


Figure 4.9: Image sample type one, from videos generated in this thesis.

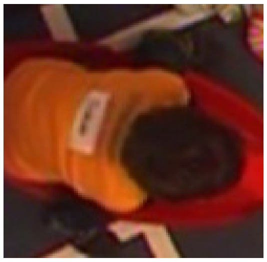


Figure 4.10: Image sample type two, from an in-group bias experiment.

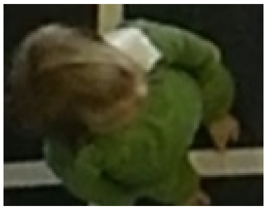


Figure 4.11: Image sample type three, from an in-group bias experiment.

intercardinal directions (north, east, south, west, northeast, southeast, southwest, northwest).

Unfortunately, not all directions had an equal number of original samples, with intercardinal directions (northeast, northwest, southeast, and southwest) having the fewest. We thus retrained networks with a few different bases. The first set of networks used the smallest number of original samples to be had for any direction that could be formed with a multiple of a 90° turn (168 images), ignoring any samples in excess of that. The second used the smallest even number possible (325 images), rotating any excess samples by whatever degree required; for example, rotating an image labeled west by 45° to form a northwest sample. This results in some added blackout due to the rotation; see Figure 4.12 for an image that might be used to form an cardinal direction, and Figure 4.13 for a cardinal direction rotated to form a northeast label. The third base uses the 168 or 325 images, plus 950 samples of bodies which came from the earlier tracker, which tracked square regions of interest. Though this was fine for generating location coordinates, for tight person bounding boxes, it tends to include large sections of floor, especially on caregivers who are off to the side of the room and therefore camera. These images, of poorer quality, but which included the (manually cropped) body images used in MKL training, were also added to the 168 or 325 datasets.

We trained several networks, the best of which are described here. Because a common and fruitful approach is to retrain the last few layers of a convolutional neural network

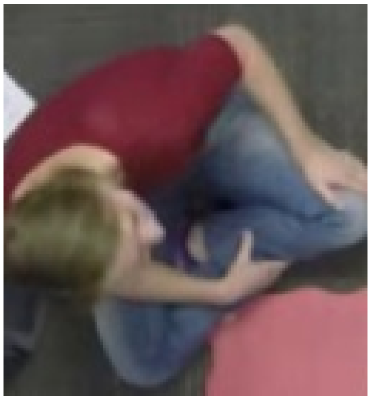


Figure 4.12: A sample image that can form any cardinal direction.

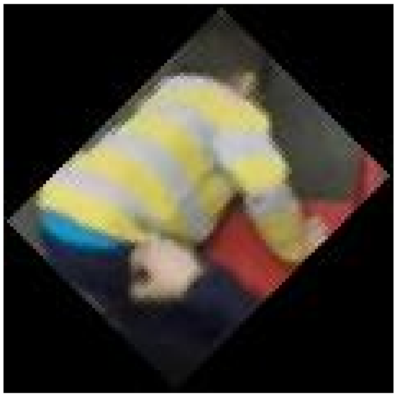


Figure 4.13: A cardinal direction rotated to bolster an intercardinal direction.

(CNN) that was already trained on a very robust dataset for many epochs, we spent minimal time crafting CNN's from scratch, and they did not perform better than the results detailed here, and details are omitted. For the following orientation-focused networks, we retrained weights from the publicly available Inception-v3 weights. Inception-v3 is trained on the ImageNet Large Visual Recognition Challenge, from 2012 data, which contains 32x32 pixel images of 1000 different classes such as zebras, dishwashers, and motorized scooters. We followed guidance from a Google Codelabs [17] to retrain the network.

We retrained networks using just the above datasets of 168 or 315 sample per orientation, as well as each of those combined with the dataset of 950 samples of bodies which include the body images used in MKL training. Retraining the inception-v3 network for $4,000$, $50,000$, and $100,000$ epochs, both with augmentations (including up to 20% random edge cropping, up to 10% random scaling, and up to 10% random brightness change) and without augmentation, we then tested the network on the original samples of images. Table 4.4 shows the confusion matrix of the network trained on a smaller but more pure number of samples, with an accuracy of 86.90%. Table 4.5 shows the confusion matrix of the network trained on a larger, bolstered network, with a similar accuracy of 86.76%.

In practice, however, this network worked poorly. Figure 4.14 shows the 325-trained

Table 4.4: Confusion matrix for orientation with only $90°$ rotations of images to form the training and testing set, at 168 samples per direction, with an accuracy of 86.90%.

	N	E	S	W	NE	NW	SE	SW
N	146	2	2	3	8	4	1	2
E	3	150	5	3	2	3	2	0
S	2	2	160	0	1	1	1	1
W	3	5	8	139	4	3	2	4
NE	5	0	3	1	151	3	2	3
NW	6	4	0	5	9	137	4	3
SE	2	10	3	2	2	3	140	6
SW	2	2	5	4	6	0	4	145

Table 4.5: Confusion matrix for orientation with a bolstered training set, in which any image could be rotated to make a new direction which might generate images with black patches (see Figure 4.13) at 325 samples per direction. The network's accuracy on the training set is 86.76%.

	N	E	S	W	NE	NW	SE	SW
N	284	5	11	7	7	6	4	1
E	5	293	7	5	2	2	9	2
S	7	3	281	14	4	1	10	5
W	11	10	5	274	7	7	4	7
NE	5	7	1	2	271	15	10	14
NW	8	3	3	7	9	287	5	3
SE	5	7	9	1	10	4	282	7
SW	1	2	4	4	11	8	11	284

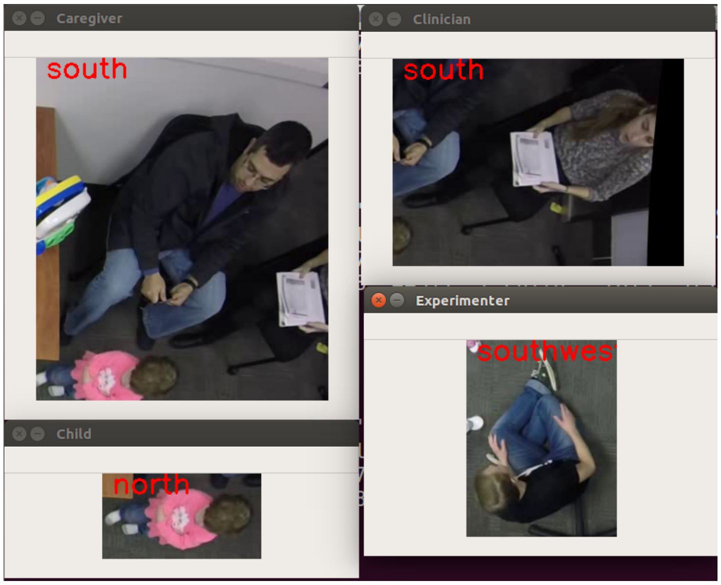


Figure 4.14: A deep learning orientation categorizer, retrained for $100,000$ epochs, applied to bounding boxes in a sample video. Some people's orientation is off by an adjacent direction, but some are completely wrong.

network applied to person bounding boxes detected from the deep-learning tracker.

4.6 Data analysis and results

4.6.1 Raw proxemics data

The data shown here stems from the raw coordinates of each actor in the room in every frame. First, we find the Euclidean distance between actors, giving us three channels of data: the distances between the child and robot, the child and caregiver, and the child and experimenter. This represents the raw distance between actors over time throughout the experiment; a sample participant's data is shown in Figure 4.15. The distance between the child and caregiver, shown in a solid red line, starts at 0 feet. Around minute 4 of the interaction, the child began to move away from the caregiver, shown in Figure 4.16, and away from the caregiver, robot, and experimenter around minute 6 of the interaction, shown in Figure 4.17.

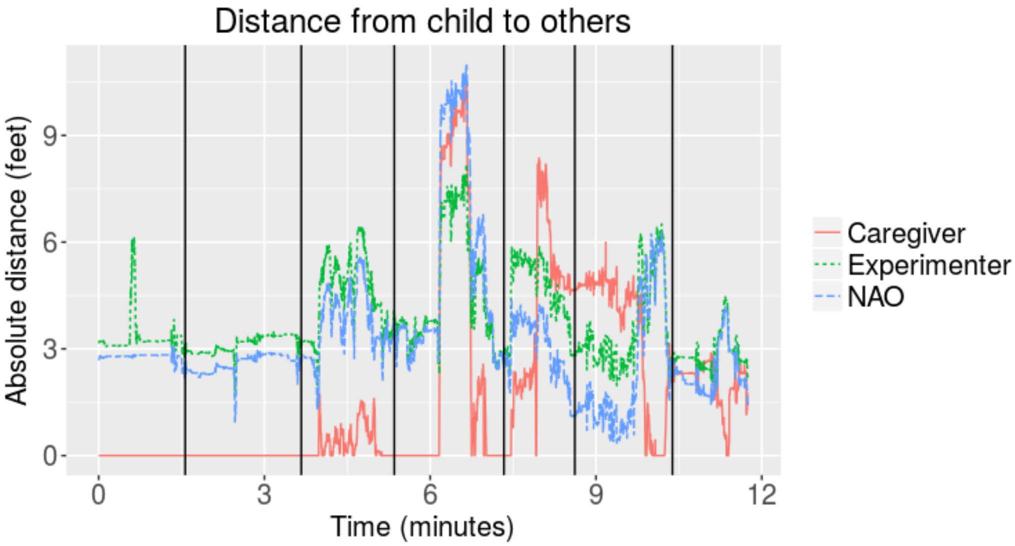


Figure 4.15: The Euclidean distances between the child and parent, child and robot, and child and experimenter (smoothed by averaging over every second). Vertical bars indicate the beginning of a press for social interaction.

In some experiments, two caregivers were present during the interaction; the minimum distance between the child and either parent was used in our data, ensuring that we can reasonably compare children with one or two caregivers present. The interactions, which are recorded at 33 frames/second, were reduced by averaging the Euclidean distances in one second windows to smooth the data slightly; we used this averaged dataset for analysis.

With distance information between all actors, we can graph the individual distances as well as the averages for all children. Figure 4.18 shows the distances from the children to other actors for the first 12 minutes of the interactions (note that many interactions finished in 9.5 minutes, and one was nearly 15 minutes). The purple lines indicated distance from child to caregiver, blue indicates distance from child to experimenter, and orange indicates distance from child to NAO. The transparent lines are individual participants; some children were on or touching their caregiver for some or all the interaction, but others were fine being across the room from their parents. A -1 on the graph indicates there is no data –

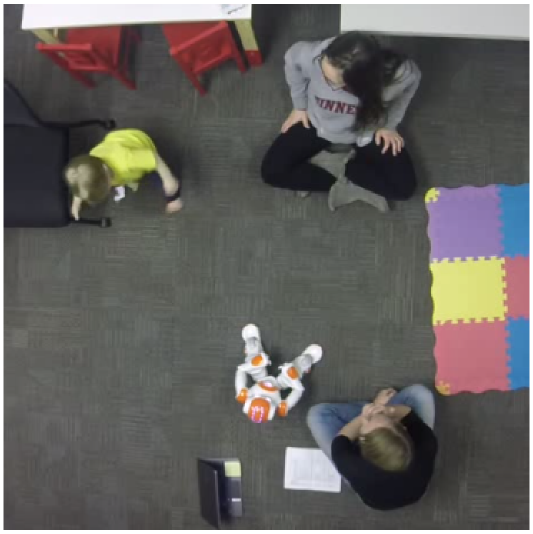


Figure 4.16: The child moves away from caregiver around minute 4.

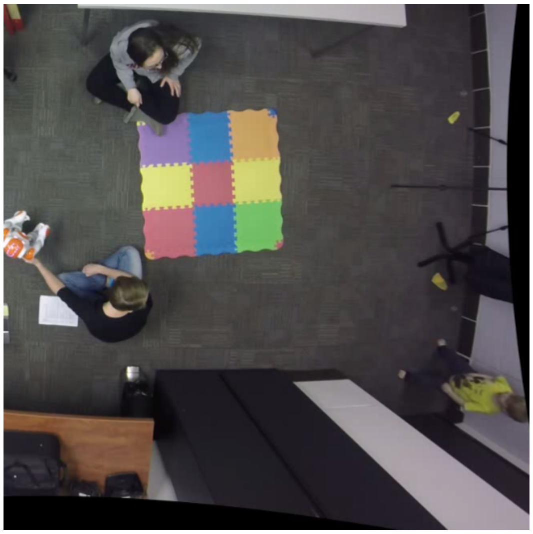


Figure 4.17: The child moves across the room around minute 6.

e.g., the caregiver left the room briefly, or the child left the room to take a bathroom break or went briefly into a connected room. The thicker lines are the average distance over all children. Figure 4.19 shows the proximity differences between the child's distance to NAO and the child's distance to the experimenter. Because the experimenter and NAO were seated very close together, the average difference is close to zero. The variation (of 2 or 3 feet) results from some children sitting on or next to the experimenter while watching the robot, and other children being comfortable sitting in front of the robot while the experimenter is behind the robot.

Figure 4.20 shows the proximity differences between the child's distance to NAO and the child's distance to their caregiver. On average, the children were closer to the NAO than a parent; the variation is much greater because the parent may be on the other side of the room when the child is comfortable interacting with the robot, or because the child may want to be close to their parent and happened to be (or intentionally was) far from the robot. This generally occurred when the child desired comfort or sought something their parent, e.g. a snack, during the interaction.

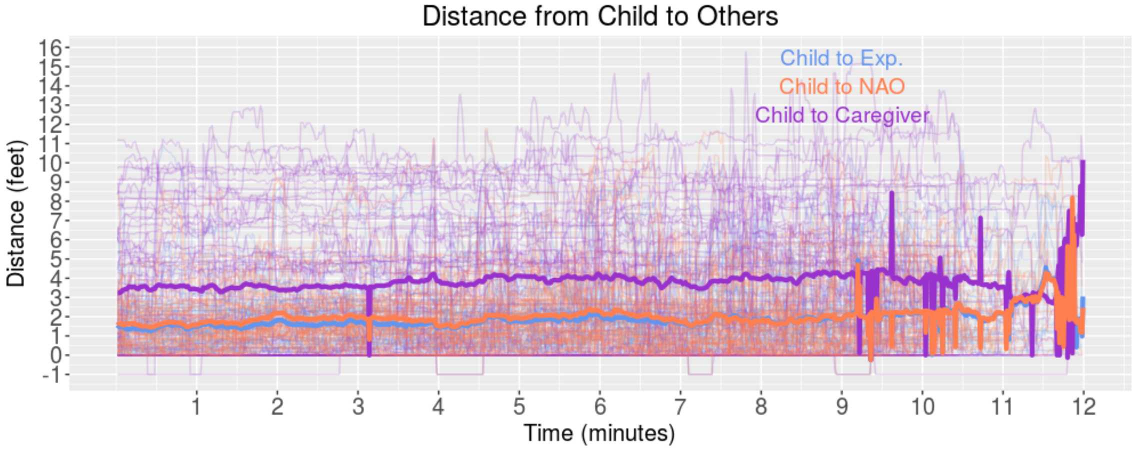


Figure 4.18: Distances from child to other actors for the first 12 minutes of the interaction. A -1 indicates the actor was not present in the room. Averages are thicker lines; individual distance graphs are transparent lines.

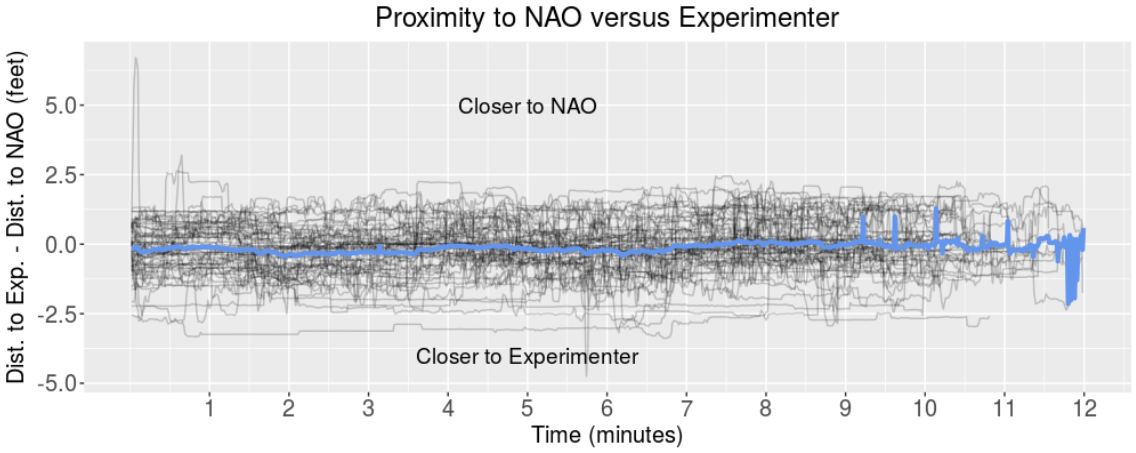


Figure 4.19: Proximity differences between the child to NAO and the child to experimenter for the first 12 minutes of the interaction; a positive number means the child was closer to NAO and a negative number means the child was closer to the experimenter.

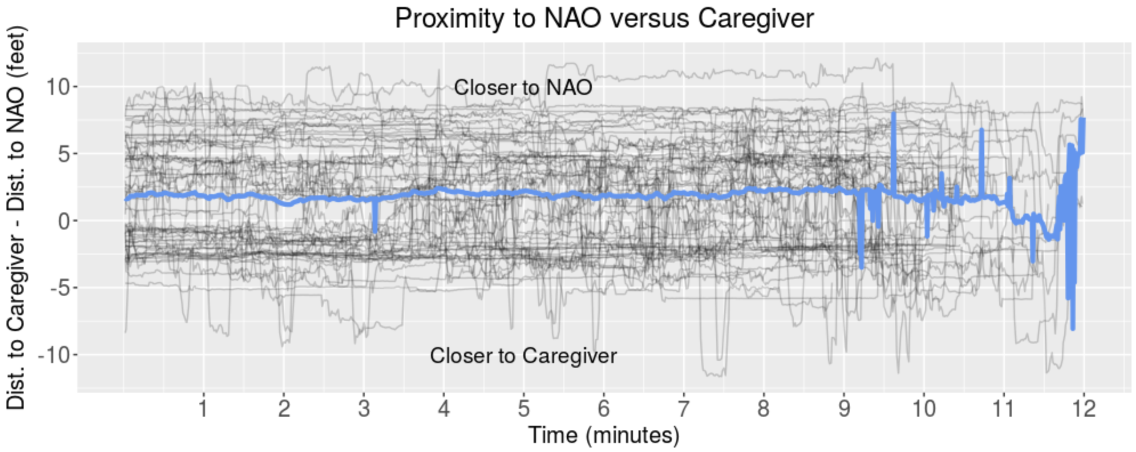


Figure 4.20: Proximity differences between the child to NAO and the child to parent for the first 12 minutes of the interaction; a positive number means the child was closer to the NAO and a negative number means the child was closer to the caregiver.

4.6.2 Data normalization

Our first data reduction method reduces the three time series per child into one, essentially normalizing the data.

Figure 4.21 shows two ways of doing this. The first reduction considers the distance to the robot, denoted N, and the distance to the caregiver, denoted CG, in Eq. (4.5).

$$N/(N+CG) \tag{4.5}$$

The second reduction also considers the distance to the experimenter, denoted E, in Eq. (4.6), so as to retain data on some cases in which participants sought the experimenter's company while interacting with the robot and not the caregiver.

$$N/(N+CG+E) \tag{4.6}$$

In either case, we have reduced multiple series of variable magnitude to a single, unit-less number scaled $[0, 1]$, where a 1 value means closest to caregiver and a 0 value means

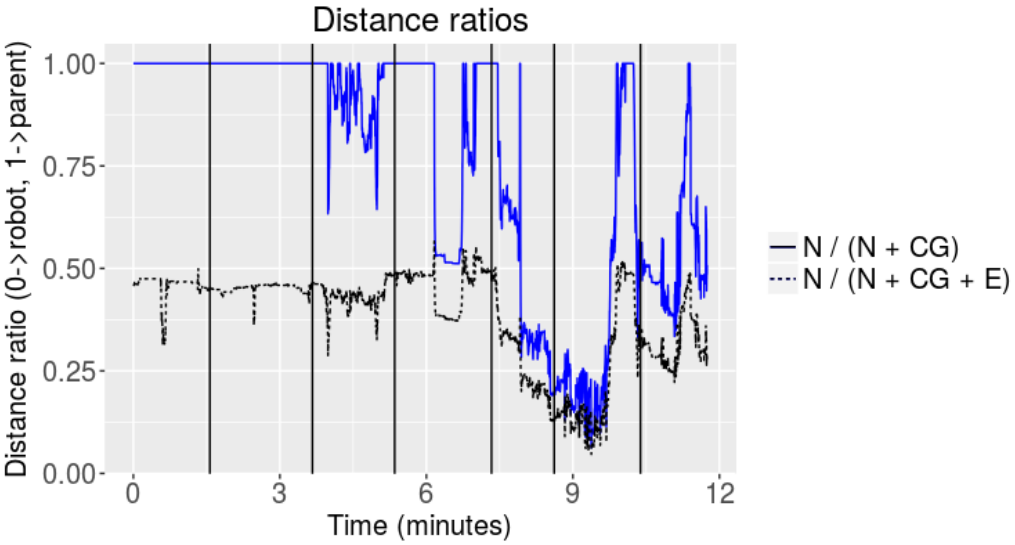


Figure 4.21: Condensing three time series to a single metric, with and without considering the distance to the experimenter (smoothed by averaging over every second). Vertical bars indicate the beginning of a press.

closest to the robot, regardless of the absolute distance between child and robot or child and caregiver.

Recall that each interaction varies in length due to potential buffer time between presses in a single interaction. The buffer time, lasting up to one minute between presses as needed, was included in the last press that occurred. For example, if Child A needed a 40 second break after a one minute press, but Child B didn't need a break and only used two seconds after a one minute press, the press lasted for 100 seconds for Child A but only 62 seconds for Child B. This flexibility in interaction time naturally raises the question of how to compare these variable length data. The time difference between presses is capped at 60 seconds, and any time between presses is used to draw the participant's attention back to the robot. By and large, the excess time was spent by the participant getting a snack or toy, playing with other items in the room, or talking to their caregiver; none of the buffer time was spent interacting with the robot, and the robot did not move autonomously during these breaks. Therefore, we consider the time between presses to be noise.

To remove this noise, we choose a simple method of aligning all presses in all interactions, and we truncate each press to the length of the shortest occurrence of that press over all participants. For example, say Child A took 60 seconds during Press 1 with a 30 second break, then 180 seconds during Press 2 with a 10 second break. Say Child B took 60 seconds during Press 1 with a 2 second break, then 180 seconds during Press 2 with a 2 second break. It should be noted that in most cases the experimenter manually starts the next press for attention with the robot, so short breaks of 1-3 seconds is simply the time taken to reach over and push buttons on the robot (or occasionally, to first re-orient the robot towards the child if they shifted position). We only want to compare participant reactions to the robot while the robot is actively moving or speaking; thus, Press 1 is truncated to 62 seconds for both participants and Press 2 is truncated to 182 seconds for both participants. Alternatively, we could have timed the presses off-line and truncated participant data with those timings, but in practice, some participants progressed through the interactions in immediate succession and this off-line timing was unnecessary.

The effect of such data loss, i.e. 28 seconds after Press 1 and 8 seconds after Press 2 in the above example, admittedly contains some distance data. Either the participant didn't move between presses or approached the robot again from somewhere else in the room. In theory, a participant might have outlasted the one minute buffer time and started the next press further from the robot than they were at the (truncated) end of the previous press. However, in practice, this did not happen; only the first two cases occurred. Thus, minimal interaction information was lost due to this truncation. These snippets of non-robot interactions could last up to a full six minutes, if some participant was willing to continue the interaction but outlasted the 60 second buffer every time; in practice, no participant used the entire buffer time between every interaction, and only 2-3 participants ever outlasted the one minute buffer for a press.

4.6.3 Developmental assessment scores

We mentioned in Section 4.1 that each participant takes multiple standardized developmental assessments. There are several specific dimensions we are interested in, each of which varies in people with autism. In the social dimension, we look at socialization behaviors, communication skills, adaptive behaviors, and early learning composite scores. Each score is standardized such that 100 is average for the population. In the physical dimension, we look at ritualized and restricted behaviors, and composite endorsed score of such behaviors. The ritual, restricted, and composite scores come from a parent-report, the Repetitive Behavior Scale- Early Childhood (RBS-EC). The communication and socialization scores come from the Vineland Adaptive Behavior Scales, and the early learning composite score comes from the Mullen Scales of Early Learning (that score combines the five areas of cognitive functioning into a composite score). The other socialization measure, noted SRS/vrRSB in this text, comes from the parent report assessments Social Responsiveness Scale- Second Edition (SRS-2) or the video-referenced rating of Reciprocal Social Behavior (vrRSB), depending on the age of the participant. For this experiment, because these participants form a robot-interaction baseline, children with notable differences from each other in all such scores are desirable, and make for a more robust comparison to atypically developing children. The scores are summarized in Table 4.6 and Table 4.7, as well as graphed in Figure 4.22, Figure 4.24, and Figure 4.23. For more details about the Vineland and Mullen, the exams we use the most scores from, see Appendix B; this appendix includes sample items that children are tested on, as well as sample outputs and how to interpret them.

4.6.4 Response categorization

Thus far, we have considered both absolute and normalized distances, even in our distance metrics in Eq. (4.5) – (4.6). Consider that shorter distances between a participant and another actor probably indicate more comfort or interest, and longer distances indicate less

Table 4.6: Assessment summary for particular scores of interest for all participants (generated from Vineland, Mullen, and other social behavior assessments). Because either the vrRSB or SRS assessment is taken, based on the child's age, participants have one or the other. This data was standardized and combined to make comparisons easier.

	Age (weeks)	Restricted	Ritual	Composite Endorsed	vrRSB	SRS	vrRSB/SRS combined
Min	110	0	0	0	9	7	0.028
Mean	143.7	2.362	2.793	9.638	19.727	30.623	0.134
St. Dev	19.974	2.058	2.007	6.794	9.707	16.126	0.093
Max	196.6	8	9	26	41	82	0.578

Table 4.7: Assessment summary for particular scores of interest for all participants (generated from Vineland, Mullen, and other social behavior assessments) in socialization and communication dimensions.

	Socialization	Adaptive	Communication	Early Learning Composite
Min	82	86	81	83
Mean	108.5	107.8	109.8	115.1
St. Dev	11.903	12.026	11.205	17.150
Max	136	133	135	148

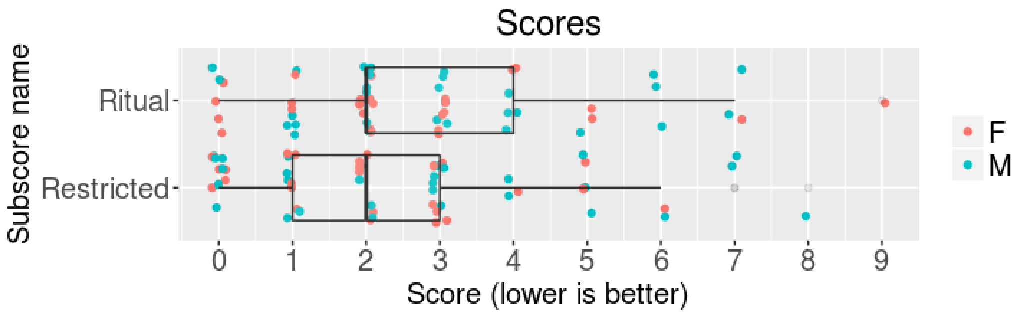


Figure 4.22: Ritual and restricted behaviors of all participants.

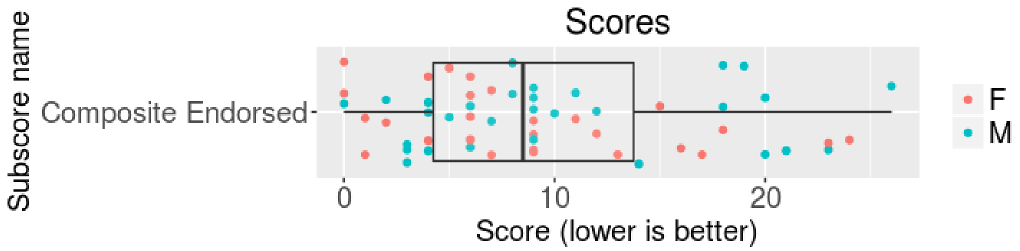


Figure 4.23: Composite endorsed score of ritualized and restricted behaviors of participants.

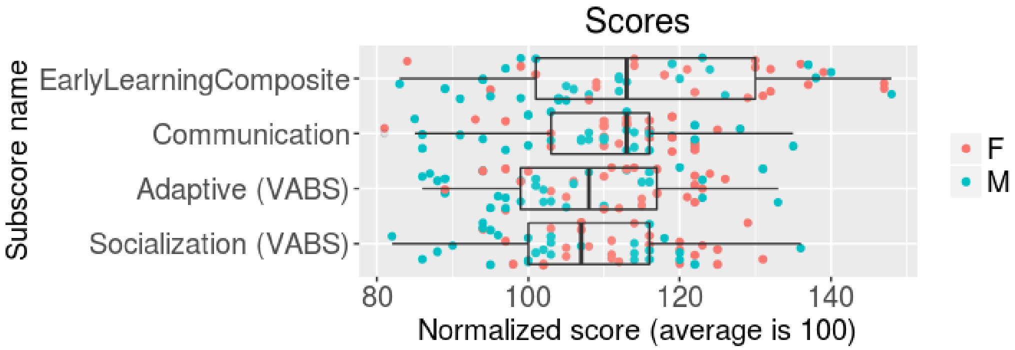


Figure 4.24: Early learning composite, communication skills, adaptive behaviors, and socialization behaviors of all participants.

comfort or interest. A participant sitting two feet away from the robot and five feet away from the caregiver probably indicates strong comfort with or interest in the robot. However, if a participant is seated on their caregiver and the caregiver is seated two feet from the robot, they probably need comfort from their caregiver while they watch the robot. Do children change their mind about the robot, or become more or less comfortable with it over time? We first graph the distance ratios of children over time, averaging the entire distance ratio per press for each child, in Figure 4.25 (best viewed in color). Examining participants one by one reveals more obvious differences: some children are always near their caregiver (12, or 20.34% of this sample), or always near the robot (22, or 37.29%), and some children move from closer to their caregiver to closer to the robot $(6$, or $10.17\%)$, all shown in Figure 4.26. There are also children that move from closer to the robot to closer to their caregiver (3, or 5.08%, shown in Figure 4.27), and some have no steady pattern (16, or 27.12%, shown in Figure 4.28).

We first check if children in these five movement pattern classes are similar to other children in the same group in terms of assessment scores. These classes by scores are shown in Figure 4.29 – Figure 4.36.

To check if participant reactions were a reasonable way to differentiate participants, we clustered participants' with the K-Nearest Neighbors algorithm over the vector of their

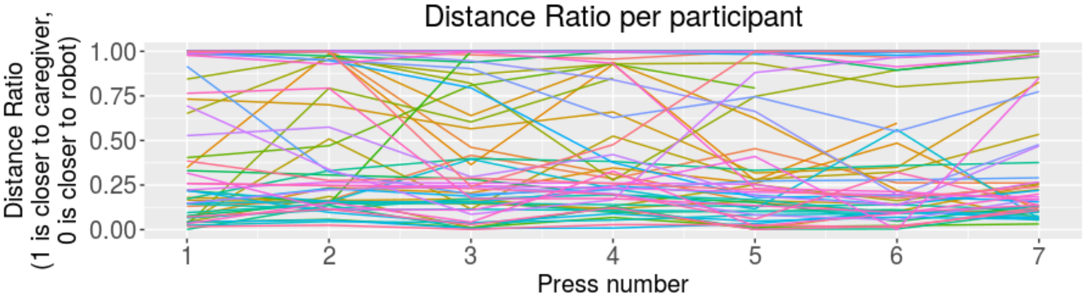


Figure 4.25: Plot of Distance Ratio per child, averaged over each press.

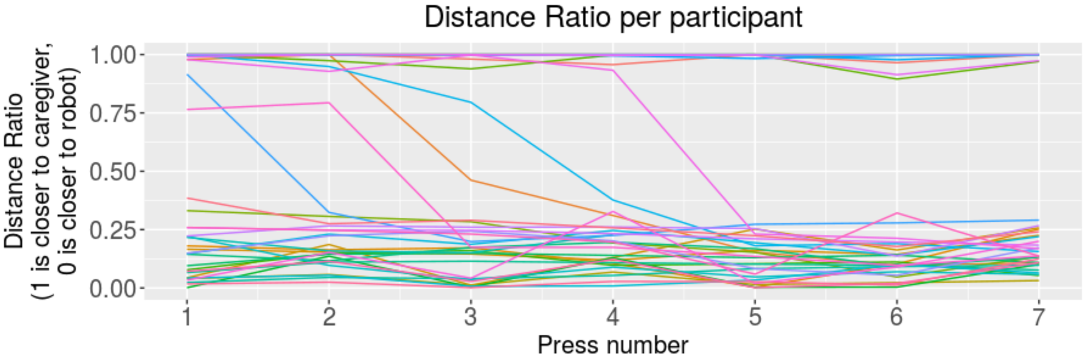


Figure 4.26: Three visible patterns of children that stay near the caregiver (12), that stay near the robot (22), and that move from caregiver to robot over time (6). The latter may also include children that stay near the robot with their caregiver, but the caregiver leaves and the child remains near the robot.

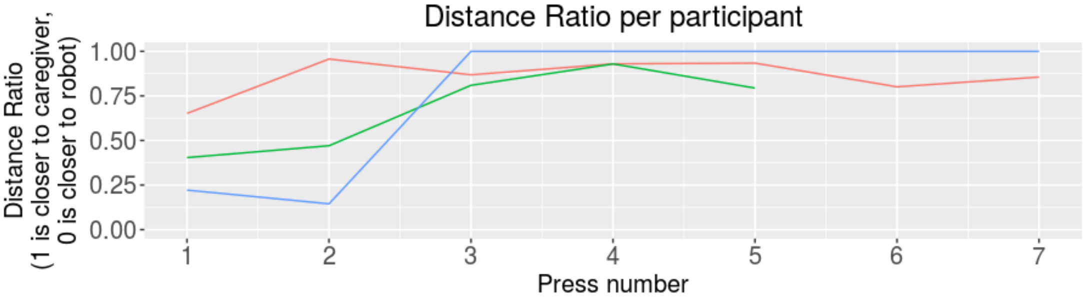


Figure 4.27: Children that initially were near the robot, but went closer to their caregiver over time (or asked their parent to come be near them as they played with the robot, 3).

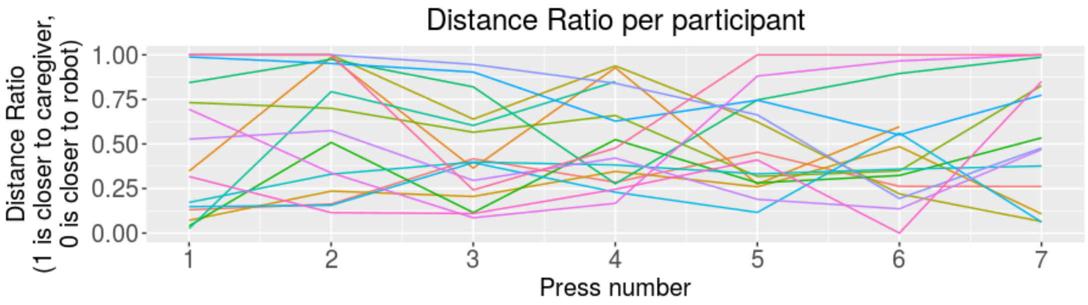


Figure 4.28: Children with no particular pattern over time (16).

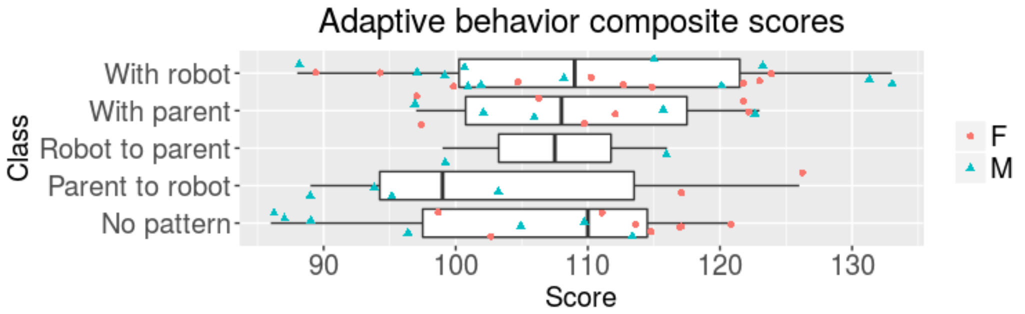


Figure 4.29: Participants' adaptive behaviors, grouped by the change in their distance ratio over time.

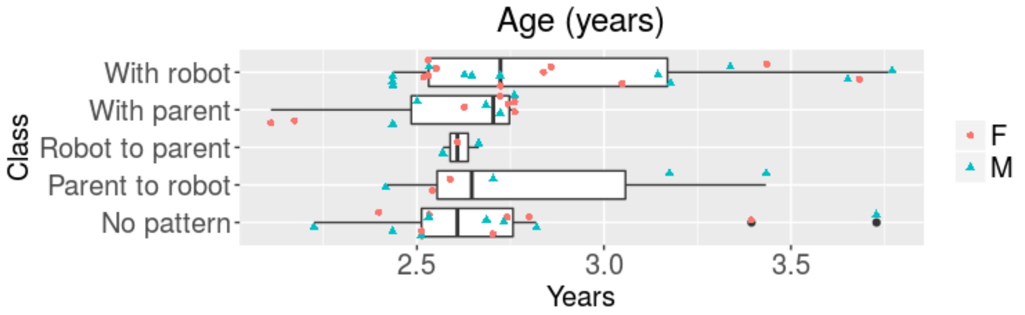


Figure 4.30: Participants' age in years, grouped by the change in their distance ratio over time.

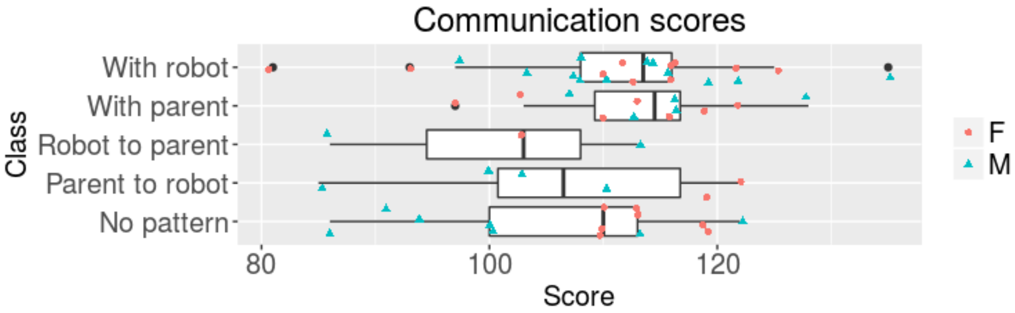


Figure 4.31: Participants' communication skills, grouped by the change in their distance ratio over time.

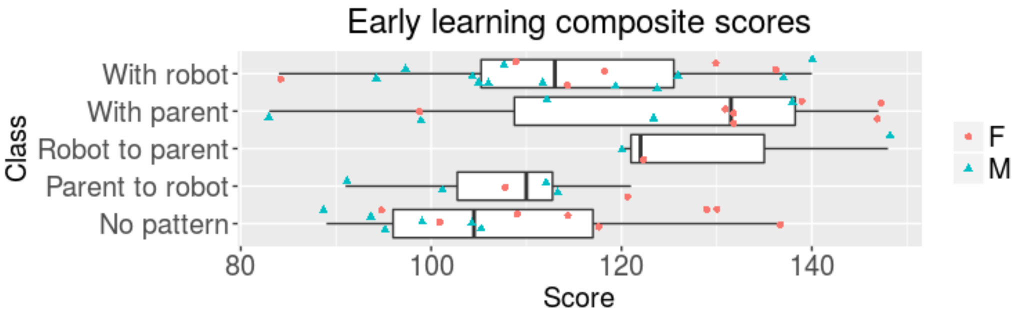


Figure 4.32: Participants' early learning composite scores, grouped by the change in their distance ratio over time.

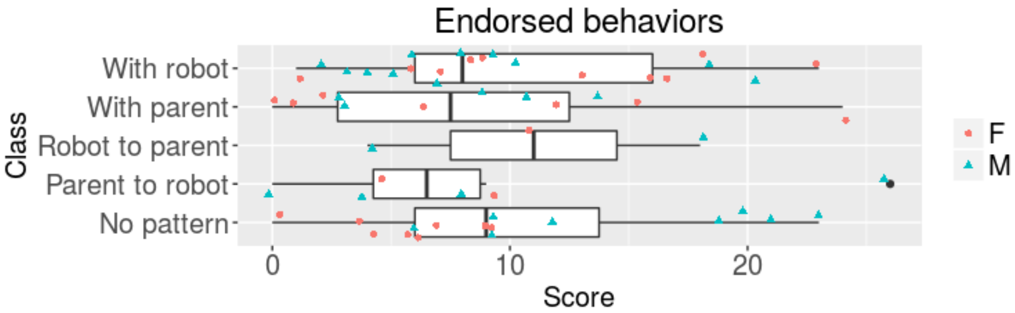


Figure 4.33: Participants' total endorsed ritual / restricted behaviors, grouped by the change in their distance ratio over time.

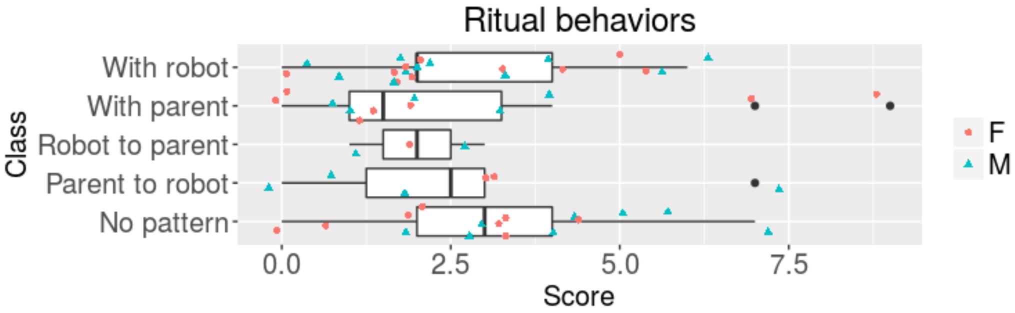


Figure 4.34: Participants' ritual behaviors, grouped by the change in their distance ratio over time.

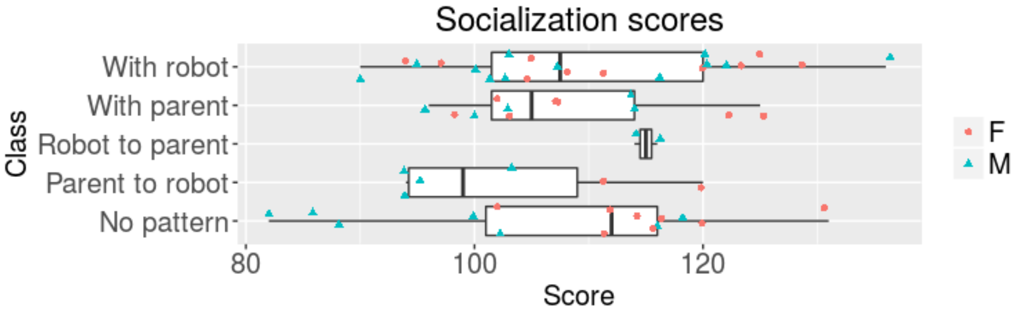


Figure 4.35: Participants' socialization skills, grouped by the change in their distance ratio over time.

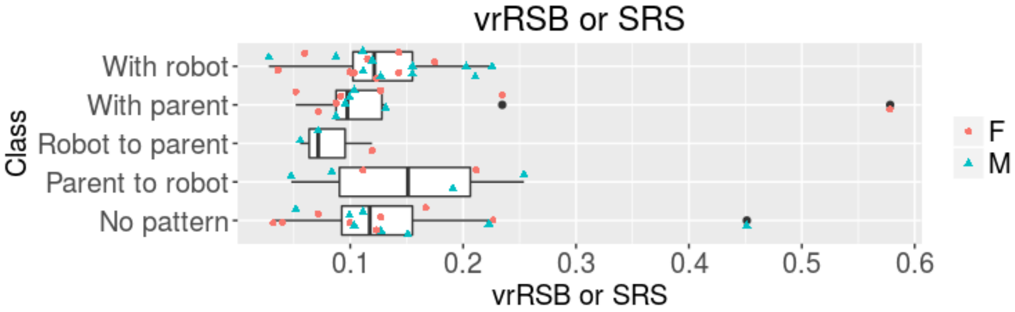


Figure 4.36: Participants' vrRSB or SRS scores (standardized and combined), grouped by the change in their distance ratio over time.

average distance ratio from Eq. (4.5) per press. Because 5 patterns (or 4 patterns and one lack-of-pattern) was visible to the human eye, we tested $K = 2$ through $K = 5$, running each number of clusters $1,000$ times. This did not yield consistent groupings for higher K's; for $K = 5$, the same groups were generated only 92 times, and for $K = 4$ the same groups were generated only slightly more, at 123 times. $K = 3$ was the most stable, at 970 consistent groupings, and $K = 2$ was slightly less so at 685.

Participants clearly vary by their behavior towards the robot, but do these behaviors ultimately relate to their development assessment scores? Strictly considering the behavior classes, we cannot say that these children have clinically significant differences. Some groups are simply too small, such as the three children who moved from robot to caregiver over time, and their scores may be an average of this behavior or it may simply be noise; one participant did not finish the interaction, so that pattern may even have changed into a 'no pattern' before the end of the interaction. We then consider combining assessment scores with the distance ratio, to see if distance ratios contribute meaningful information when looking at differences between participants.

4.6.5 Participants in clusters

Pruett and Povinelli suggest that though autism is currently considered a spectrum, clustering or grouping children based on proxemics, eye contact and joint behavior, and communication timing in dyadic interactions may differentiate typically developing and atypically developing children [81]. Though they hypothesize that those three social-communicative areas may differentiate and cluster children with attention-deficit/hyperactivity disorder, autism, and Williams syndrome, they also call out "the significance of restricted and repetitive behaviors" and state "it would be interesting to study correlations between these and our proposed key variables of dyadic interacting." Following their lead, we ask if clustering children on one physical dimension, one social dimension, and the distance ratio results in groups with

Table 4.8: Groups generated from K-Nearest Neighbors, with $K = 2$.

Group	1	2
Size	37	22
Sex	17 F, 20 M	11 F, 11 M
Age mean (yrs)	2.764	2.743
Age std. dev	0.412	0.338

statistically significant score differences. To that end, we look at the ritual behaviors scores, socialization scores, and distance ratio. Using K-Nearest Neighbor clustering, from 2 to 5 possible groups, results show that clustering with $K = 2$ groups of participants is better than larger numbers of clusters. For 1000 trials, $k = 2$ results in completely stable clustering – participants were grouped exactly the same way every time. These participants were 37 in Group 1 (20 M) and 22 in Group 1 (11 M), described in Table 4.8.

Given this grouping of participants, we want to know if the children also vary by another physical and social dimension, as well as other composite physical behaviors and early learning scores. For these validations, we use the restricted behavior score, communication score, vrRSB or SRS (depending on age of participant), and early learning composite scores.

After grouping participants, we test the differences between the two groups of children with Student's T-Test: do the groups vary by development assessment scores with clinical significance? Results are summarized in Table 4.9 and Table 4.10, as well as shown in Figure 4.37 – Figure 4.39 (all of which are statistically significant at $p < 0.001$), Figure 4.40 and Figure 4.41, (all of which are statistically significant at $p < 0.05$) The scores between the groups' vrRSB and SRS scores and the scores between the groups' early learning composite scores were not statistically significant (not shown).

4.6.6 Participants on a spectrum

Autism spectrum disorder is, by definition, a spectrum of behaviors, so we would be remiss not to explore how our typically developing participants fall on a spectrum. For this we

Table 4.9: Score differences in distance ratio (averaged over the entire interaction), ritualized behavior scores, and socialization scores between the two groups. Student's T-Test shows the scorings are different at a p-value < 0.001 for each.

	Distance Ratio		Ritual		Socialization	
Group	G1	G2	G1	G2	G1	G2
Min.	0.029	0.06513	0	0	82	94
1st Qu.	0.2501	0.12628	1	2.75	97.75	109
Median	0.654	0.21115	2	4	103	118
Mean	0.6053	0.29191	1.946	4.25	104.19	115.8
3rd Qu.	0.9707	0.33477	3	6	112.5	122.5
Max.	1	1	5	9	125	136

Table 4.10: Score differences restricted behaviors and communication scores between the two groups. Student's T-Test shows the scorings are different at a p-value < 0.05 for each.

	Restricted		Communication	
Group	G1	G2	G1	G2
Min.	0	0	81	85
1st Qu.	0	1	100	109
Median	2	3	110	113
Mean	1.946	3.1	107.6	113.2
3rd Qu.	3	5	116	120.5
Max.	8	7	128	135

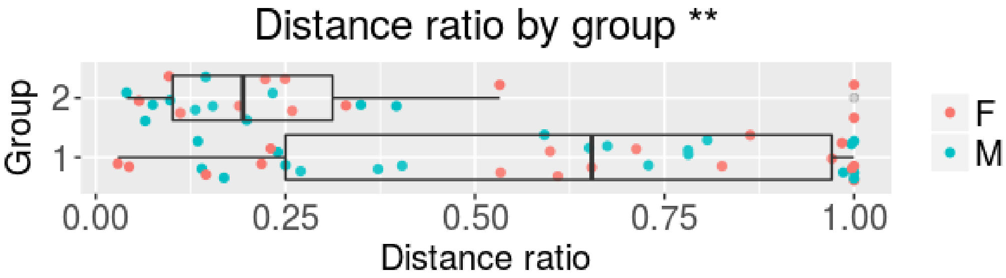


Figure 4.37: Participants' distance ratio, averaged over the entire interaction, by group; t-test indicates significant differences, at $p < 0.001$.

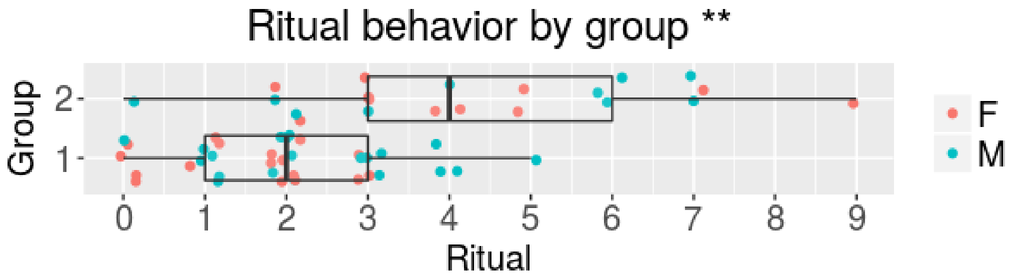


Figure 4.38: Participants' ritual behavior score, by group; t-test indicates significant differences, at $p < 0.001$.

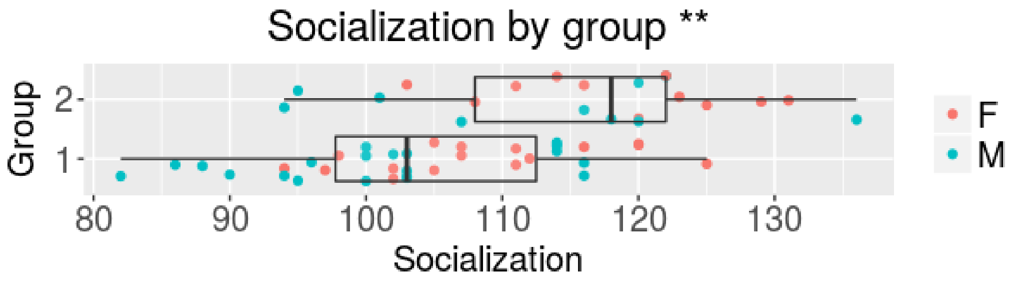


Figure 4.39: Participants' socialization skills, by group; t-test indicates significant differences, at $p < 0.001$.

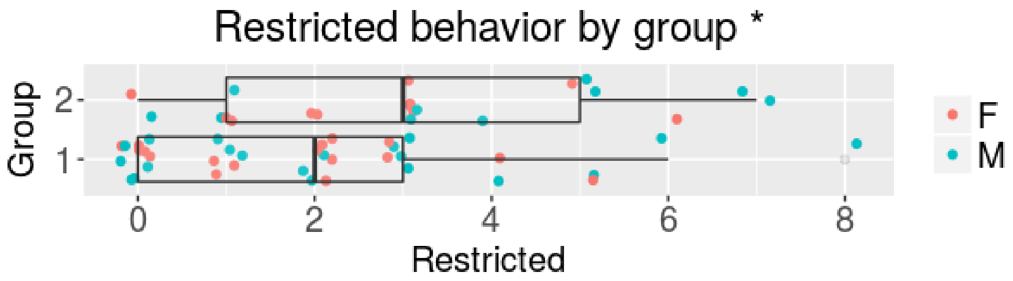


Figure 4.40: Participants' restricted behavior score, by group; t-test indicates significant differences, at $p < 0.05$.

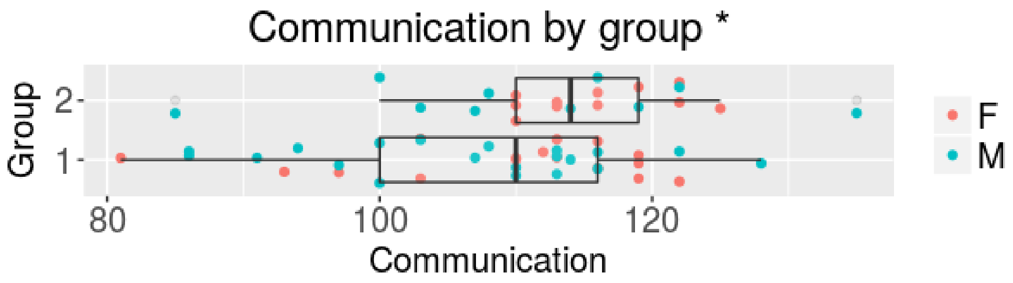


Figure 4.41: Participants' communication scores, by group t-test indicates significant differences, at $p < 0.05$.

perform Principal Component Analysis (PCA) on several different subsets of the data. In all cases, for the sake of data inclusion, a few unknown values (e.g., distance information from a child who did not finish the interaction and only has the first few of the seven presses, or a child who did not have a particular assessment score), were replaced by the participant-wide average. We then show the grouping results in Section 4.6.4 and Section 4.6.5 overlaid with the top two PCA components.

First, we perform PCA on the average Distance Ratio 1 (Eq. (4.5)) per press, shown in Figure 4.42. The top two components explain 79% and 11% of variance; a large grouping is visible near $(0, -1)$, with about $1/3$ of participants spread across the rest of the axes.

Next we perform PCA on the participants' over-all interaction average distances to NAO, Caregiver, and Experimenter, as well as average Distance Ratio 1 (Eq. (4.5)) and Distance Ratio 2 (Eq. (4.6)) in Figure 4.43. The top two components account for 68% and 21% of the variance of these averages between participants. Though the top component is smaller than the top component of PCA on all 7 presses, the progressive nature of the interaction (considering the known groups of children that for example 'stayed near NAO' or 'moved from parent to NAO' as shown in color) are much more visible. We take from this that a single metric, Distance Ratio 1, is a reasonable choice to represent the entire interaction.

Next, we ignore distances for a moment and just perform PCA on the various physical and social dimension assessment scores we discussed in Section 4.6.3, shown in Figure 4.44.

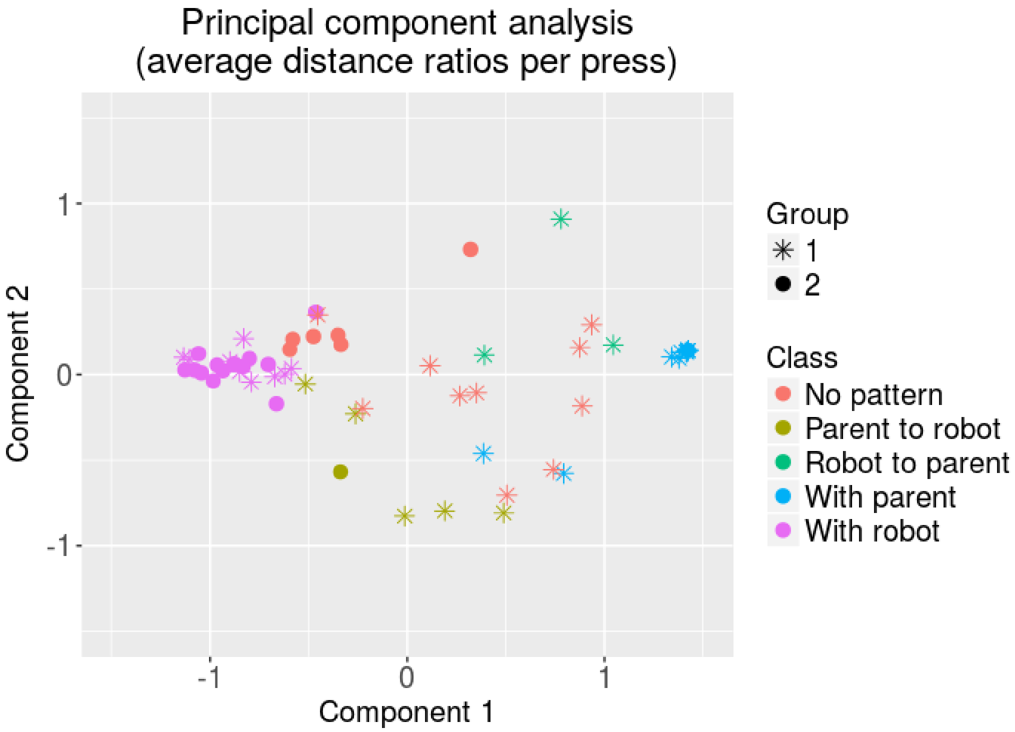


Figure 4.42: Principal component analysis on participants' average Distance Ratio 1 for each of the 7 presses over time. Group clusters from clustering on Distance Ratio 1, ritual behaviors, and socialization scores shown in 'Group' closed circles and stars, and categorization of reaction to robot shown in 'Class' colors (best viewed in color). Component 1 accounts for 79% of variance, and component 2 accounts for 11% of variance.

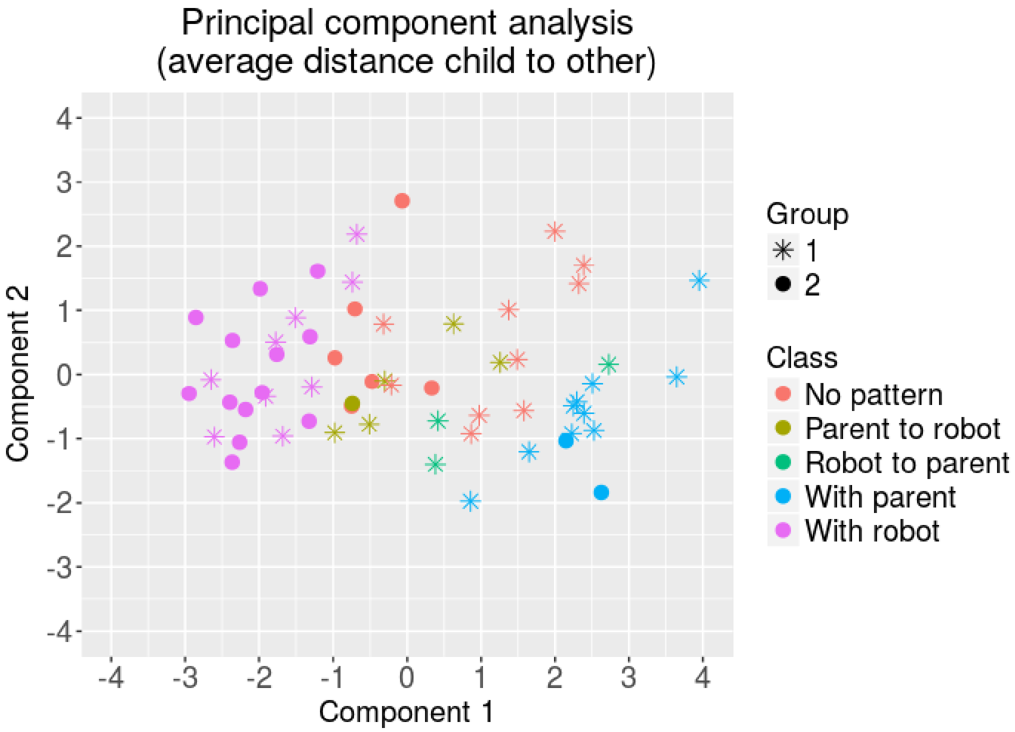


Figure 4.43: Principal component analysis on participants' average distance to NAO, caregiver, and experimenter, and average Distance Ratio 1 and average Distance Ratio 2. Group clusters from clustering on Distance Ratio 1, ritual behaviors, and socialization scores shown in 'Group' closed circles and stars, and categorization of reaction to robot shown in 'Class' colors (best viewed in color). Component 1 accounts for 68% of variance, and component 2 accounts for 21% of variance.

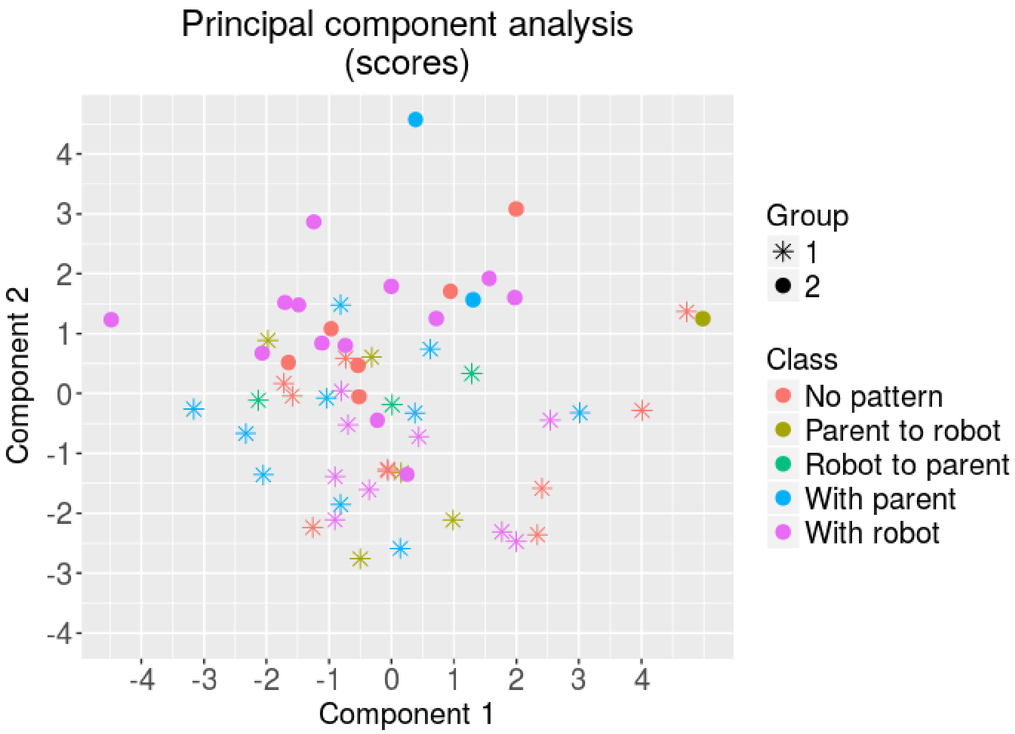


Figure 4.44: Principal component analysis on the participants' scores. Group clusters from clustering on Distance Ratio 1, ritual behaviors, and socialization scores shown in 'Group' closed circles and stars, and categorization of reaction to robot shown in 'Class' colors (best viewed in color). Component 1 accounts for 41% of variance, and component 2 accounts for 29% of variance.

There is no nice combination of the original scores that can form any small number of components to account for the variability in the participants – even the best component only accounts for 41% of the variance. This means no overt score simplification can categorize or place these participants on a spectrum. Intuitively, this makes sense – ASD is a multi-faceted disorder, and any combination of physical and social disfunctions may result in a positive diagnosis.

The last PCA analysis run combines the distance information and assessment scores, shown in Figure 4.45. Again, there is no one good component; even the top component only accounts for 44% of the variance in the data, probably driven by the number of assessment

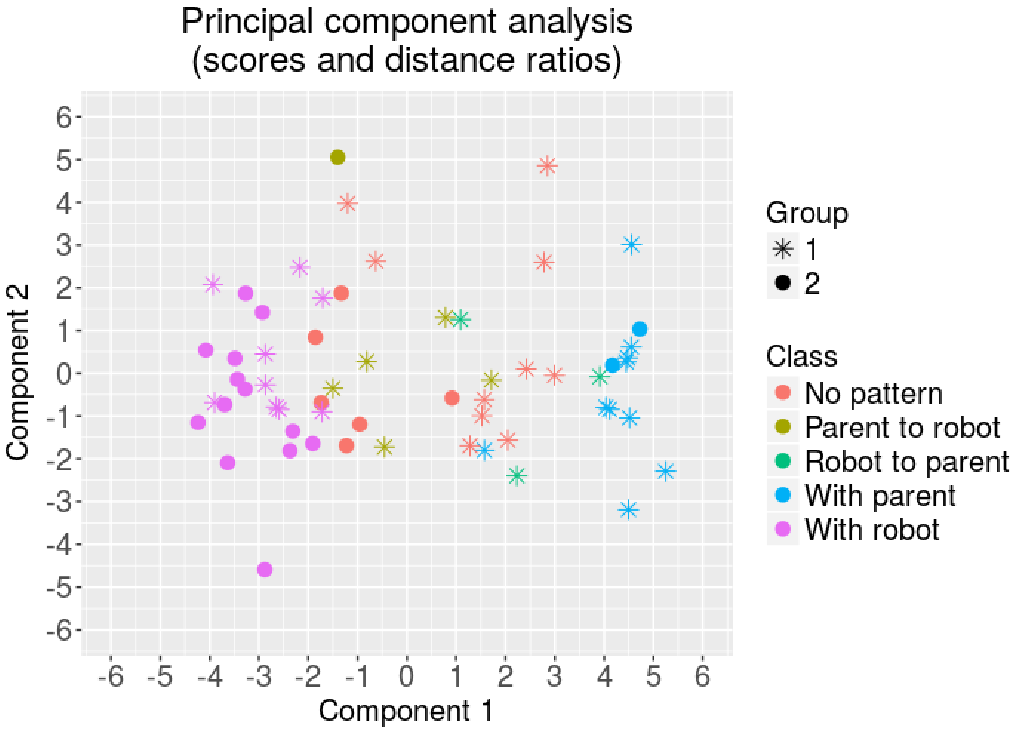


Figure 4.45: Principal component analysis on the participants' scores, average distances, and distance ratios. Group clusters from clustering on Distance Ratio 1, ritual behaviors, and socialization scores shown in 'Group' closed circles and stars, and categorization of reaction to robot shown in 'Class' colors (best viewed in color). Component 1 accounts for 44% of variance, and component 2 accounts for 17% of variance.

scores we concern ourselves with.

That PCA on scores and distance ratios, shown in Figure 4.45, also shows the pattern of the children's distance ratio classes is more interesting; this shows that distances still combine and appear as a top component (even though the component only accounted for 44% variance), and tells us that the Distance Ratio is an important feature in classifying children. Figure 4.42, which shows PCA on just Distance Ratio 1 in each press and does not consider assessment scores, shows the K-Means groups along Component 1, which accounted for 79% variance. Examining just PCA on the scores in Figure 4.44 does show a rough linear separation of the KNN groups 1 and 2. This tells us that the response to robot is an important component in the groupings, and distance adds a unique feature just like participants' assessment scores.

4.7 Summary

In Section 4.6.1, we discussed how to calculate the raw distances between child and robot, caregiver, and experimenter to form several time series. Section 4.6.2 discussed two ways to condense the absolute distances to two different metrics; Distance Ratio 1 combines distance to robot or caregiver, and Distance Ratio 2 combines distance ratio to robot, caregiver, and experimenter. Section 4.6.3 discussed the results of development assessment exams performed with each child, and showed that variable results still constitute typically developing children; this robust baseline of children should therefore represent typical reactions of a $2-4$ year old child to a social robot. We first discussed basic reactions in Section 4.6.4, which include children that are comfortable being closer to the robot the entire time, that remain close to their caregiver the whole time, that move from closer to caregiver to closer to the robot, that move from closer to robot to closer to caregiver, and those that do not have a definitive pattern.

In Section 4.6.5 we discussed how to cluster our diverse but typically developing participants into groups with statistically different scores in physical and social dimensions, namely ritual and restricted behaviors and socialization and communication scores. These children also differed in their reaction to the robot, shown through their different physically proximities to their caregiver and the novel robot. Section 4.6.6 discussed a way to simplify the different metrics through principal component analysis, showing that the dimension reduction method can not find one or two very good components that group children similarly. PCA did, however, reflect the different groups when considering just average Distance Ratio 1 per press, even though clustering did not use individual press distance data, it used the entire interaction's average ratio.

We may draw several conclusions from these data and results. Typically developing children are varied enough in their responses to a novel robot interaction to be differentiable from each other by their distances to caregiver and robot. Adding proxemics to assessment scores gives unique information that can also separate children into clinically different groups.

4.8 Lessons learned in child-robot interactions

What can we learn from child-robot experiments in this young age-group? Pilot experiments showed that 18-24 months of age was too young for a child to comfortably interact with a robot. Very young participants were willing to watch the robot, but not actually interact with it – they did not follow along with the looking or imitation games, and did not like to stand to dance with the robot. They preferred to sit with their parents while watching the robot move. They did enjoy looking at and touching the robot, and one participant enjoyed interacting with the experimenter to choose which colors the robot's eyes should become. Recall that previous research found that children with autism were more willing to interact with adults when technology was involved, even when it was clear the experimenter was the one controlling the robot [111]. Similarly, very young typically developing children were more willing to interact with a strange human when the novel robot was involved.

A very real possibility is that the sheer size of the robot interacted with age. Participants younger than two, and even some two year olds, were roughly the same height as the robot. When the robot stands for the first time, it is likely taller than a young child seated on the ground in front of it. In the first 20-30 participants in this study, proximity graphs revealed a dip in distance to parent around the 2 minute mark, which is about when the robot stood up for the first dance. This effect disappeared by the last of the experiment participants, however. Previous research includes much older children – a 5 or 7 or 11 year old is usually much taller than a NAO, so there is likely zero intimidation factor. Future research should investigate whether very young children in this age range prefer to play with a robot that is smaller than they are (for example, half their size).

That age has an effect on how closely children will interact with a robot should not come as a huge surprise. Though age did not prove to have a significant relationship in the data presented here, these participants were chosen at an age range larger and slightly older than the participants in the pilot experiment. Older children, too, were slightly more likely to want to use the robot as a toy – pushing buttons, looking around the robot to see if anything was attached to it, and generally understanding that while it was acting in a social way, it was still a controllable piece of technology. They were also more willing to interact with the robot without their caregiver.

There may prove to be a sweet spot of age in which typically developing children are more likely to interact with the robot like it is also a social being. We would need to recruit more participants of ages outside our $2-4$ years to be sure. How would this compare to the same exact age of children with autism? Only future research will tell us this – it may be that a wider age range in children with autism react like their typically developing $2-4$ year old counterparts, or that they react like a different age group of typically developing children. Though we have a solid and robust baseline of typically developing children, we do not yet know how atypically developing children compare to this baseline.

There are also lessons to be learned from these experiments simply regarding interdisciplinary research between computer scientists and psychologists. This experiment was made possible by working closely with and incorporating feedback from experts in autism, child development, and children's social preferences. The experiment changes made included adding a dance more likely to be known by very young children that includes instructions ('If you're happy and you know it'), and adding another prompt to form an even number of motions to each of 'I spy' and 'Simon Says.' This enables the experimenter to react to exactly half of the robot's prompting. That is, on the very first Simon Says prompt, 'I touch my nose. Can you touch your nose?' the experimenter responds affirmatively and imitates the motion, and encourages the child to do the same. In the second prompt, ‘I bump my fists. Can you?’ the experimenter instead disengages and looks at a nearby document, ignoring both the robot and the child. This pattern repeats – the experimenter follows along, is disengaged, follows along, is disengaged, follows along – for each of the six prompts in each of the games. In any future research that looks at the contingent responses of the child, we can now evenly compare the responses of the child to an engaged experimenter, and the contingent responses to just the robot’s prompting.

Chapter 5

Data Relationships between Eye Tracking and Human-Robot Interactions

5.1 Participants

A subset of participants from the eye tracking task in Chapter 3 also performed the human-robot interaction experiment in Chapter 4. Using valid data criterion from Chapter 3, we look only at participants who contributed valid data in both eye tracking tasks, and interacted with the NAO robot for which we also had overhead video footage. These 27 total participants were 15 males, 12 females, $25 - 45$ months old, mean age 34.7 months old (2.89 years), standard deviation 6.14 months (0.51 years).

5.2 Results

We first explore relationships between the main dependent measures of the two eye tracking tasks and the human-robot interaction experiment. The eye tracking task measures we examine for task one, dynamic video comparison, is ratio on Social versus NAO, ratio on Social versus Others, and ratio on NAO versus Others; the correlation between these measures and the average Distance Ratio 1 in the HRI experiment is shown in Table 5.1.

Table 5.1: Correlations between three Dynamic Video Comparison measures (Task One) and the Distance Ratio from the human-robot interaction experiment.

HRI measure	Task One measure	Correlation
Ave Dist. Ratio	Ratio on Social vs Others	0.01490516
Ave Dist. Ratio	Ratio on NAO vs Others	-0.3154637
Ave Dist. Ratio	Ratio on NAO vs Social	-0.328696

Table 5.2: Correlations between three Static Image Search measures (Task Two) and the Distance Ratio from the human-robot interaction experiment.

HRI measure	Task Two measure	Correlation
Ave Dist. Ratio	Proportion on NAO	0.1072499
Ave Dist. Ratio	Proportion on Social	-0.1008959
Ave Dist. Ratio	Proportion on Others	0.1201873

For task two, static image search, we examine the proportion of first fixations on NAO, the proportion of first fixations on Social and the proportion of first fixations on Others; the correlation between these measures and the average Distance Ratio in the HRI experiment is shown in Table 5.2.

The correlation coefficients show a weak correlation (above 0.30) between Average Distance Ratio and NAO versus Other robots gaze preferences, (Figure 5.1) and Average Distance Ratio and NAO versus Social gaze preferences, (Figure 5.2). The gaze preferences have been rescaled from 0 to 1 such that 0.5 is actually a preference for neither item on-screen (recall the original vectors, for clarity, were scaled from -100 to 100 to indicate a complete percent preference for NAO, Social, or Other robots). In the gaze preferences versus Distance Ratio plots, in all cases, there is a slight tendency for preference to gaze at NAO to be associated with being closer to NAO.

We also calculated the correlation between age and the ratios from Task One, the proportions of first looks in Task Two, and the Distance Ratio from the HRI experiments, but only age and Distance Ratio had a weak relationship, with a correlation of -0.3725553. This relationship is shown in Figure 5.3.

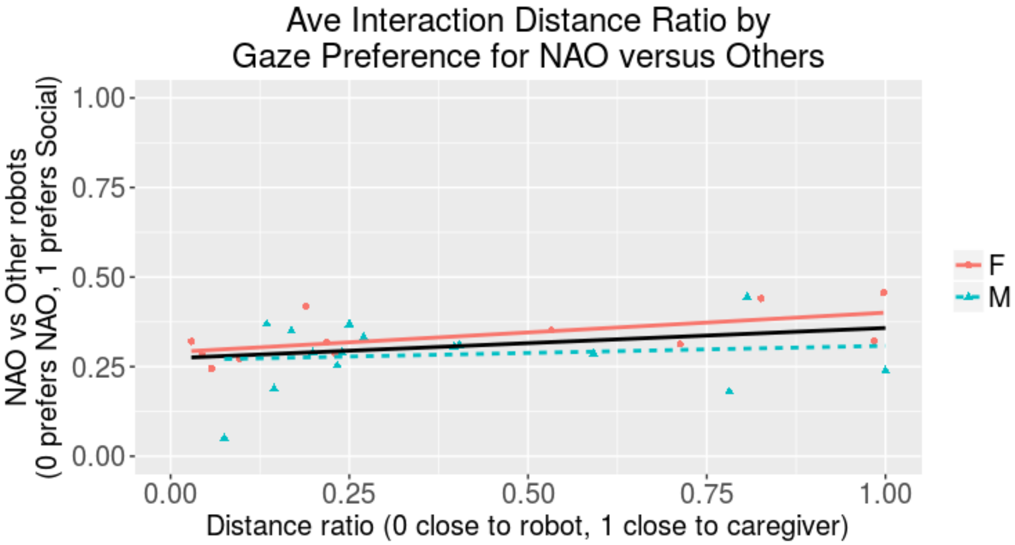


Figure 5.1: Plot of average Distance Ratio versus dynamic gaze preference between NAO and Other types of robots. Using linear regression over the entire dataset (black solid line), males (green dashed line), and females (pink solid line), no real trend is visible.

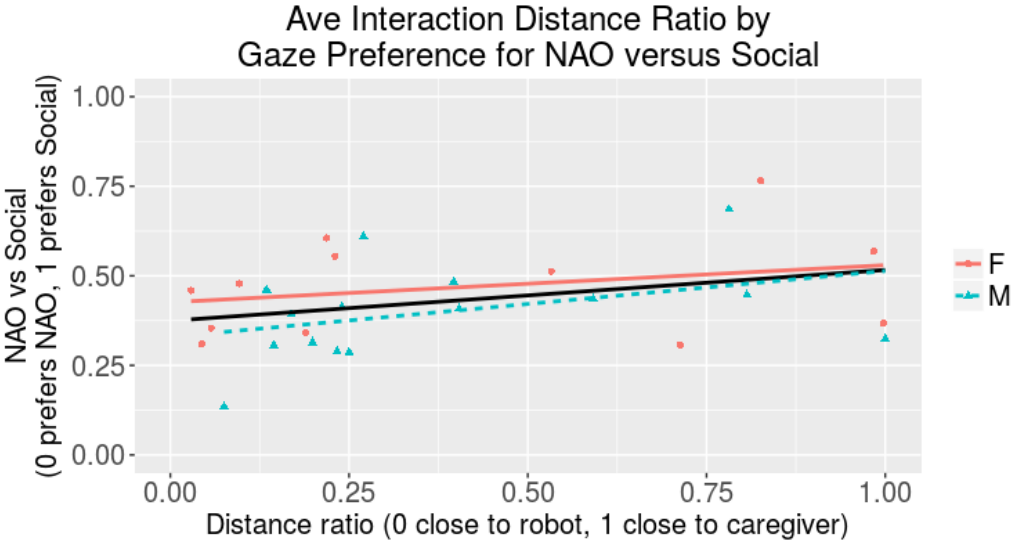


Figure 5.2: Plot of average Distance Ratio versus dynamic gaze preference between NAO and Social. Using linear regression over the entire dataset (black solid line), males (green dashed line), and females (pink solid line), no real trend is visible.

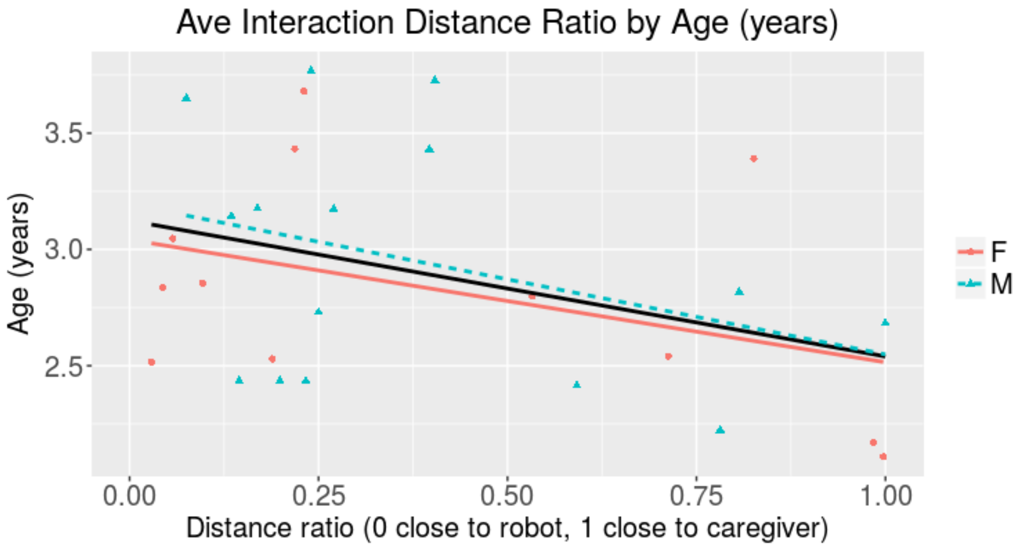


Figure 5.3: Average Distance Ratio plotted against age in years. Older children tend to be closer to NAO during the interaction.

We also explored whether there was an effect of the eye tracking test itself; if a child saw the NAO robot during the eye tracking experiments, are they more likely to be comfortable being closer to the robot in the early presses of the HRI experiment? To test this hypothesis, we compared the average Distance Ratio of participants who finished the first eye tracking task and therefore saw the NAO moving on-screen before seeing it in person, and those who only saw the NAO in person and did not see it move until the interaction. Using the Wilcoxon signed-rank t-test, a non-parametric statistical hypothesis test, we looked at the average Distance Ratio for each press, as well as the average Distance Ratio throughout the entire interaction, but did not find significant p-values.

5.3 Discussion

No strong relationship was found between the eye tracking study dependent measures and proxemics during the human-robot interaction task. Age has a mild effect, shown in a tendency for older children to be closer to NAO relative to their caregivers during the interaction. There is also a weak tendency for children who prefer to look at the NAO during the dynamic eye tracking task when compared with any other robot or Social being to be nearer the NAO during the interaction. No strong effect of sex was found in any task metric comparison.

We can conclude that at this young age, while children have strong preferences in gazing at humanoid robots and social beings over other robot types, this preference does not directly translate into preferring to be closer to the robot relative to caregiver during an in-person interaction. There may be a sex effect that can only be detected with more participants; with 27 participants, male and female tendencies are very similar.

Chapter 6

Applicability to Other Research

We now explore the applicability of this work to other projects, by applying the person tracking software and proxemics analyses to two other research projects that test child development psychology theories. The first research project asks children to open a box that they have opened previously but which, unknown to the participant, is no longer possible to open. The unsolvable box tests how long children will attempt to solve an unsolvable task until they request help from a caregiver or the expert that previously showed them how to open the box. The second research project tests whether children automatically sort themselves by in-group and out-group, as physically indicated by color-coded clothing [86]. The free-play scenario tests whether unacquainted children automatically show an in-group bias by sorting themselves in in-group / out-group, or us-versus-them, groups while playing.

6.1 Unsolvable box

The 'Unsolvable Box' research project asks children to solve an impossible task to determine how long the child is willing to try to solve the task before initiating social interaction [87]. A sample overhead view in Figure 6.1 shows the experiment's environment.

The experimenter shows the participant how to open and close a large, clear, square box. The box has a toy inside; the experimenter first demonstrates how to open the box

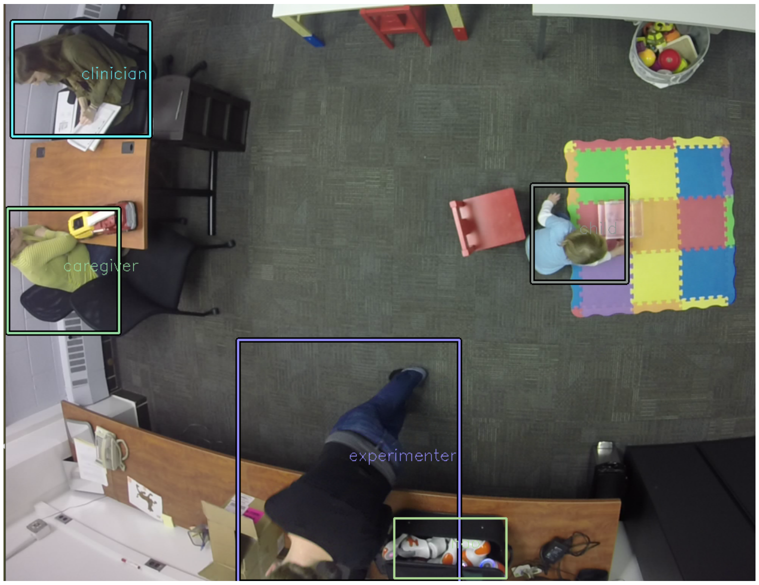


Figure 6.1: A child begins to play with a locked box to retreive the toy inside.

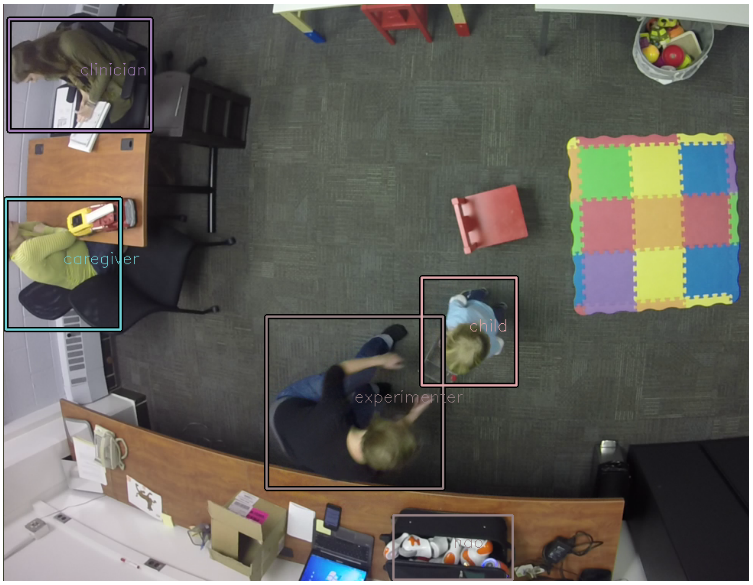


Figure 6.2: The child brings the box to the experimenter, who is known to be able to open the box.

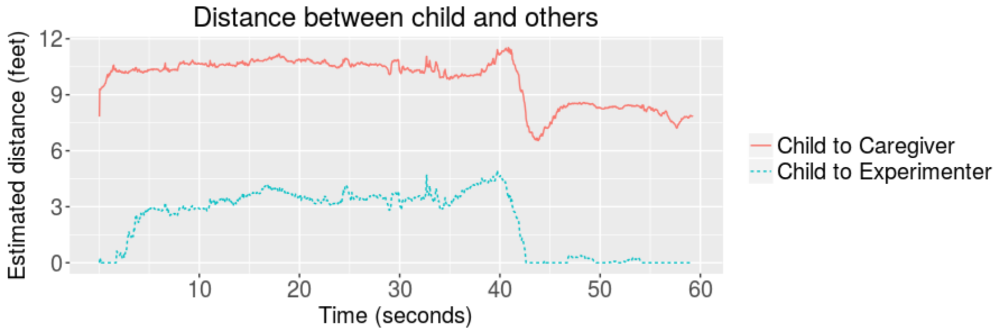


Figure 6.3: Distance between child and experimenter and child and caregiver for the 60 second experiment duration. This child approached the experimenter for help.

and retrieve the toy inside. The experimenter then asks the child to do the same; after the child has demonstrated they are able to open the box and retrieve the toy three times, the experimenter moves the box to an obscured location, switches the toy inside, and locks the box. The experimenter then gives the box back to the child, who is unaware the box is now locked and cannot be opened. The experimenter asks the child to open the box and play with the toy inside, and then removes herself to a distant location and becomes unavailable to answer questions or help. The experiment begins when the experimenter walks away after asking the child to open the locked box, and ends after one minute has passed or the child has insistently asked for help to open the box from the experimenter or a caregiver and the adult has begun responding.

We used the second version of the automated person detector on several sample videos of participants. The tracker runs on 60 second videos, beginning when the experimenter puts the box down in front of the child. The raw distance graphs show a few common patterns: the child approaches the experimenter for help (Figure 6.3 shows the child in Figure 6.2 approach the experimenter around the 40 second mark), the child approaches their caregiver for help (Figure 6.4), or the child does not move to request help for the duration of the experiment (Figure 6.5).

In the original experiment, the distance of the child to others is not considered, only

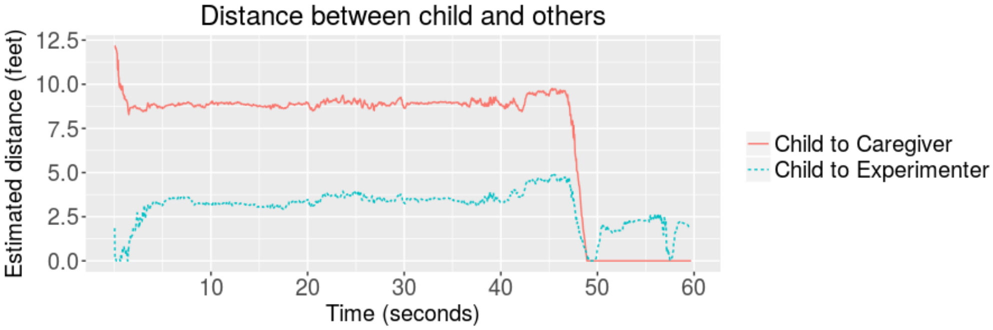


Figure 6.4: Distance between child and experimenter and child and caregiver for the 60 second experiment duration. This child approached their caregiver.

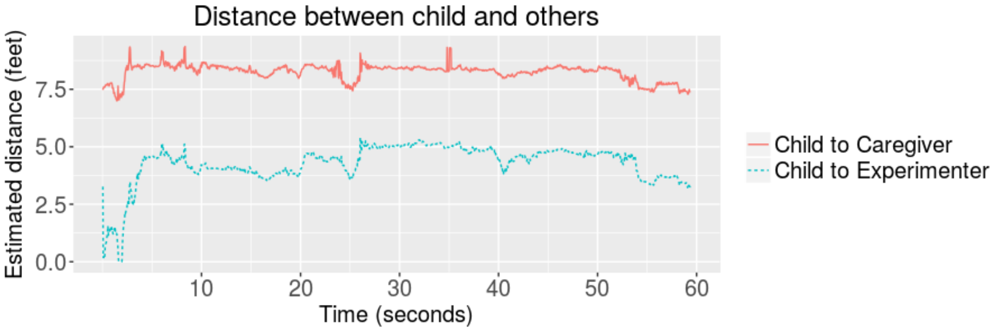


Figure 6.5: Distance between child and experimenter and child and caregiver for the 60 second experiment duration. This child continued to try to open the box for the entire duration.

vocalizations (e.g. the experimenter begins talking to the child). The hand-coding timings are taken from videos on the side of the room, which include reasonable audio, not the footage from the ceiling perspective. However, the proximity of the child to caregiver or experimenter would be reflected in the child's distance shrinking to 0 if the caregiver or experimenter approached the child (to explain how to open the box), or the child approached the caregiver or experimenter (if the child physically brings the box to an adult to fix the box). If verbal discussion happens before proximity shrinks, then the distance data can be considered a maximum bound on experiment time. Given data from entire experiment videos, we could examine the average distances between child and others throughout time, as done in Chapter 4, and examine how lower-stakes games (playing with a clear box with a new toy, rather than playing with a novel robot) affect proximities.

Future work in this vein would include running the tracker on all 84 Unsolvable Box participants' video footage, then testing if the hand-coded experimental length correlates highly with the proxemics generated from the overhead view footage. Positive or strongly correlated results would indicate if, at least at the age of these participants of $1.5-3.75$ years, proxemics might be reasonable stand-ins for audible conversation during social interactions that require help from experts (known and trusted caregivers, or someone who has shown knowledge about how to do an action). Furthermore, though the ages of the participants in the Unsolvable Box were lower than in the HRI experiment (at $1.5 - 3.75$ years), some children participated in both the Unsolvable Box experiment and the HRI experiment. Further examination might show that some behaviors towards the robot correlate to how willing the child is to try to continue solving the Unsolvable Box. For example, children that always stay near their caregiver during an interaction may be more likely to physically approach their caregiver for help after they determine the box cannot be opened, or approach their caregiver faster.

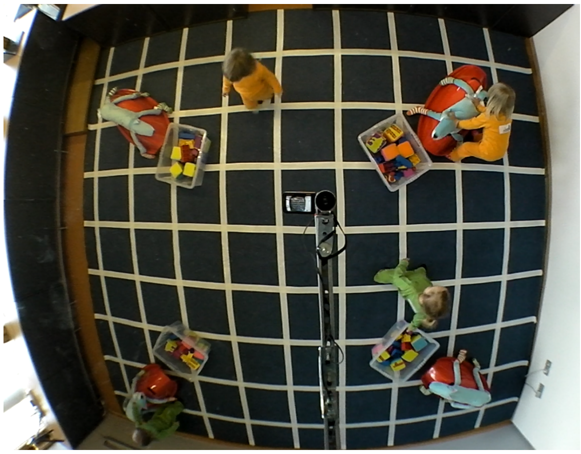


Figure 6.6: A sample image from the beginning of an in-group / out-group research video; four of the randomly color coded children, two in green and two in orange, play in a room without adult direction.

6.2 In-group out-group task

The ‘In-group Out-group’ research project, presented in Dr. Nadja Richter’s thesis, [86] explores whether children will reflect in-group bias in free-play scenarios based on the subtle cue of what color they are wearing. In-group bias is the tendency for people to behave more favorably towards their own group of people, for example those with similar political or religious beliefs, or ethnic or sex similarities. Adults reflect in-group bias, and Dr. Richter examined at what age and what cue subtlety children will also show such bias. The experiments place unacquainted children (always the same age, e.g. 2, or 5, or 8) in a room with several identical sets of toys placed around the room. Before entering the playroom, children put on either an orange or a green jumpsuit. Children are not told that they are part of a ‘green’ or ‘orange’ group, or cued in any way about groupings. A camera is placed, facing up, towards a mirror on the ceiling. A sample of this ceiling footage is shown in Figure 6.6; four children are clearly color coded in orange or green and have free options to play with each other or with the toys.

In the original project, the researchers took a single video frame every 10 seconds for the duration of the experiment (less than half an hour each). The room was taped end-to-end at every meter, and each child's position was calculated based on the closest tape hash-mark. The researchers then calculated the mean distance between pairs of in-group children, such as three children in orange or four children in green, and the distance between pairs of out-group children, which is any pair of children wearing different colors (green and orange). If the mean distance between children wearing the same color was smaller than children wearing different colors, and this difference was statistically significant across the different groups of children and different ages of children, then an in-group bias effect would be proven for children as young as 5 and 8 years old, based on physical similarities and differences.

This manual coding is time intensive; a single experiment with 2 year olds produced $1,517$ frames to annotate for child identity and location. The same experiment, with locations calculated every 1 second, would produce $15,170$ frames. Coding this data for multiple age groups and multiple groups within an age is obviously ripe for automation. The original data did not show an in-group bias. However, it's possible that with more data, patterns or biases might appear. With that in mind, we applied the deep learning tracker to a sample session of 2.5 year old children. A sample frame is shown in Figure 6.7.

We first apply the automated tracker to the entire video session of 2.5 year old children. We calculated the raw distance between the center of each child's found bounding box in pixels, calculated the distance between children in different groups and same groups, and found the means of pairs of in-groups and out-groups; Figure 6.8 shows the resulting means found at a frame rate of 30 fps. The raw pixel distance was scaled by dividing by 125 for visibility on the graph.

We also compared the true, manually coded distances found between participants with the automated distances, shown in Figure 6.9. Even without accounting for the slight fish-eye effect of distorted video footage on the calculated distances, the distances found look close – there are peaks and valleys in roughly the same spots, even without performing any sort of

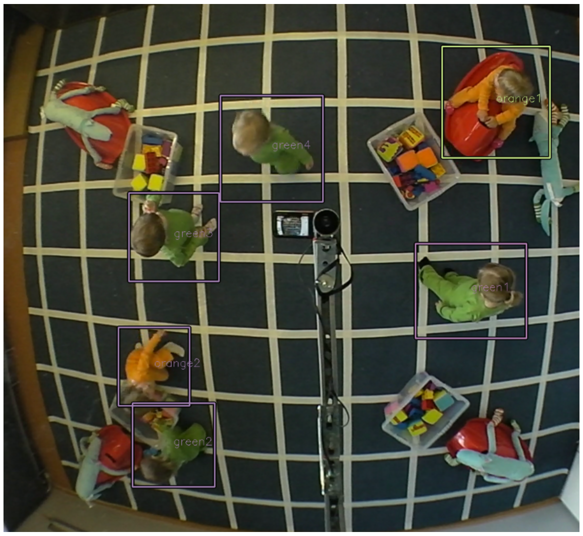


Figure 6.7: A sample image from an in-group / out-group research experiment after the deep learning tracker is applied.

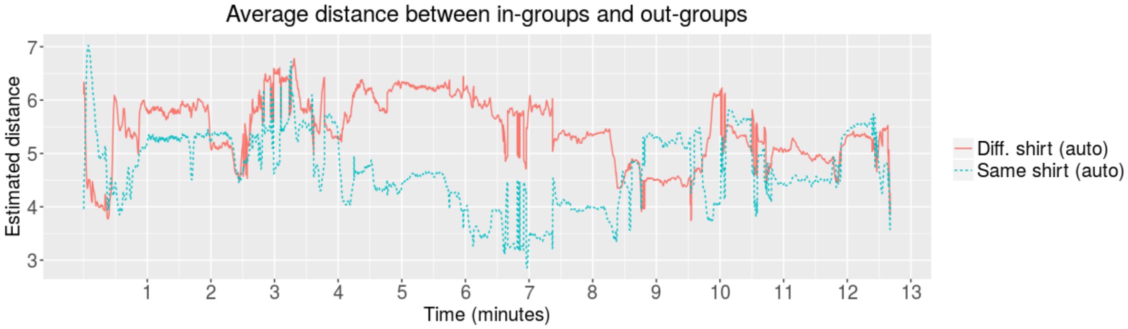


Figure 6.8: The average distance between in-groups (pairs of children in green, and pairs of children in orange) found from the automated tracker for the first 13 minutes (every frame at 30 fps) of an experiment session. The estimated distance is the Euclidean distance in pixels (scaled by 125), between the centers of the bounding box found for each child; it is not a true distance.

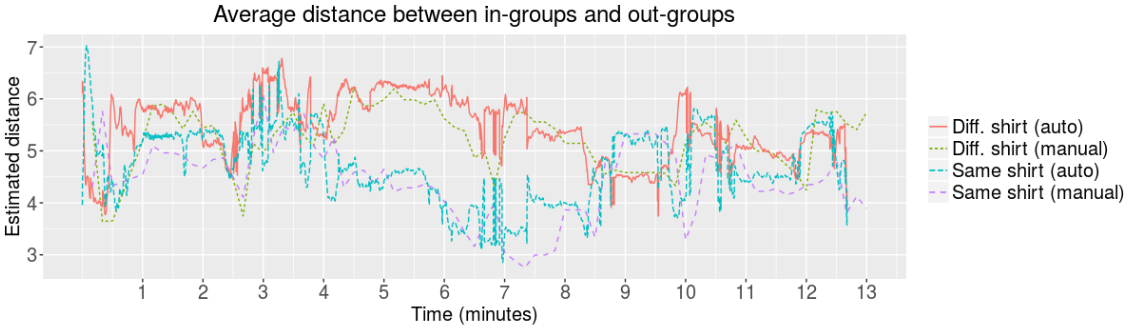


Figure 6.9: The original manually coded distance information (in meters) versus the calculated pixel distance (divided by 125) from the automated tracker.

smoothing on the child coordinates found from the tracker.

Because the automated data reflects so much more data than the hand-coded results, we also reduced the automated information by taking only distances found every 10 seconds. We then find the correlation between the time series – ideally, the distances found by hand will correlate highly with the automated distances. We computed the Pearson correlation coefficient between the means for children in different-color shirts in the manual and automated data, and the correlation coefficient between the means for children in same-color shirts in the manual and automated data. For children in different colors, the correlation was strong, at 0.52; Pearson's product moment correlation found a highly significant p-value of < 0.001 ($t = 5.2398$, $df = 74$). For children in the same colors, the correlation was even stronger, at 0.68; Pearson's product moment correlation found a highly significant p-value of < 0.001 ($t = 7.91$, $df = 74$). The manually coded distances and the reduced automated distances are shown in Figure 6.10.

What these results tell us is that even without precise distance unit measurements and without un-distorting overhead video footage, we can generate measurements similar to hand-annotated ground-truth. Future work in this direction would include adding functionality to estimate the child's nearest meter location (instead of using bounding box centers), then running the tracker on all 12 videos from the in-group bias experiment, then testing if

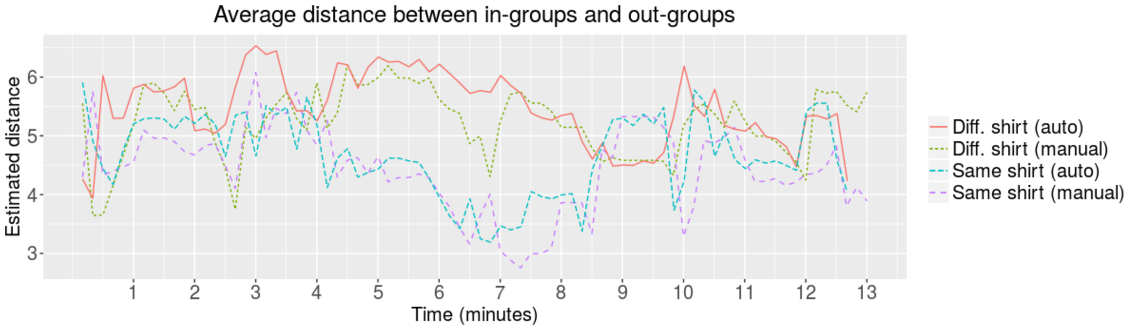


Figure 6.10: The original manually hand-coded distance information (in meters) versus the calculated pixel distance (divided by 125) from the automated tracker.

the results are the same as in the original experiment. The original research found no difference in in-group and out-group physical distances in the free-play, color-coded child scenario. With more measurements across time, however, such differences may become more apparent.

Chapter 7

Summary and Conclusions

This thesis began with the goal of characterizing and autism phenotype in very young children. To this end, we designed and ran two end-to-end experiments. The first experiment, detailed in Chapter 3, investigated if children showed statistically significant preferences in a paired visual comparison task and preferential orientation in a visual search task between three types of robots and a social peer. The robot types were anthropomorphic (having human qualities), zoomorphic (having animal qualities), and caricatured (having animated or cartoon qualities), and the social peer was a 4 year old girl.

This experiment revealed that in the paired visual comparison of social being to NAO, NAO to other robots, and the social being to the other (non-NAO) robots, children strongly preferred to look at NAO over the social peer, NAO over other robots, and the social peer over other robots. In the visual search task, children preferred to look immediately at the social peer, but by the third fixation, they were just as likely to look at the NAO as the social peer. In the first five fixations, children were more likely to look at NAO and the social peer than any of the other robots. Overall, the eye tracking experiments tell us that children this age, $1.5 - 3.75$ years, show a preference for the humanoid robot NAO that is similar to the preference for live human peers, and that the preference for the NAO is greater than for other robots.

Future work in this area should explore if there is a preference for the gender of the social peer. This research used a female social peer, but boys and girls may have different preferences for the peer's gender. Though there was no statistically significant effect of participant sex on the results, males showed stronger preference for gazing at the NAO than females. This effect may be more pronounced, or prove to not exist, with a larger number of participants. There was also a non-significant but visible effect that boys prefer looking at the humanoid robot after the first few fixations on the social peer, whereas girls looked at the social being for more first fixations than the boys. This, too, might be impacted by the perceived gender of the social peer.

Having shown that the humanoid robot is a highly interesting (gaze-worthy) object, the second experiment, described in Chapter 4, investigated how children react to that same novel humanoid robot during a 9-15 minute interaction. The robot played different game or dances; each game was really a press for social interaction, like imitating behaviors or looking for items. Starting with open-source tracking software and moving to state-of-the-art techniques in deep learning to track people and robots across time, this experiment showed that proxemics, or distances between people, differed in children during a humanoid robot interaction. Typical reactions included staying near the caregiver throughout the experiment, staying near the robot, moving from closer to the caregiver to closer to the robot, moving from closer to the robot to closer to the caregiver, or no apparent preference or movement towards a person or robot. We then showed that combining these proxemics with a physical and social dimension that are both classically associated with autism, ritualized behaviors and socialization skills, produced groups of children that were statistically different, even though these children are all typically developing. We tested that these groups of children differ in two different physical and social dimensions associated with autism, restricted behaviors and communication skills.

We next investigated the relationships between the eye tracking results and the human-robot interaction experiment in Chapter 5. Though 27 children participated in both research projects, we found no strong, statistically significant relationships in the data. A larger sample size of children with results from both experiments, as well as accounting for potential effect of the gender of the social peer in eye tracking experiments, may reveal solid relationships between the data, and such an investigation is left to future work.

Proxemics, the study of social distances between people, was a core area investigated in the human-robot interaction study. We thus applied the same methods used to track children in the robot interaction experiment to two other research projects, discussed in Chapter 6. In the first project, we showed how children trying to open an unsolvable, locked box show proxemics differences during their attempts to unlock the box. The second project looks at how children in an in-group bias experiment differ from each other. We showed how the automated tracker, without any undistortion on the video or specific unit measurements applied to the tracked children, can show in-group and out-group differences. Our results were highly correlated to the hand-coded results from the original experiment, only our data was far more specific, generated for each frame rather than at the 10-second intervals used in the original experiment.

The largest open issue found from this work, approached in two different ways in Section 4.5, is automatically determining orientation based on just a overhead picture of a person. Though both multiple-kernel learning and deep learning methods showed good performance against training databases, neither performed well enough in practice to apply to the person-tracking experiments to automatically detect where a child was facing. Knowing if a child is facing the robot, their caregiver, the experimenter, or none of the above would allow us to build a more robust offline tracker that can potentially identify contingent looks. A child that continually looks at the robot during an experiment but sits on or near their parent is likely different from a child that spends half their time looking around the room or playing with a toy while they sit on or near their parent. Even a thoroughly trained neural network on carefully cultivated samples from different experiments was not enough to reliably generate orientation on an overhead photo. This is highly recommended as an area for future work, and would help research into this space of automating reactions and behaviors in developmental disorders such as autism spectrum disorder.

Appendix A

Programming the NAO Robot

The NAO robot can be programmed with proprietary software called Choregraphe, which comes with the purchase of any NAO robot. The Choregraphe version used in this thesis was version 2.1.4, on a NAO robot version H25, V4, purchased in 2013. The NAO's software operating system was version 2.1.4.13. Choregraphe is a drag-and-drop user interface that comes with many pre-programmed actions for the robot, including sitting, standing, waving one arm, doing a Tai Chi dance with accompanying music, and speaking (given text). Choregraphe also allows a user to program the NAO through its interface in Python, or by training the NAO's motors into specific configurations.

The NAO interaction program built for this thesis used all three methods. Figure A.1 shows an overview of the program containing the games the NAO plays with children. It includes commands (written in python) to record the audio and video from the robot's perspective, introducing itself, blocks containing "Simon Says" (called "I do this" for simplicity's sake), blocks containing the "I spy" (called "I see this" for simplicity's sake), blocks containing the "If you're happy and you know it," "Bye Bye Bye," and "Tai Chi" dances, and other blocks to speak, wait for a touch from the experimenter, and blocks to adjust timing or robot perspective. Figure A.2 shows the python program interface; Figure A.3 shows just one piece of a "Simon Says" action, a toe touch. Figure A.4 shows the "key frames," or joint positions, during the toe touch. The "Simon Says" actions, the

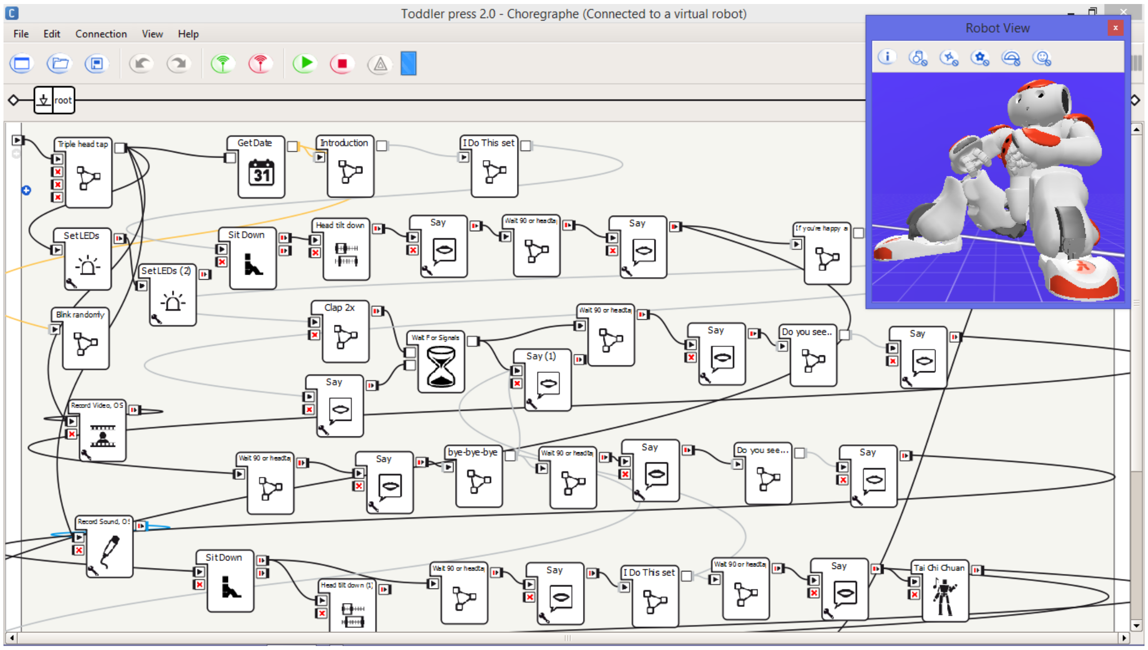


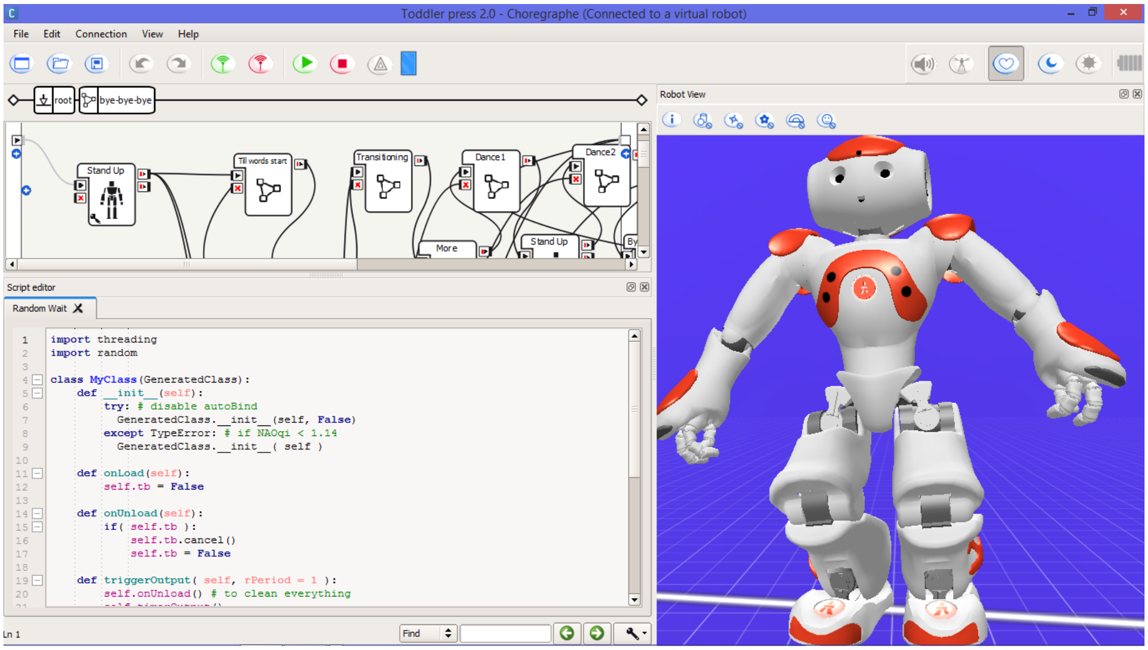
Figure A.1: An overview of the toddler games used in the human-robot interaction sequence.

looking-around behavior exhibited during the “I spy” presses, and the “If you’re happy and you know it” dance were programmed by the author and experimenter. The “Bye Bye Bye” dance sequence was programmed by an undergraduate student at the University of Minnesota, and the standing up, sitting down, and Tai Chi movements were already available in Choregraphe.

The full program text is given in Table A.1. Between each press is additional encouragement from the robot, such as “you’re doing a great job. Let’s keep playing!” after which it waits for the experimenter to press a button on it to start the next press.

Table A.1: The program text and actions. The program takes about 9.5 minutes to run if the experimenter rapidly progresses between presses.

Press	Action or speech
Introduction	(wave hello) My name is Robbie. I like to play games. Can you play with me? Try to follow along.
1	I touch my nose. Can you touch your nose? I bump my fists. Can you? I clap my hands. Can you clap your hands? I put my hands on my head. Can you put your hands on your head? I clap like this. Can you clap? I touch my toes. Can you touch your toes?
2	Would you like to dance? I know a song called If you're happy and you know it. (Dance)
3	Let's play I spy. Can you look around with me? Do you see a red chair? Do you see a cat? Do you see a blicket? Do you see a robot? Do you see a door? Do you see a window?
4	I want to show you a dance I like to do. Have you heard this song? It's called bye bye bye. (Bye Bye Bye dance)
5	Do you see a computer? Do you see a shoe? Do you see a phone? Do you see a table? Do you see a black chair? Do you see a cabinet?
6	Let's play again and you can copy what I do. Look over here. (looks right) Can you look over there? I wave like this. Can you wave? I look over here. (looks left) Can you look there? I nod my head. Can you nod? I point over there. Can you point? I move my fingers. Can you move your fingers?
7	(Tai chi dance)



```
import threading
import random

class MyClass(GeneratedClass):
    def __init__(self):
        try: # disable autoBind
          GeneratedClass.__init__(self, False)
        except TypeError: # if NAOqi < 1.14
          GeneratedClass.__init__( self )

    def onLoad(self):
        self.tb = False

    def onUnload(self):
        if( self.tb ):
            self.tb.cancel()
            self.tb = False

    def triggerOutput( self, rPeriod = 1 ):
        self.onUnload() # to clean everything
```

Figure A.2: An example of the python code available to edit or add in the Choregraphe interface.

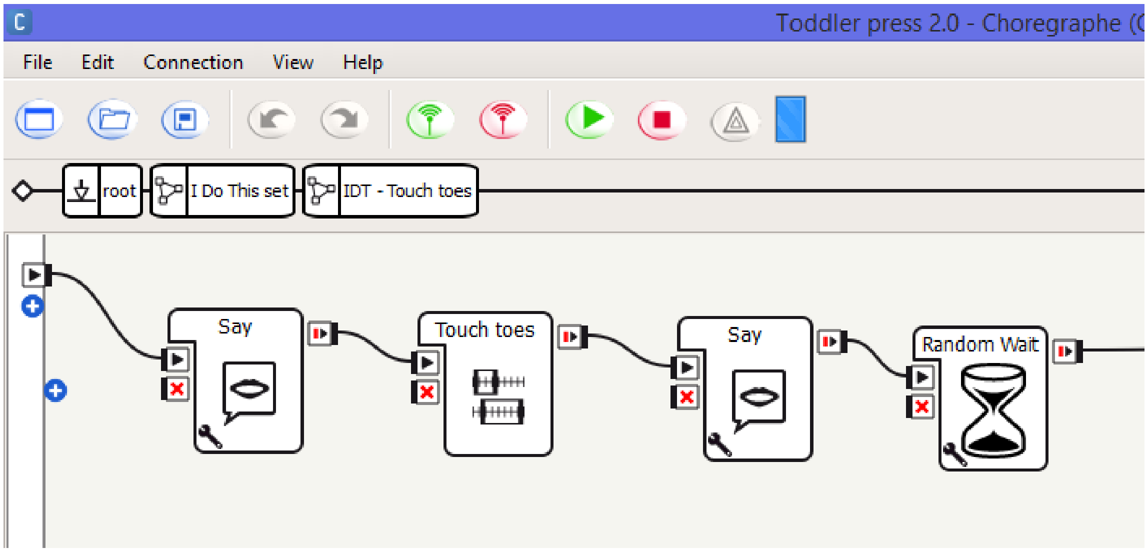


Figure A.3: A sample item, the toe touch, in a "Simon Says" set. The robot says "I touch my toes," leans forward to touch the tip of one foot, sits upright again, says "can you?" and waits for a random amount between $1 - 3$ seconds for the child to react.

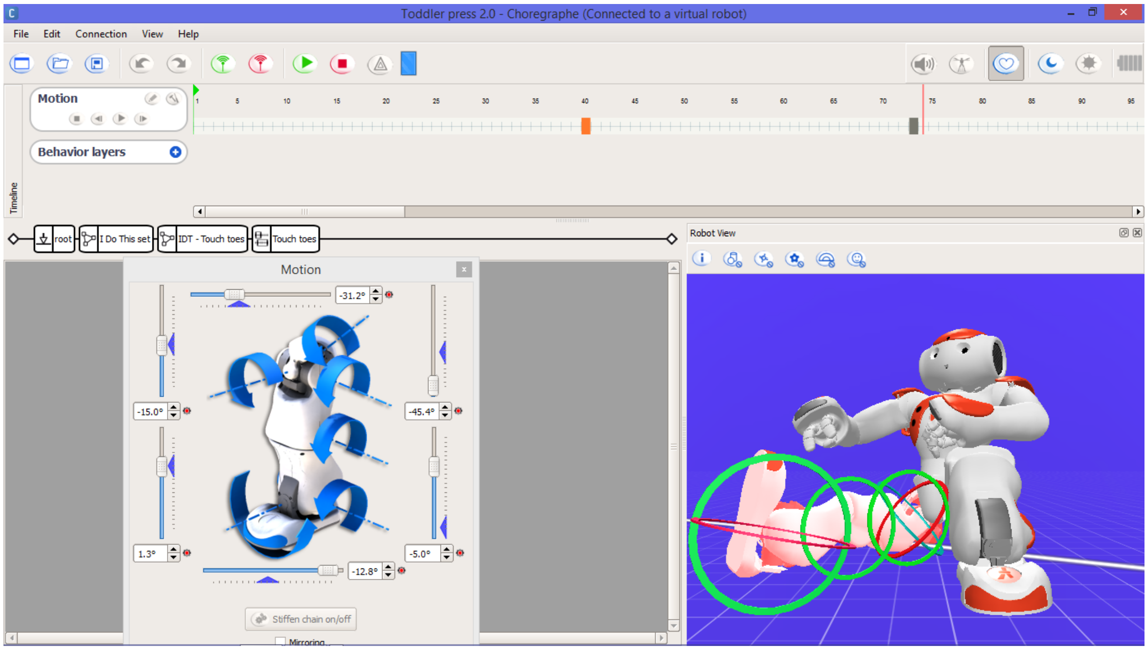


Figure A.4: A sample view of the joint positions and motion programming available through the Choregraph interface.

Appendix B

Child Development Assessments

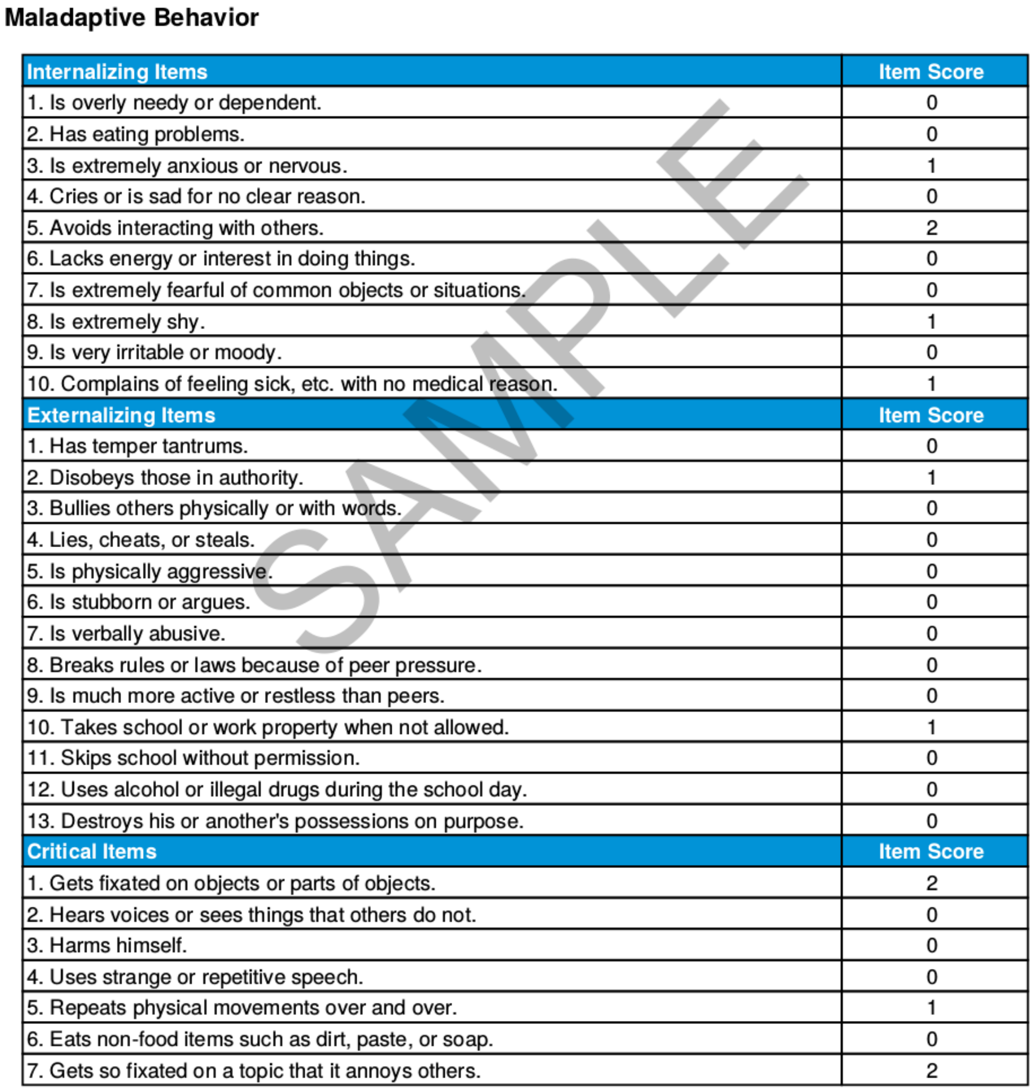
B.1 Vineland Adaptive Behavior Scales

This thesis uses real participant scores generated from the Vineland Adaptive Behavior Scales [107], or Vineland, or VABS. This development assessment gives participants an overall rating of adaptive functioning, based on how often the participant shows some action or behavior in real life, as reported by someone who knows the participant well such as a caregiver or teacher. Behaviors are not rated on how the child performs on a particular day of testing; the assessment asks the reporter how frequently a child shows a particular behavior in daily living, on a $0-2$ or never / sometimes (partially independent) / usually (without help, independent) scale. The questions ask if the person has *seen* the child do the item in question, not if the child is *capable* of doing it. The Adaptive Behavior Composite is a score based on three areas, or domains: communication, daily living, and socialization scores. All four scores, the three domains and the composite score, are standardized to a mean of 100 and a standard deviation of 15. An adequately developing child would fall somewhere within one standard deviation of the mean, or within the $86-114$ point range on the composite and the domain scores.

Each domain measures a particular area of daily functioning. The communication domain measures how well a child listens, understands, reads, writes, and expresses themselves through speech. The daily living domain measures performance in common, practical tasks at home or school, such as zips zippers that are fastened on the bottom, as on a backpack. The socialization domain measures skills in social situations, like playing cooperatively with more than one other child for more than 5 minutes. The actual exam questions vary by age of the participant; some sub-domains, like communication, can be assessed in children less than one year old, but some sub-domains do not apply to children younger than three years old.

Each area of assessment for adaptive behaviors has a list of questions in order of complexity; for each set of questions, the examiner will ask if the child does it on the $0-2$ scale. The base scale is 4 consecutive items marked with 2 (child usually does it without help) and the ceiling is 4 consecutive items marked 0 or ‘does not do.’ A three year old child might sometimes ‘follow instructions in “in-then” form’, and sometimes ‘listen to a story for at least 15 minutes,’ but they do not ‘listen to a story for at least 30 minutes,’ ‘follow 3-part instructions,’ ‘follow instructions or directions heard 5 minutes ago,’ or understand idioms, at which point the examiner will stop assessing the receptive section of the communication domain.

The maladaptive behaviors section, unlike adaptive behavior sections, is asked from start to finish, and looks at internalizing behaviors, externalizing behaviors, and critical items. These questions look for problematic, interrupting behaviors in life or school settings, such as destroying possessions intentionally or extreme anxiety. A sample scoring sheet for the maladaptive behavior section of the assessment for a purely fictitious 8 year old male is in Figure B.1 [113]. The $0-2$ scoring is the same as the adaptive behaviors; in the sample scoring sheet, 0 indicates none of the behavior, and 1 or 2 indicates the presence of severely maladaptive behaviors, which are causes for concern.



Maladaptive Behavior

Internalizing Items	Item Score
1. Is overly needy or dependent.	0
2. Has eating problems.	0
3. Is extremely anxious or nervous.	1
4. Cries or is sad for no clear reason.	0
5. Avoids interacting with others.	2
6. Lacks energy or interest in doing things.	0
7. Is extremely fearful of common objects or situations.	0
8. Is extremely shy.	1
9. Is very irritable or moody.	0
10. Complains of feeling sick, etc. with no medical reason.	1
Externalizing Items	**Item Score**
1. Has temper tantrums.	0
2. Disobeys those in authority.	1
3. Bullies others physically or with words.	0
4. Lies, cheats, or steals.	0
5. Is physically aggressive.	0
6. Is stubborn or argues.	0
7. Is verbally abusive.	0
8. Breaks rules or laws because of peer pressure.	0
9. Is much more active or restless than peers.	0
10. Takes school or work property when not allowed.	1
11. Skips school without permission.	0
12. Uses alcohol or illegal drugs during the school day.	0
13. Destroys his or another's possessions on purpose.	0
Critical Items	**Item Score**
1. Gets fixated on objects or parts of objects.	2
2. Hears voices or sees things that others do not.	0
3. Harms himself.	0
4. Uses strange or repetitive speech.	0
5. Repeats physical movements over and over.	1
6. Eats non-food items such as dirt, paste, or soap.	0
7. Gets so fixated on a topic that it annoys others.	2

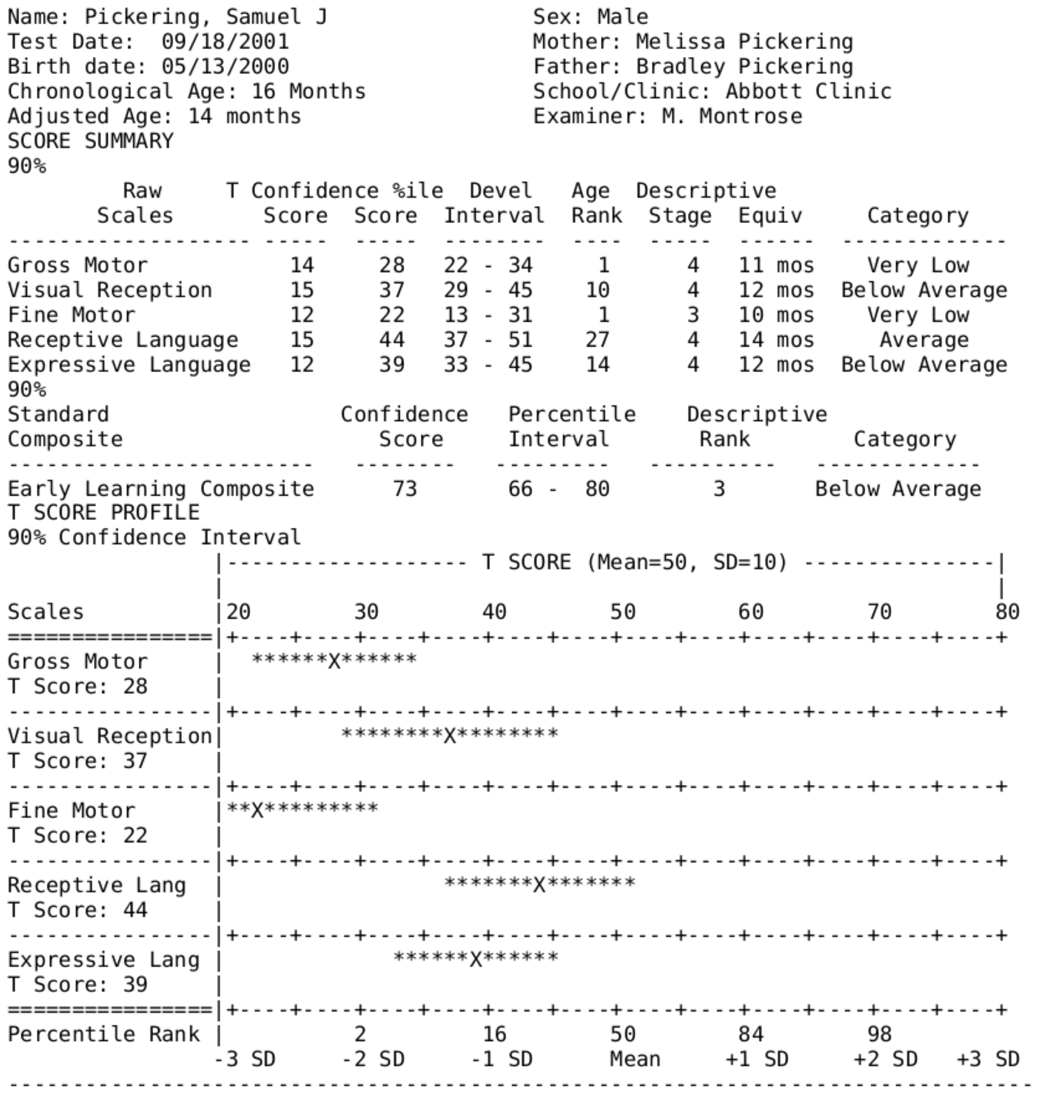
Figure B.1: Part of a sample scoring sheet for a fictitious 8 year old male. This sample scoring sheet comes from Vineland, 3rd edition.

B.2 Mullen Scales of Early Learning

This thesis uses real participant scores generated from the Mullen Scales of Early Learning [74], or Mullen. This exam is delivered to the child over a 1-2 hour time period by a trained examiner, unlike the previous caregiver interview assessment. There are five main areas assessed: Gross Motor, Visual Reception, Fine Motor, Expressive Language, and Receptive Language. Combined, these form an Early Learning Composite score. In each area, the examiner asks the child questions, or asks them to demonstrate physical activities, which get progressively more demanding, until the child is unable to complete three tasks in a row.

The question format depends on the area; for example, Gross Motor Scale requires standing, walking, and running. The Visual Reception scale requires matching, sorting, and nesting cups, whereas the Fine Motor Scale requires stacking blocks, drawing, and stringing beads. The Receptive Language Scale requires recognizing body parts and following commands. The Expressive Language scale requires answering questions and completing analogies. Each area question is formatted such that it does not require a different area; for example, receptive language does not need the participant to speak. A completely successful task might be scored from 1 - 5; some actions, like unscrewing and screwing a large plastic nut and bolt, is a 1 (does it) or 0 (does not do it) score, but stacking blocks vertically might score 0 (no blocks), 1 (3 - 5 blocks stacked), 2 (6 - 8 blocks stacked), or 3 (9 or more blocks stacked).

A sample plain-text scoring sheet for the Mullen assessment for a purely theoretical 16 month old male is in Figure B.2 [112]. It shows the score for each domain, the composite score, and the equivalent developmental age in each area for this child.



```
Name: Pickering, Samuel J                Sex: Male
Test Date:  09/18/2001                   Mother: Melissa Pickering
Birth date: 05/13/2000                   Father: Bradley Pickering
Chronological Age: 16 Months             School/Clinic: Abbott Clinic
Adjusted Age: 14 months                  Examiner: M. Montrose
SCORE SUMMARY
90%
         Raw     T Confidence %ile  Devel   Age  Descriptive
       Scales       Score  Score  Interval  Rank  Stage  Equiv     Category
------------------- -----  -----  --------  ----  -----  ------  -------------
Gross Motor           14     28   22 - 34     1      4   11 mos    Very Low
Visual Reception      15     37   29 - 45    10      4   12 mos  Below Average
Fine Motor            12     22   13 - 31     1      3   10 mos    Very Low
Receptive Language    15     44   37 - 51    27      4   14 mos     Average
Expressive Language   12     39   33 - 45    14      4   12 mos  Below Average
90%
Standard                  Confidence   Percentile    Descriptive
Composite                    Score     Interval       Rank        Category
------------------------   --------    ---------  ----------   -------------
Early Learning Composite      73       66 -  80        3       Below Average
T SCORE PROFILE
90% Confidence Interval
                |------------------ T SCORE (Mean=50, SD=10) ---------------|
                |                                                           |
Scales          |20        30        40        50        60        70        80
================|+----+----+----+----+----+----+----+----+----+----+----+----+
Gross Motor     |  ******X******
T Score: 28     |
----------------|+----+----+----+----+----+----+----+----+----+----+----+----+
Visual Reception|         ********X********
T Score: 37     |
----------------|+----+----+----+----+----+----+----+----+----+----+----+----+
Fine Motor      |**X*********
T Score: 22     |
----------------|+----+----+----+----+----+----+----+----+----+----+----+----+
Receptive Lang  |                *******X*******
T Score: 44     |
----------------|+----+----+----+----+----+----+----+----+----+----+----+----+
Expressive Lang |            ******X******
T Score: 39     |
================|+----+----+----+----+----+----+----+----+----+----+----+----+
Percentile Rank |         2         16        50        84        98
                -3 SD     -2 SD     -1 SD      Mean     +1 SD     +2 SD   +3 SD
--------------------------------------------------------------------------------
```

Figure B.2: Sample final score tally of a fictitious 16 month old male.

References

[1] TOBII Pro.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] Thomas Achenbach. *Manual for the Child Behavior Checklist*. University of Vermont Department of Psychiatry, Burlington, Vt, 1991.

[4] Dagmara Annaz, Ruth Campbell, Mike Coleman, Elizabeth Milne, and John Swettenham. Young children with autism spectrum disorder do not preferentially attend to biological motion. *Journal of Autism and Developmental Disorders*, 42(3):401–408, 2012.

[5] Kosuke Asada, Yoshikuni Tojo, Hiroo Osanai, Atsuko Saito, Toshikazu Hasegawa, and Shinichiro Kumagaya. Reduced personal space in individuals with autism spectrum disorder. *PLOS One*, 11(1), 2016.

[6] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[7] Jon Baio, Lisa Wiggins, Deborah L. Christensen, Matthew J Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, Maureen S. Durkin, Pamela Imm, Loizos Nikolaou, Marshalyn Yeargin-Allsopp, Li-Ching Lee, Rebecca Harrington, Maya Lopez, Robert T. Fitzgerald, Amy Hewitt, Sydney Pettygrove, John N. Constantino, Alison Vehorn, Josephine Shenouda, Jennifer Hall-Lande, Kim Van Naarden Braun, and Nicole F. Dowling. Prevalence of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, united states, 2014. *Morbidity and Mortality Weekly Report Surveillance Summaries*, 67(6):1 – 23, 2018.

[8] Gail A. Bernstein, Tasoulla Hadjiyanni, Kathryn R. Cullen, Julia W. Robinson, Elizabeth C. Harris, Austin D. Young, Joshua Fasching, Nicholas Walczak, Susanne Lee, Vassilios Morellas, and Nikolaos Papanikolopoulos. Use of computer vision tools to identify behavioral markers of pediatric obsessivecompulsive disorder: a pilot study. *Journal of Child and Adolescent Psychopharmacology*, 27(2):cap.2016.0067, 2016.

[9] Jonathan Bidwell, Irfan a. Essa, Agata Rozga, and Gregory D. Abowd. Measuring child visual attention using markerless head tracking from color and depth sensing cameras. *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pages 447–454, 2014.

[10] Pieter Blignaut and Daniël Wium. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior research methods*, 46(1):67–80, 2014.

[11] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and

Chris Moore. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages 1–143, 1998.

[12] Alice S. Carter and Margaret J. Briggs-Gowan. Manual of the infant-toddler social-emotional assessment. *New Haven, CT: Yale University*, 2000.

[13] Thierry Chaminade and Maria M. Okka. Early processes of social attention elicited by a humanoid robot. In *The 22nd IEEE International Symposium on Robot and Human Interactive Communication*, pages 436–440. IEEE, 2013.

[14] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

[15] T. Charman. Why is joint attention a pivotal skill in autism? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1430):315–324, 2003.

[16] Pauline Chevalier, Gennaro Raiola, Jean-claude Martin, Brice Isableu, Christophe Bazile, and Adriana Tapus. Do sensory preferences of children with autism impact an imitation task with a robot? *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 177–186, 2017.

[17] Google codelabs. Tensorflow for poets.

[18] Kenneth Mark Colby. The rationale for computer-based treatment of language difficulties in nonspeaking autistic children. *Journal of Autism and Developmental Disorders*, 3(3):254–260, 1973.

[19] TensorFlow Community. Tensorflow object detection API, 2018.

[20] John N. Constantino and Christian P. Gruber. *Social Responsiveness Scale, Second Edition*. Western Psychological Services, Los Angeles, CA, 2012.

[21] John N Constantino, Stefanie Kennon-McGill, Claire Weichselbaum, Natasha Marrus, Alyzeh Haider, Anne L Glowinski, Scott Gillespie, Cheryl Klaiman, Ami Klin, and Warren Jones. Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature*, 547(7663):340, 2017.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *European Conference on Computer Vision*, 23(6):484–498, 1998.

[23] D Cristinacce and T Cootes. Feature detection and tracking with constrained local models. *Proceedings of the 2006 British Machine Vision Conference*, pages 929–938, 2006.

[24] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:886–893, 2005.

[25] Kirsten A Dalrymple, Marie D Manner, Katherine A Harmelink, Elayne P Teska, and Jed T Elison. An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in psychology*, 9, 2018.

[26] Kerstin Dautenhahn and Iain Werry. Towards interactive robots in autism therapy: background, motivation and challenges. *Pragmatics and Cognition*, 12(1):1–35, 2004.

[27] Geraldine Dawson, Emily J H Jones, Kristen Merkle, Kaitlin Venema, Rachel Lowy, Susan Faja, Dana Kamara, Michael Murias, Jessica Greenson, and Jamie Winter. Early behavioral intervention is associated with normalized brain activity in young

children with autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51(11):1150–1159, 2012.

[28] Joshua J Diehl, Lauren M Schmitt, Michael Villano, and Charles R Crowell. The clinical use of robots for individuals with autism spectrum disorder: a critical review. *Research in Autism Spectrum Disorders*, 6:249–262, 2012.

[29] Jed T. Elison, Jason J. Wolff, Debra C. Heimer, Sarah J. Paterson, Hongbin Gu, Heather C. Hazlett, Martin Styner, Guido Gerig, and Joseph Piven. Frontolimbic neural circuitry at 6 months predicts individual differences in joint attention at 9 months. *Developmental Science*, 16(2):186–197, 2013.

[30] Gary W Evans and Roger B Howard. Personal space. *Psychological bulletin*, 80(4):334, 1973.

[31] Torbjörn Falkmer, Katie Anderson, Marita Falkmer, and Chiara Horlin. Diagnostic procedures in autism spectrum disorders: A systematic literature review. *European Child and Adolescent Psychiatry*, 22(6):329–340, 2013.

[32] Joshua Fasching, Nicholas Walczak, Gail A Bernstein, Tasoulla Hadjiyanni, Kathryn Cullen, Vassilios Morellast, and Nikolaos Papanikolopoulos. Automated coding of activity videos from an OCD study. *IEEE International Conference on Robotics and Automation*, pages 5638–5643, 2016.

[33] Joshua Fasching, Nicholas Walczak, Vassilios Morellas, and Nikolaos Papanikolopoulos. Classification of motor stereotypies in video. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4894–4900, 2015.

[34] Joshua Fasching, Nicholas Walczak, Ravishankar Sivalingam, Kathryn Cullen, Barbara Murphy, Guillermo Sapiro, Vassilios Morellas, and Nikolaos Papanikolopoulos.

Detecting risk-markers in children in a preschool classroom. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1010–1016, 2012.

[35] David J Feil-Seifer and Maja J Matarić. Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features. *Proceedings of the 6th International Conference on Human-Robot Interaction*, pages 323–330, 2011.

[36] David J Feil-Seifer and Maja J Matarić. Distance-based computational models for facilitating robot interaction with children. *Journal of Human-Robot Interaction*, 1(1):55–77, 2012.

[37] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4):143–166, 2003.

[38] Centers for Disease Control and Prevention. What is fragile x syndrome?

[39] Yoav Freund and Robert E. Schapire. A short introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[40] Francesca Fulceri, Annarita Contaldo, Ilaria Parrini, Calderoni Sara, Antonio Narzisi, Raffaella Tancredi, Fabio Apicella, and Filippo Muratori. Locomotion and grasping impairment in preschoolers with autism spectrum disorder. *Clinical Neuropsychiatry*, 12(4):94–100, 2015.

[41] Michael L Ganz. The lifetime distribution of the incremental societal costs of autism. *Archives of Pediatrics and Adolescent Medicine*, 161(4):343–349, 2007.

[42] Erica Gessaroli, Erica Santelli, Giuseppe di Pellegrino, and Francesca Frassinetti. Personal space regulation in childhood autism spectrum disorders. *PLOS One*, 8(9):e74959, 2013.

[43] JC Gilliam. Gars-3: Gilliam autism rating scale–third edition, 2014.

[44] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[45] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[46] Robert Goodman. Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(11):1337–1345, 2001.

[47] Michael a. Goodrich, Mark a. Colton, Bonnie Brinton, and Martin Fujiki. A case for low-dose robotics in autism therapy. *Proceedings of the 6th International Conference on Human-Robot Interaction*, page 143, 2011.

[48] Quentin Guillon, Nouchine Hadjikhani, Sophie Baduel, and Bernadette Roge. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience and Biobehavioral Reviews*, 42:279–297, 2014.

[49] Edward Twitchell Hall. *The hidden dimension*. Doubleday & Co, 1966.

[50] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavioral markers. *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pages 1–33, 2012.

[51] Jordan Hashemi, Mariano Tepper, Thiago Vallin Spina, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, Helen Egger, Geraldine Dawson, and Guillermo Sapiro.

Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism Research and Treatment*, 2014.

[52] Roy S Hessels, Richard Andersson, Ignace TC Hooge, Marcus Nyström, and Chantal Kemner. Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633, 2015.

[53] Chris Plauché Johnson, Scott M Myers, et al. Identification and evaluation of children with autism spectrum disorders. *Pediatrics*, 120(5):1183–1215, 2007.

[54] Rebecca M. Jones, Audrey Southerland, Amarelle Hamo, Caroline Carberry, Chanel Bridges, Sarah Nay, Elizabeth Stubbs, Emily Komarow, Clay Washington, James M. Rehg, Catherine Lord, and Agata Rozga. Increased Eye Contact During Conversation Compared to Play in Children With Autism. *Journal of Autism and Developmental Disorders*, 0(0):0, 2016.

[55] Canan Karatekin. Eye tracking studies of normative and atypical development. *Developmental Review*, 27(3):283–348, 2007.

[56] Min Gyu Kim, Emilia Barakova, and Tino Lourens. Rapid prototyping framework for robot-assisted training of autistic children. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 353–358, 2014.

[57] Ami Klin, Warren Jones, Robert Schultz, Fred Volkmar, and Donald Cohen. Defining and quantifying the social phenotype in autism. *American Journal of Psychiatry*, 159(6):895–908, 2002.

[58] Ami Klin, Warren Jones, Robert Schultz, Fred Volkmar, and Donald Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives Of General Psychiatry*, 59(9):809, 2002.

[59] Ami Klin, David J Lin, Phillip Gorrindo, Gordon Ramsay, and Warren Jonas. Two-year-olds with autism orient to nonsocial contigencies rather than biological motion. *Nature*, 459(7244):257–261, 2009.

[60] Lauren A. Kryzak and Emily A. Jones. The effect of prompts within embedded circumscribed interests to teach initiating joint attention in children with autism spectrum disorders. *Journal of Developmental and Physical Disabilities*, 27(3):265–284, 2015.

[61] Olalekan Lanihun, Bernie Tiddeman, Elio Tuci, and Patricia Shaw. Improving active vision system categorization capability through histogram of oriented gradients. *Towards Autonomous Robotic Systems*, pages 143–148, 2015.

[62] Alain Legendre and Dominique Munchenbach. Two-to-three-year-old children's interactions with peers in child-care centres: Effects of spatial distance to caregivers. *Infant Behavior and Development*, 34(1):111–125, 2011.

[63] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[64] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223, 2000.

[65] Catherine Lord, Michael Rutter, Susan Goode, Jacquelyn Heemsbergen, Heather Jordan, Lynn Mawhood, and Eric Schopler. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2):185–212, 1989.

[66] Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5):659–685, 1994.

[67] Molly Losh, Patrick F Sullivan, Dimitri Trembath, and Joseph Piven. Current developments in the genetics of autism: from phenome to genome. *Journal of Neuropathology & Experimental Neurology*, 67(9):829–837, 2008.

[68] Rhiannon Luyster, Katherine Gotham, Whitney Guthrie, Mia Coffing, Rachel Petrak, Karen Pierce, Somer Bishop, Amy Esler, Vanessa Hus, Rosalind Oti, Jennifer Richler, Susan Risi, and Catherine Lord. The autism diagnostic observation schedule-toddler module: a new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(9):1305–1320, 2009.

[69] Marie Manner, Ming Jiang, Qi Zhao, Maria Gini, and Jed Elison. Determining child orientation from overhead video: A multiple kernel learning approach. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3477–3482, 10 2017.

[70] Marie D Manner. Using small humanoid robots to detect autism in toddlers. *Proceedings of International Joint Conferences on Artificial Intelligence*, pages 4383–4384, 2015.

[71] Marie D. Manner, Jed Elison, and Maria Gini. Leveraging computer vision and humanoid robots to detect autism in toddlers. In *Workshop at the 25th International Joint Conference on Artificial Intelligence*, IJCAI'16. IJCAI Press, 2016.

[72] Natasha Marrus, Anne L. Glowinski, Theodore Jacob, Ami Klin, Warren Jones, Caroline E. Drain, Kieran E. Holzhauer, Vaishnavi Hariprasad, Rob T. Fitzgerald,

Erika L. Mortenson, Sayli M. Sant, Lyndsey Cole, Satchel A. Siegel, Yi Zhang, Arpana Agrawal, Andrew Heath, , and John N. Constantino. Rapid video-referenced ratings of reciprocal social behavior in toddlers: a twin study. *Journal of child psychology and psychiatry*, 56(12):1338–46, 2015.

[73] Ross Mead, Amin Atrash, and Maja J Matarić. Automated proxemic feature extraction and behavior recognition: applications in human-robot interaction. *International Journal of Social Robotics*, 5(3):367–378, 2013.

[74] Eileen M Mullen and Others. *Mullen scales of early learning*. AGS Circle Pines, MN, 1995.

[75] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 331–338. ACM, 2011.

[76] U.S. Department of Health and Human Services. Autism spectrum disorder (ASD), 2017.

[77] Roald A Øien, Synnve Schjølberg, Fred R Volkmar, Frederick Shic, Domenic V Cicchetti, Anders Nordahl-Hansen, Nina Stenberg, Mady Hornig, Alexandra Havdahl, Anne-Siri Øyen, et al. Clinical features of children with autism who passed 18-month screening. *Pediatrics*, 141(6):e20173596, 2018.

[78] Andreea Peca, Ramona Simut, Sebastian Pintea, Cristina Costescu, and Bram Vanderborght. How do typically developing children and children with autism perceive different social robots ? *Computers in Human Behavior*, 41:268–277, 2014.

[79] Paola Pennisi, Alessandro Tonacci, Gennaro Tartarisco, Lucia Billeci, Liliana Ruta, Sebastiano Gangemi, and Giovanni Pioggia. Autism and social robotics: A systematic review. *Autism Research*, 9(2):165–183, 2016.

[80] Paola Pennisi, Alessandro Tonacci, Gennaro Tartarisco, Lucia Billeci, Liliana Ruta, Sebastiano Gangemi, and Giovanni Pioggia. Autism and social robotics: A systematic review. *Autism Research*, 9(2):165–183, 2016.

[81] John R. Pruett and Daniel J. Povinelli. Commentary - autism spectrum disorder: Spectrum or cluster? *Autism Research*, 9(12):1237–1240, 2016.

[82] Shyam Rajagopalan, Abhinav Dhall, and Roland Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 755–761, 2013.

[83] Melinda Randall, Kristine J Egberts, Aarti Samtani, Rob JPM Scholten, Lotty Hooft, Nuala Livingstone, Katy Sterling-Levis, Susan Woolfenden, and Katrina Williams. Diagnostic tests for autism spectrum disorder (ASD) in preschool children. *Cochrane Database of Systematic Reviews*, 7, 2018.

[84] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children's social behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3414–3421, 2013.

[85] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):11371149, Jun 2017.

[86] Nadja Richter. *The use of physical self-similarity in social assortment in children*. PhD thesis, Leipzig University, 2014.

[87] Nadja Richter, Kirsten Dalrymple, Olivia Engel, Samantha Kopf, and Jed Elison. Young children spontaneously initiate social interaction in response to an unsolvable task. Society for Research in Child Development, 2017.

[88] Daniel J. Ricks and Mark B. Colton. Trends and considerations in robot-assisted autism therapy. *IEEE International Conference on Robotics and Automation*, pages 4354–4359, 2010.

[89] Ben Robins and Kerstin Dautenhahn. The role of the experimenter in HRI research - a case study evaluation of children with autism interacting with a robotic toy. *The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 44(0):646–651, 2006.

[90] Ben Robins and Kerstin Dautenhahn. Tactile interactions with a humanoid robot: novel play scenario implementations with children with autism. *International Journal of Social Robotics*, 6(3):397–415, 2014.

[91] Ben Robins, Kerstin Dautenhahn, and Paul Dickerson. From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot. In *Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 205–211. IEEE, 2009.

[92] Ben Robins, Kerstin Dautenhahn, and Janek Dubowski. Robots as isolators or mediators for children with autism? a cautionary tale. In *Proceedings of the AISB 2005 Symposium on Robot Companions*, pages 82–88. AISB, 2005.

[93] Ben Robins, Kerstin Dautenhahn, and Janek Dubowski. Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies*, 7(3):509–542, 2006.

[94] Diana L Robins, Deborah Fein, Marianne L Barton, and James A Green. The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 31(2):131–144, 2001.

[95] Adrian Rosebrock. *Deep Learning for Computer Vision with Python*, volume 3. PyImageSearch, 1.3.0 edition, 2017.

[96] Noah J Sasson and Jed T Elison. Eye tracking young children with autism. *Journal of visualized experiments: JoVE*, 61, 2012.

[97] Brian Scassellati. Quantitative metrics of social response for autism diagnosis. In *IEEE International Workshop on Robot and Human Interactive Communication*, volume 2005, pages 585–590, 2005.

[98] Brian Scassellati. How social robots will help us to diagnose, treat, and understand autism. *Robotics Research*, 28:552–563, 2007.

[99] Brian Scassellati, Maja J Matarić, and Henny Admoni. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14(1):275–294, 2012.

[100] Eric Schopler, Robert Jay Reichler, and Barbara Rochen Renner. *The Childhood Autism Rating Scale (CARS): For diagnostic screening and classification of autism*. Irvington New York, 1986.

[101] Syamimi Shamsuddin, Hanafiah Yussof, Luthffi Idzhar Ismail, Salina Mohamed, Fazah Akhtar Hanapiah, and Nur Ismarrubie Zahari. Initial response in HRI- a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot NAO. *Procedia Engineering*, 41(Iris):1448–1455, 2012.

[102] Frederick Shic, W Jones, A Klin, and Brian Scassellati. Swimming in the underlying stream: Computational models of gaze in a comparative behavioral analysis of autism. In *28th Annual Conference of the Cognitive Science Society*, volume 28, pages 780–785, 2006.

[103] Elaine Short, Katelyn Swift-Spong, Jillian Greczek, Aditi Ramachandran, Alexandru Litoiu, Elena Corina Grigore, David J Feil-Seifer, Samuel Shuster, Jin Joo Lee,

Shaobo Huang, Svetlana Levonisova, Sarah Litz, Jamy Li, Gisele Ragusa, Donna Spruijt-Metz, Maja J Matarić, and Brian Scassellati. How to train your dragonbot: socially assistive robots for teaching children about nutrition through play. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 924–929, 2014.

[104] Ravishankar Sivalingam, Anoop Cherian, Joshua Fasching, Nicholas Walczak, Nathaniel Bird, Vassilios Morellas, Barbara Murphy, Kathryn Cullen, Kelvin Lim, Guillermo Sapiro, and Nikolaos Papanikolopoulos. A multi-sensor visual tracking system for behavior monitoring of at-risk children. In *IEEE International Conference on Robotics and Automation*, pages 1345–1350. IEEE, 2012.

[105] David Skuse, Richard Warrington, Dorothy Bishop, Uttom Chowdhury, Jennifer Lau, William Mandy, and Maurice Place. The developmental, dimensional and diagnostic interview (3di): A novel computerized assessment for autism spectrum disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(5):548–558, 2004.

[106] Leann E Smith, Jinkuk Hong, Marsha Mailick Seltzer, Jan S Greenberg, David M Almeida, and Somer L Bishop. Daily experiences among mothers of adolescents and adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 40(2):167–178, 2010.

[107] Sara S Sparrow, David A Balla, Domenic V Cicchetti, Patti L Harrison, and Edgar A Doll. *Vineland adaptive behavior scales*. American Guidance Service Circle Pines, MN, 1984.

[108] Wendy L Stone, Elaine E Coonrod, and Opal Y Ousley. Brief report: screening tool for autism in two-year-olds (stat): development and preliminary data. *Journal of Autism and Developmental Disorders*, 30(6):607–612, 2000.

[109] P. Szatmari, M. B. Jones, L. Zwaigenbaum, and J. E. MacLean. Genetics of autism: Overview and new directions. *Journal of Autism and Developmental Disorders*, 28(5):351–368, 1998.

[110] Deniz Tahiroglu, Louis J. Moses, Stephanie M. Carlson, Caitlin EV Mahy, Eric L. Olofson, and Mark A. Sabbagh. The childrens social understanding scale: Construction and validation of a parent-report measure for assessing individual differences in childrens theories of mind. *Developmental psychology*, 50(11):2485, 2014.

[111] Michael Villano, Charles R Crowell, Kristin Wier, Karen Tang, Brynn Thomas, Nicole Shea, Lauren M Schmitt, and Joshua J Diehl. DOMER: a wizard of oz interface for using interactive robots to scaffold social skills for children with autism spectrum disorders. *Proceedings of the 6th International Conference on Human-Robot Interaction*, page 279, 2011.

[112] Pearson VUE. Mullen sample report. https://www.pearsonschool.com/index.cfm?locator=PSZ3Mo.

[113] Pearson VUE. Vineland sample report. https://images.pearsonclinical.com/images/Assets/vineland-3/Vineland-3Domain-LevelTeacherFormSampleReport.pdf.

[114] Joshua Wainer, Kerstin Dautenhahn, Ben Robins, and Farshid Amirabdollahian. A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with autism. *International Journal of Social Robotics*, 6(1):45–65, 2014.

[115] Joshua Wainer, Ben Robins, Farshid Amirabdollahian, and Kerstin Dautenhahn. Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 6(3):183–199, 2014.

[116] Nicholas Walczak, Joshua Fasching, William D. Toczyski, Vassilios Morellas, Guillermo Sapiro, and Nikolaos Papanikolopoulos. Locating occupants in preschool classrooms using a multiple RGB-D sensor system. *International Conference on Intelligent Robots and Systems*, pages 2166–2172, 2013.

[117] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616, 2015.

[118] Zachary E Warren, Zhi Zheng, Amy R Swanson, Esubalew Bekele, Lian Zhang, Julie A Crittendon, Amy F Weitlauf, and Nilanjan Sarkar. Can robotic interaction improve joint attention skills? *Journal of Autism and Developmental Disorders*, 45(11):3726–3734, 2015.

[119] Sam V Wass, Tim J Smith, and Mark H Johnson. Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1):229–250, 2013.

[120] Amy M Wetherby and Barry M Prizant. *Communication and symbolic behavior scales: Developmental profile*. Paul H Brookes Publishing, 2002.

[121] Lorna Wing, Susan R. Leekam, Sarah J. Libby, Judith Gould, and Michael Larcombe. The diagnostic interview for social and communication disorders: background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry*, 43(3):307–325, 2002.

[122] Jason J Wolff, Brian A Boyd, and Jed T Elison. A quantitative measure of restricted and repetitive behaviors for early childhood. *Journal of neurodevelopmental disorders*, 8(1):27, 2016.

[123] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28, 2014.

[124] Tian (Linger) Xu, Hui Zhang, and Chen Yu. See you see me: the role of eye contact in multimodal human-robot interaction. *ACM Transactions on Interactive Intelligent Systems*, 6(1):1–22, 2016.

[125] Robyn L Young, Neil Brewer, and Clare Pattison. Parental identification of early behavioural abnormalities in children with autistic disorder. *Autism*, 7(2):125–143, 2003.

[126] Lonnie Zwaigenbaum, Susan Bryson, Tracey Rogers, Wendy Roberts, Jessica Brian, and Peter Szatmari. Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, 23(2-3 SPEC. ISS.):143–152, 2005.