Machine-assisted Technologies for Young Children with Autism Spectrum Disorder: Novel Platforms for Early Detection and Intervention

By
Zhi Zheng

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Electrical Engineering
December, 2016
Nashville, Tennessee

Approved:

Nilanjan Sarkar, Ph.D.
Zachary E. Warren, Ph.D.
Gabor Karsai, Ph.D.
Robert J. Webster III, Ph.D.
D. Mitchell Wilkes, Ph.D.
Amy S. Weitlauf, Ph.D.

The presented work is dedicated to

my coming baby,

my husband Kai Ni,

and my parents Jinbo Zheng and Shuling Duan.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor, Dr. Nilanjan Sarkar, for his invaluable guidance and support during my Ph.D. study. His continued advice in every aspect of my study was fundamental in building me as a professional researcher and engineer. I would also like to thank Dr. Zachary Warren and Dr. Amy Weitlauf. They helped me conduct meaningful interdisciplinary research in technology, psychology, and pediatrics. Furthermore, I would like to express my appreciation to other members of my dissertation committee: Dr. Gabor Karsai, Dr. Robert Webster, and Dr. Mitch Wilkes, for their insight and feedback that improved this dissertation work significantly.

Second, I would like to acknowledge Ms. Amy Swanson and Dr. Julie Crittendon at the Vanderbilt Kennedy Center Treatment and Research Institute of Autism Disorder (TRIAD). Their continuous help on user study design, data analysis, and writing was essential in this work. I also much acknowledge my colleagues, Joshua Wade, Esube Bekele, Eric Young, Shuvajit Das, Lian Zhang, Huan Zhao, Jing Fan, Dayi Bian, Guangtao Nie, Zhaobo Zheng, and many others from the Robotics and Autonomous Systems Lab who provided consistent encouragement, dedicated collaboration, and resourceful conversation. The collaboration with these friends was one of my best experiences in these years. I would like to thank Nicole Bardett, Rebecca Johnston, and Amy Nicolson at the TRIAD for their professional help in the user studies.

Finally, I would like to thank my family. My parents, Jinbo Zheng and Shuling Duan, have encouraged me so much in my struggling days, and taught me to be strong. They also reminded me to be humble and stay thirsty for knowledge in my bright days. My husband, Kai Ni, has always been there with me whenever I needed him. My family provided me the strongest support during this long journey.

This dissertation work have been supported in part by the Vanderbilt University Innovation and Discovery in Engineering and Science (IDEAS) grant, a grant from the Vanderbilt Kennedy Center (Hobbs Grant), the National Science Foundation under Grant 0967170 and 1264462, and the National Institute of Health under Grant 1R01MH091102-01A1 1R21MH103518-01, and 5R21MH103518-02. Work also includes core support from NICHD (P30HD15052) and NCATS (UL1TR000445-06).

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Chapter I. Introduction

1.1 Autism Spectrum Disorder and Technology

Autism spectrum disorder (ASD) is a common disorder associated with enormous individual, familial, and social cost across the lifespan [1]. Its core deficits include impairments in social communication as well as repetitive and atypical patterns of behavior [2]. According to the Centers for Disease Control and Prevention (CDC), an estimated 1 in 68 children in the United States have ASD [3]. The cumulative ASD literature suggests earlier and more intensive behavioral interventions are efficacious for many children [4]. However, many families and service systems struggle to provide intensive and comprehensive evidence-based early intervention due to extreme resource limitations [5, 6]. The average lifetime cost of care for individuals with autism is estimated to be around $3.2 million [1] with associated annual care costs estimated to exceed $35 billion in the United States [7]. In conjunction with the individual, familial, and societal impact associated with ASD, these alarming figures underscore that effective identification and treatment of ASD is a public health emergency [8]. As such, there is an urgent need for more efficacious and less expensive treatments whose realistic application will yield more substantial impact on the neurodevelopmental trajectories of young children with ASD within resource strained environments. It is within this context, we propose the design and development of machine-assisted intervention technologies for three core impairments of ASD.

The first core impairment that we have attempted to address is the imitation skill. Imitation involves translating from the perspective of another individual to oneself, and creating representation of this individual's primary representation of the world [9]. Although the exact reasons of the imitation impairment associated with ASD are still unclear, evidence suggests that this imitation impairment may be related with the basic capability to map actions of others onto an imitative match by oneself [10]. Imitation is a critically important social communication skill that emerges early in life and it is recognized to play an important role in the development of cognitive, language, and social skills [11]. Children with ASD show powerful impairments in imitation and such deficits have been tied to a host of associated neurodevelopmental and learning challenges over time [12].

The second core impairment that we have looked at is the social orienting skill. Among multiple aspects of social communication development, social orienting is one of the most fundamental and critical skills that naturally develops in children [13]. Social orienting indicates spontaneous orientation to naturally occurring social stimuli in one's environment [14], which is closely related to other important social communication skills such as joint attention. Unfortunately, children with ASD usually show powerful deficits in this development. My research focuses on Response to Name (RTN), which is an important social orienting skill. RTN, as the name suggests, is a task that assesses how a child responds when his name is called. A decreased tendency to RTN is one of the most sensitive and specific predictors of whether an infant will later be diagnosed with ASD when he is old enough for a definitive diagnosis (typically 24 months or later) [15, 16]. RTN is also a key measurement of the standard ASD diagnostic assessment such as the Autism Diagnostic Observation Schedule (ADOS) [17].

The third impairment that we have investigated is the joint attention skill. Joint attention skills are thought to be fundamental, or pivotal, social communication building blocks that are central to the etiology and treatment of ASD [18, 19]. At a basic level, joint attention refers to the development of specific skills that involve sharing attention with others (e.g., pointing, showing objects, and coordinating gaze). These exchanges enable young children to socially coordinate their attention with other people to more effectively learn from their environments. Fundamental differences in early joint attention skills have been demonstrated to underlie the deleterious neurodevelopmental cascade of the disorder and successful treatment of these deficits has been demonstrated to substantially improve numerous developmental skills across settings [19-21].

In recent years, researchers have proposed advanced technologies, including robotic systems [22, 23], virtual reality environments [24, 25], interaction games [26, 27], etc., as potential solutions for addressing these limits of ASD intervention. Despite hypothesized theoretical benefits of such applications, major challenges exist regarding implementing such systems. They must be (1) capable of robust autonomous functioning, (2) relevant and important to the core features of ASD at appropriate points in development, and (3) potentially realistic as cost-effective intervention systems outside of highly specialized research environments. In an effort to address these challenges, my research mainly focuses on robot-mediated and computer-assisted interventions for children with ASD.

Researchers have concluded that adaptive autonomous technology has the potential for improving social communication abilities for children with ASD [23, 28]. However, only a few studies of adaptive technological and robotic interaction with children with ASD have appeared in the literature: proximity-based closed-loop robotic interaction [29, 30], haptic interaction [31], adaptive robot-assisted play [32], video-game responses to physiological signals [27], and turn-taking imitation interactions in school-aged children [30]. Although all of these works described robust systems for adaptive interaction, the paradigms explored had limited direct relevance to the core deficits of ASD at young ages and instead focused on proof-of-concept task and game performance or school-aged children. In this chapter, we introduce the background and development status of robot-mediated and computer-assisted intervention for children with ASD and a summary of my research.

1.2 Human-machine Interaction for Children with ASD

Human-machine interaction is the interaction and communication between a human user and a machine via a human-machine interface. The “machine” usually indicates a system with a certain level of ability which provides the user with proper feedback. These systems can be built with a robot [33], a computer [34], or a smart phone [35]. This work mainly focused on robot-mediated and computer-assisted intervention. The background of these two research topics are introduced below.

1.2.1 Robot-mediated intervention for children with ASD

Robotic technology is gaining momentum as an intervention platform for children with ASD. Since 1976, when Weir and Emanuel [36] found that robots could improve social interaction for children with ASD, a plethora of works have been published that demonstrated the potential and promise of intervention based on systems for children with ASD. Such work includes exploring the response of children with ASD to robot-like characteristics; eliciting specific behaviors; modeling, teaching, and practicing skills; as well as providing feedback and encouragement during interactions [22, 37]. Robots have several advantages over traditional human-led interactions, including precise control of intervention modality, robust consistency, simplified features, autonomous operation, and potential cost effectiveness. Initial results obtained from applying robotic technology to ASD intervention have consistently shown a unique potential to elicit interest and attention in young children with ASD [23, 38, 39]. The emerging robotic and technological literature has demonstrated that many individuals with ASD show a preference for robot-like characters over non-robotic toys [40, 41] and in some circumstances even responded faster when cued by robotic movement rather than human movement [42, 43]. Although most of these researches have focused on school-aged children and adults, the downward extension of this preference for robotic and technological stimuli is promising. Many very young children with and at-risk for ASD often preferentially orient to nonsocial contingencies, videos, and arrays rather than biological motion or video [44, 45]. Further, a number of studies have indicated the advantages of robotic systems over animated computer characters for skill learning and optimal engagement. This is likely due to the capability of robotic systems to utilize physical motion in a manner not possible in screen-based technologies [46, 47].

There is significant heterogeneity in studies conducted to date regarding sample size, interaction type, and ages of participants that creates challenges when summarizing the literature. The age of participants spreads from preschoolers [48] to teenagers [42]. The user group ranges in size from one [49] to dozens [50, 51], and to hundreds [52]. The interaction patterns mainly include free interactions [30] as well as task specific interactions [29, 53]. There were both longitudinal, multi-session studies [49] and short term, single session studies. The results in many multi-session studies showed that the children's performance and attention on the robot progressed across sessions [49, 54]. Because ASD is four times more prevalent in males than females, the majority of participants were males, with some exceptions that included equal numbers of both sexes [42]. Although most studies conclude that their findings need more extensive testing in the future, this did not often happen, as testing beyond a pilot study is a time, cost and labor intensive process.

1.2.1.1 Robotic platform development

There are three main categories of robots that have been utilized to interact with children with ASD. The first category is a traditional machine-like robot. Costa et al. [49] used a LEGO robot to help a child with ASD learn how to share objects and fulfill orders. Pierno et al. [42] used a robot arm to study the imitation behavior of children with ASD. Liu et al. [55] developed an interactive basketball playing robot, which could adapt its behavior based on the participant's emotional state. However, none of these robots resembled living creatures.

The second category is an animal-like robot. Several studies have found that robots with animal features can elicit social behaviors from children with ASD. For example, Stanton et al. [56] tested the interaction between a robotic dog AIBO and children with ASD. The results showed that the participants were engaged and showed fewer autism symptoms such as verbal engagement and reciprocal interaction. Dickstein-Fischer et al.[57] built a robotic penguin for the same purpose. Kim et al.[58] studied how a robotic dinosaur Pleo could elicit social behaviors form children with ASD. Even though the outward appearances were very different, each of these different animalistic designs showed promise for social intervention purposes.

The third category is a humanoid robot, which currently is the most widely used robot for ASD related studies with robots. Even though the appearances of most of the humanoid robots are kept relatively simpler than a real person, their functionalities have been dramatically improved in recent years. Pioggia et al. [59] developed a robotic face (called FACE) to interact with participants based on their facial expressions, body gestures and psychophysical signals. Goodrich et al. [60] developed a robot with two arms and a flat screen face, which presented different facial expressions. Kozima et al. [61] developed an upper body humanoid robot, Infanoid, to investigate its ability to affect social intentionality, identification and communication as part of ASD intervention. Some researchers created full body robots, as well. Fujimoto et al. [62] used a full body humanoid robot to teach children with ASD to imitate arm gestural skills. Amirabdollahian et al. [31], Wainer et al. [63], and Dautenhahn et, al. [64] developed a full body child-like robot, KASPAR, which is capable of eliciting different social communication behaviors such as joint attention, imitation, tactile exploring, and collaborative game playing. Still other researchers have used full body robots within modifiable environments, such as Feil-Seifer et al. [30] who set a humanoid robot on top of a mobile robot and used it to measure children's interaction pattern with respect to distance-based features.

Each of these robots was developed within research labs. However, there are also commercial robotic platforms that have been broadly applied to work with children with ASD, such as the full body humanoid robots NAO [29, 65] and Zeno [66, 67]. An advantage of using commercial robots is that a design based on them can be easily reproduced and improved by different groups. However, a general robot platform may not be as flexible and cost effective as a task-customized robot developed as part of research.

During a robot-mediated intervention, the robot usually serves as a game mediator and promoter [49], and many functionalities are targeted as such. For example, a primary parameter when designing a robot for ASD intervention is the motion flexibility, which depends on the intervention needs (and the robot's role). A simple robot with only a few degrees of freedom such as "Keepon" [68] is enough to catch children's attention, while a more complex robot like NAO [54] is necessary for explicit imitation intervention. Another parameter is the need to include other peripheral sensing technologies to track the participant's behavior. For example, Boccanfuso et al. developed a doll-like robot CHARLIE [69] with computer vision-based hand and face tracking functions. Chuah et al. developed a robot LILI [70] that had embedded functionalities including gesture recognition and speech recognition. Finally, Ravindra et al. [71] developed a robot with gaze tracking function for teaching joint attention skills for children with ASD.

As seen from the above discussion, given the wide range of designs, functions, and capabilities, it is very important to consider a robot's role within an intervention when deciding on its ability to move, catch and direct a child's attention, and gather data about the children with ASD they are helping to treat. It may be more appropriate for certain features to be important in some intervention designs (e.g., the prominence of robot FACE for promoting facial expressions) than others (e.g., the hand-tracking abilities of robot CHARLIE). Although all of these technologies have their own limitations in terms of detection accuracy, robustness, and detection range, they represent the general movement of the field toward ideal robotic platforms for different environments and treatment goals.

1.2.1.2 Patterns of robot-children interactions

A primary question raised regarding using robotic technology for children with ASD is whether the children will accept the robots and interact with them as opposed to humans, due to the differences in the robots' physical appearance, voice, and behavior. In 2005, Robins et al. [72] conducted a study where the same person interacted with children with ASD in two ways, first while dressing and acting like a robot and then while dressing and acting normally. Similar to other works [40, 54, 58], their results showed that children with ASD interacted more with the person when he had a simplified robotic appearance. Indeed, children with ASD are interested in robots with a simple appearance. For example, a very simple small yellow snowman-like robot "Keepon", developed by Kozima et al. [68] was successful in eliciting positive responses from children with ASD. The most obvious appearance features of Keepon are simply an elastic body, two eyes and one nose. However, the robot successfully elicited positive responses from children with ASD such as playful behaviors with Keepon and induced relaxed mood. Even today, the design principle for robot appearance in many studies still follows the principle of simplicity [73].

Currently, data collection on social interactions for robot-mediated intervention is performed using a combination of manual video coding as well as by autonomous methods within the system. In the early days of robotic intervention work, most studies were forced to utilize manual video coding to analyze data, most of which consisted of gaze pattern, tactile pattern, gestures, speech and other interaction specific behaviors [31, 42, 49, 56]. Later on, as autonomous technologies have improved dramatically, more and more children's behaviors and responses can be tracked and recorded automatically. This laid the groundwork for the development of closed-loop, adaptive interaction systems using increasingly complex, integrated technologies. This includes tracking participants' movements within the test environment. For example, in Pierno et al. [42], the participant's arm motion was recorded in 3D space by using infrared cameras that tracked passive markers put on the participants' arms. Bekele et al. [65] applied a camera based head tracking system to approximate children's gaze direction. Greczek et al. [29] used Microsoft Kinect to track the motion of the participants.

Both manual and automated data recording methods have their own advantages and disadvantages. The manual video coding method can be used for any visually measurable parameter that cannot be effectively and automatically detected by current technology. However, video coding is impacted by the personal bias of the coder, and it is also labor and time intensive. Technology based data recording methods reduce labor time and costs, and avoid personal biases, such as the coder's attention and habits. However, the signals that can be detected automatically are still limited. For instance, camera based methods are sensitive to illumination and occlusion. Limited tracking rules may not be flexible enough to accommodate different types of participants and for complex interaction recognition. Even so, the development of these technologies represents a primary push for future research trends. Based upon recent advancement, we are optimistic that more and more methods will be developed to ameliorate these current limitations.

Despite this hypothesized advantage, there have actually been relatively few systematic and adequately controlled applications of robotic technology to investigate the impact of directed intervention and feedback approaches [58, 74-76]. There is a need for robotic systems in terms of application to intervention settings necessitating extended and meaningful adaptive interactions. We believe that adaptive interaction, operationalized as within system changes in response to measured behaviors, is important for individualization of intervention and ultimately addressing core deficits of ASD [77]. Only a few adaptive robotic interaction works with children with ASD have appeared in the literature: proximity-based closed-loop robotic interaction [30], haptic interaction [78], and closed-loop adaptive interaction work based on affective cues inferred from physiological signals [79]. While all of these works were able to put forth robust systems for adaptive interaction, the paradigms explored had very little direct relevance to the core deficits of ASD in that they are focused on simple task and game performance. Recent work has suggested that robotic intervention systems designed to operate in "closed-loop" form [80] and targeted to early pivotal skills for children with ASD [81] may represent more promising paradigms for extended and meaningful interaction. The development of adaptive interaction realizing within-system changes in response to detected and measured behaviors is extremely important for individualization of meaningful technological interventions, as effective treatment of young children with ASD often requires extended and meaningful adaptive interactions [82].

Therefore, developing autonomous closed-loop robotic systems that can adaptively adjust their behavior based on the children's response in a core-deficit related intervention is one of the most important goals in my research. The robotic systems built in my research are based on principles inspired by Scassellati et al. [23] that a robot applied in interventions should possess three important features: 1) the robot must be either controlled remotely or programmed to autonomously observe physical behavior; 2) the robot must know when to begin and sense the child's response with sufficient accuracy, and 3) the robot must be able to map those response to its own, potentially limited, effectors in order to replicate the behavior as closely as possible, in a recognizable fashion.

The content of this section in inherited from a book chapter I composed with my advisors and collogues [83].

1.2.2 Extended computer-assisted intervention for children with ASD

Besides robots, computer-assisted intervention is another important type of technological application on children with ASD. In general, this application includes any intervention conveyed by one or more computers. Human-computer interaction has been widely applied as an assistive technology for neurobehavioral studies, especially for reinforcing desirable behaviors [84-87]. Children with ASD usually show impairments in sensory perception, which may lead to difficulties in discriminating and screening out unnecessary information from overall meaning [88-90]. Computer-assisted technology has the advantage to be designed in a way that only primary information is presented to the children. Furthermore, studies have shown that when stimuli are more predictable, the responsiveness of children with ASD increases accordingly [91]. Similar to robotic technology, computer-assisted technology is capable of providing high controllability, precision, consistency, and robustness. This advantage is one of the main reasons why computer-assisted intervention is attractive in ASD research. Computer-assisted intervention for children with ASD can be traced back to decades ago. A representative work was conducted in 1973 by Colby [92], where computer programs were used to stimulate language development in children with ASD. In this work, the participants pressed letters on a keyboard, and then audio as well as animations associated with the letter would show up to mimic spontaneous language acquisition. Since then, a great number of works have explored computer-assisted intervention for children with ASD. Most of these designs took the form of games or educational tools to enhance the engagement of participants.

1.2.2.1 Conventional computer-assisted intervention

Early works mainly emphasized on multimedia when it became popular in computer applications. Williams et al. [88] compared computer-assisted reading (reading material presented on computer screen) with traditional book reading by children with ASD. Eight children aged 3-5 years participated in an experiment for 10 weeks, where they were randomly allocated to the computer or book condition. The results showed that the participants were less resistant to reading and spent more time on the reading task in the computer condition than that in the book condition. Heimann et al. [93] studied the effect of using Alpha, which was an interactive and child-initiated computer program, on helping children with ASD with reading and communication skills. This computer program was used by the students as supplementary materials to their regular reading and writing activities. This program was used to practice individual words, create sentences, attest words and test sentences. A user study with 11 children with ASD showed that using Alpha significantly improved their word reading and phonological awareness. Hagiwara and Myles [94] investigated the effects of a computer-based story intervention on youth with ASD. The author claimed that this was the first attempt to implement a multimedia social story intervention. The multimedia social story computer programs contained text of social stories, movies of the participants' actions corresponding to social story sentences, audio that read aloud sentences using a synthesized computer voice, and a navigational clickable button. Even though these computer programs and interaction activities looked fairly simple, these pioneering works indicated great potential of using computer as an effective intervention tool.

Along with the development of computer technology, computer was increasingly accepted by people both as part of their daily life and a potential tool for neural rehabilitation. New technologies also offered more possibility and complex functionality for ASD intervention. For example, Hetzroni and Tannous [95] investigated a computer-assisted intervention for enhancing communication skills in children with ASD. This work applied a software that provided conversational interaction simulation regarding play, food, and hygiene. Each of the tree topics included short animations as well as question asking and answering. Five school-aged children with ASD tried the software. The user study results showed that the practice in this controlled and structured interaction provided by the software had a great potential to teach new knowledge to children with ASD. Most importantly, they could transfer this knowledge to the natural classroom environment. Passerino et al. [96] conducted a 3.5 years multi-case longitudinal user study on the impact of a digital learning environment, named Eduquito, for students with ASD. Eduquito were used for mail, chat, bulletin board, forum, collaborative stories, etc. The results showed that, after the longitudinal usage of the digital learning environment, the participants improved in their communication levels of autonomy, self-regulation, indirect self-control, and social interaction. The author claimed that the controllable levels of complexity which allowed the system to be adjusted based on the need of each subject contributed primarily to this positive result. Silver and Oakes [97] introduced a computer program for teaching children with ASD recognizing and predicting emotional responses in others. Twenty-two children participated in a user study where they practiced recognizing, correlating, and predicting pictures of facial expressions, verbal description and pictures correlated to different emotions on a computer screen. Results showed improvement in task performance.

1.2.2.2 Computer-assisted intervention based on virtual reality techniques

In recent years, virtual reality (VR) based intervention emerged as a particularly promising new technology for children with ASD, where interactive environment and peers are created to help children with ASD build communication skills. VR (or virtual environment) is a computer generated 3D simulation of real or imaginary environment [98]. A major part of VR based studies for children with ASD put emphasis on simulating real-life scenarios. For instance, Wallace et al. [99] investigated whether an immersive VR environment could simulate ecologically valid situation. Ten children with ASD and 14 typically developing (TD) adolescents participated in a user study, where each of the participant experienced a virtual residential street, a school playground, and school corridor scenes. Each of these scenes was embedded with realistic acoustical background and avatars. Results showed that the children with ASD had similar levels of presence compared with TD peers and did not have negative sensory experiences. The immersive VR environment was realistic enough to simulate authentic social situation. Mineo et al. [100] studied the engagement potential of animated video, video of self, video of a familiar person engaged with an immersive VR, and immersion of self in the VR game. Fourty-two participants' engagement was measured by gaze duration and vocalization. In general, results showed that all of these stimuli held students' visual attention well, at least for a short period of time. In some cases, seeing oneself in the stimuli generated greater gaze duration as compared to seeing another person. Besides, the VR stimulated more vocalization than the traditional video sessions.

A few studies applied avatar as interaction cues. Tartaro and Cassell [101] developed a 3D life size animated virtual peer for children with ASD which was language enabled and could share toys with the children and respond to children's input. This virtual figure, 'Sam', interacted with children by using eye-gaze, body and head postures, and speech to negotiate turns. The behavior of Sam was controlled by a system operator or the child who interacted with him. Improved engagement and social communication behaviors were stimulated during the interaction. Bernardini et al. [102] developed ECHOES, which was an intelligent serious game for practicing social communication skills in children with ASD. ECHOES used a 3D intelligent virtual character as an interaction agent. In different computer simulated social situations, the avatar acted as a peer or a tutor and interacted with the children with the help a 2D sensory garden. Twenty-nine children with ASD tried ECHOES, and the results showed that ECHOES has a potential to be applied in real-world and school environment for teaching communication skills to children with ASD.

There were also other VR studies oriented toward particular skills. Moore et al. [103] and Bekele et al. [25] developed VR-based program to evaluate how children with ASD identify and make inferences from different facial expressions. In these two works, different facial expressions of emotions were displayed by avatars, such as happiness, sadness, and anger. The participant was required to identify the corresponding emotion based on the facial expressions. Patterns of the participants gaze were analyzed and compared with TD peers. The results provided quantitative description of the difference between the two groups. Wade et al. [104] developed VR driving simulations to help adolescences with ASD learn how to drive a car and analyze their driving behaviors. In this work, the participants operated a physical steering wheel, gas and brake pedals navigated through a virtual city. The participant's affective status and gaze information were monitored during the whole interaction. Results showed that the driving performance of the participants improved after multi-session interaction.

1.2.2.3 Discussion on computer-assisted intervention

As we can see from the examples listed above, the key interaction cues affiliated with computer-assisted technology for children with ASD include verbal communication, eye gaze, body gesture, emotion recognition, etc. A review article by Grynszpan et al. [105] evaluated innovative technology-based interventions of ASD. This article systematically summarized research that used pre-post design to assess the effect of computer programs, virtual reality, and robotics on children with ASD. The author concluded that in general, technology-based interventions significantly improved the participants' performance compared with that before intervention, and the mean effect size was medium. This summary revealed the importance of investigating machine-assisted interventions for children with ASD.

Although the works reviewed indicate successful application of computer-based intervention, these systems have several limitations, which led to our goal of designing an enhanced human-computer interaction method for young children with ASD to address the following isssues:

1) Many of the previous studies used Wizard of Oz method (for example, an experimenter controlled the behavior of avatars), where a human operator was needed during interaction. However, one of the ultimate goals of applying technology in ASD intervention is to reduce human labor and the corresponding cost. Thus developing autonomous closed-loop system which does not require Wizard of Oz operation is critical. The system developed shall be able to adapt its behavior based on the participant's performance in real-time.

2) Most of the existing studies are oriented towards school-age and/or older children. However, evidence suggests that early detection and intervention is critical to optimal treatment for ASD [106, 107], when brain is still malleable. Therefore, developing technology with proper functionality and interaction protocol for toddlers with ASD and even younger infants at risk of ASD will be of great benefit. Traditional computer-assisted intervention uses conventional interfaces such as a keyboard and a mouse. However, toddlers and infants are too young to operate these hardware. Besides, operating computer accessories is not typical in daily communication, either. Some VR-based interaction applied projected immersive environment which could over stimulate toddlers and infants, and some other studies used the head mounted display (HMD) which is usually too heavy and too big to be worn by young children. Thus extending interaction interface that is not limited to keyboard, mouse, over stimulated projection, and HMD is necessary.

3) Similar as robotic intervention systems, computer-assisted intervention systems provide objective quantitative measurements regarding the participants' behavior and performance. Compared with traditional methods (e.g. human observation and manual video coding), objective measures do not include personal bias introduced by human examiners, and thus are more reliable in many cases.

1.3 Summary of the Dissertation Research

This dissertation focuses on the intervention of three core deficits of ASD: 1) imitation skill; 2) social orienting skill; and 3) joint attention skill. This research addresses a few challenges in this field that prevent the development of fully autonomous intervention systems. These challenges are: 1) how to target the core deficits of ASD using machine-assisted technologies; 2) how to make the system adaptive based on children's real-time response; 3) how to detect interaction cues non-invasively; and 4) how to validate skill generalization from machine-assisted intervention to human-human interaction. Each of these studies is briefly introduced in this chapter, while the details of each study are discussed in Chapter II to Chapter VII.

1.3.1 Autonomous robot-mediated imitation intervention

As literature has demonstrated that many individuals with ASD show a preference for robot-like characteristics over non-robotic toys [40, 41] and in some circumstances even respond faster when cued by robotic movement rather than human movement [42, 108], we were interested in developing a fully autonomous robotic system to help teach children with ASD motor imitation skills. Specifically, we studied how children with ASD imitated a humanoid robot's arm gestures.

1.3.1.1 Single gesture imitation intervention

Chapter II is a comprehensive assemblage of two published papers [53, 109]. This chapter describes the development and initial application of a non-invasive intelligent robotic intervention system, Robot-mediated Imitation Skill Intervention Architecture 1 (RISIA1), which was capable of dynamic and individualized interaction with potential relevance to improve imitation skills for young children with ASD. RISIA1 was embedded within the robot NAO and a Microsoft Kinect [110]. The Kinect was placed in front of the participant to track his/her body movement, and the tracked data were sent to a newly designed FSM-based gesture recognition module for computing imitation performance in real time. Based on the performance of the participant, the robot could even recognize partially finished gestures and give feedback to the child accordingly.

Eight children with ASD and 8 typically developing (TD) children participated in a user study. Each participant had two human administered sub-sessions and two robot administered sub-sessions under an adaptive interactive protocol. Four different gestures, one gesture per session, were exhaustively tested in a randomized order. We compared participants' imitation performance and attention towards the robot/human administrator between the robot sessions and the human sessions.

In general, the group with ASD looked at the robot longer compared with that of the human therapist, while TD group paid similar attention to robot and human therapist. The group with ASD required a similar amount of time to complete the tasks in both robot and human sessions, while the TD group spent more time to complete the robot session. Given that impairments in imitation representing a core symptom of ASD, the TD group were more successful than the group with ASD across both conditions and did not demonstrate much difference in performance between the robot and human sessions. Within the group with ASD, children were far more successful imitating the target gestures in the robot session than that in the human session.

1.3.1.2 Mixed gesture imitation intervention

In children's daily life, activities such as playing a game, playing a musical instrument, and dancing require combination of simple gestures to accomplish a task. Mixed gesture was defined as simultaneous execution of multiple simple gestures from a participant. Chapter III is based on a published article [111] on mixed gesture detection. This chapter describes a new system, Robot-mediated Imitation Skill Intervention Architecture 2 (RISIA2), which extended RISIA1 by introducing a new Mixed Gesture Recognition and Spotting (MGRS) algorithm. MGRS is capable of detecting mixed gestures, as well as identifying ("spotting") the start and the end of each detected gesture. A new intervention protocol was also designed to test this algorithm. Under this protocol, a preliminary user study was conducted with children with ASD and their typically developing (TD) peers to show the feasibility and potential usefulness of RISIA2.

Four children with ASD and 2 TD children with different imitation abilities tested the accuracy of MGRS algorithm and the extended interaction design. The MGRS algorithm was validated by the gesture data collected from the participants. The gesture recognition and spotting results computed from MGRS was compared with human coded result on the same experimental data. The comparison showed that MGRS was accurate in both recognizing the type of the gestures performed and spotting the start and end time point of the performed gestures.

1.3.2 Autonomous computer-assisted social orienting intervention

Chapter IV was based on one published article [28] and one accepted article [112]. The primary objective of this chapter is to present a novel autonomous social orienting intervention system (ASOTS) with potential for aiding in both screening whether a child has such a deficit and improving his/her skill when needed. ASOTS was designed for social orienting skill intervention that can be adapted for various paradigms. In this chapter, we focus on an important social orienting skill, Response to Name (RTN), to demonstrate the usefulness of this novel system. RTN, as the name suggests, is a task that assesses how a child responds when his/her name is called.

ASOTS enables computer-based name calling from a wide range of angles around a participant by providing a distributed display mechanism, allows real-time attention inference of the participant through gaze tracking using a distributed array of cameras and offers an adaptive attention guiding mechanism to shape his/her response. The ultimate goal of the RTN intervention is to have the child successfully respond to caregivers' attempts to garner attention by calling his/her name from a variety of locations within the learning environment. This new system was tested via a set of experiments with 10 typically developing (TD) infants as well as 10 toddlers with ASD.

Experimental results showed that the gaze tracking method applied in ASOTS was accurate and successfully detected the attentional preference of the participant. The participants showed great engagement during the interaction and spent the majority of the session focusing on the interaction environment. The participants also successfully responded to the name calling in almost all the sessions, with the help of the attention attractor.

1.3.3 Semi-autonomous robot-mediated joint attention intervention

Bekele et al. [65] proposed an adaptive and individualized robot-mediated system, ARIA, for teaching joint attention skills to children with ASD. The system was composed of the humanoid robot NAO, with its vision augmented by a network of cameras for real-time head pose tracking. Based on the child's head movement, the robot intelligently adapted itself to generate prompts and reinforcements to promote joint attention skills. Results showed that the participant achieved a high target hit rate in robot administrated session. The participants with ASD also had a statistically significant preferential orientation towards the robot as compared to the human therapist. While ARIA was an autonomous system, only 60% of the participants with ASD could finish the user study. 75% of the participants who could not complete the study refused to wear a sensory hat which was necessary for head pose estimation. Furthermore, this study was a single session study and as such did not provide any indication on whether the children would respond similarly over multiple sessions.

Therefore, in order to further investigate the impact of robot-mediated joint attention intervention, we modified the ARIA architecture in Chapter V. The content of this chapter is based on a published article [54]. In this study, an eye tracker was embedded in the robotic system to monitor the gaze of the child on the robot non-invasively. A human therapist was involved in the loop to replace the hat and the camera system that determined when and how the child responded to robotic prompts. A longitudinal user study was conducted with 6 children with ASD. Each participant had 4 sessions on different dates. In each session, a similar interaction protocol to that in the work of ARIA was applied, where repeated trials were provided to the participant.

This small user study indicated that the children with ASD documented sustained interest with the humanoid robot NAO over several sessions and demonstrated improved performance within system regarding joint attention skills. From session 1 to session 4, the participants needed lower prompt levels to hit the target. Meanwhile, there was no significant difference on the time they spent on looking at the robot administrator.

1.3.4 Autonomous robot-mediated joint attention intervention

The small scale longitudinal user study discussed in Chapter V encouraged us to improve the system into a fully autonomous robot-mediated joint attention intervention system and conduct a larger scale longitudinal user study. The composition of Chapter VI is based on a submitted journal manuscript [113]. In this study, we developed a new non-invasive autonomous robot-mediated joint attention intervention system, named Norris. Norris inherited the advantages and eliminated the disadvantages of our previous works. Norris was embedded with a new non-invasive gaze-tracking method to form a close-loop interaction, and thus no invasive physical sensors or human operations were needed. The new system also worked autonomously following the Least-to-Most (LTM) interaction protocol. LTM is a widely applied methodology and is not limited to joint attention intervention. However, how to formally model LTM so that it can be easily implemented in a robotic system was still unanswered. Such a model, as a general guideline, will benefit the design and implementation of other robotic systems, regardless of which particular skill a system is designed for. Therefore, we proposed a LTM-based robot-mediated interaction (LTM-RI) model to solve this problem. In this chapter, LTM-RI is used to describe the interaction logic of Norris.

We tested Norris and LTM-RI in a pilot longitudinal user study with 14 young children with ASD. As the study discussed in Chapter V, each participant in the current study experienced 4 intervention sessions. The completion rate was 100%. We also found that the participants' initial interest on the robot held over the sessions, and their within-system joint attention skills improved significantly. The results also proved the effectiveness of the LTM-RI model: i.e., the higher the joint attention prompt level, the higher the probability that the participants hit the target; and the participants could hit the target eventually in almost all the trials, given the designed LTM-RI prompt hierarchy.

1.3.5 Exploration of the Generalization from Robot-mediated Joint Attention Intervention to Human-Human Interaction

Although the initial results in Chapter VI are promising, it is unclear whether the robot-mediated intervention has a positive impact on children with ASD in their daily interactions with humans. This is a fundamental question that needs to be answered, because their potential for generalization to human-human interactions reflects the ultimate value of robot-mediated interventions. In Chapter VII, we explored whether the robot-mediated intervention provided by Norris could help improve the performance of children with ASD in human-human interaction. In other words, we investigated whether the joint attention skills learnt by children with ASD within the Norris system were generalized to social communication with humans.

We conducted a more rigorous pilot randomized control study with 11 children with ASD. Children were randomized into an immediate participation group or a waitlist control group and experienced strictly scheduled robot-mediated interventions. Their performance in both human-robot interaction and human-human interaction was assessed. This group of participants had a lower baseline within-system performance compared to the participants in Chapter VI. The results showed that participants who received the robot-mediated intervention improved more in human-human interaction, compared with the participants who did not receive the intervention. In addition, the participants who improved in human-robot interaction gained more improvement in human-human interaction, compared with the participants who did not improve in human-robot interaction. Even though we did not observe significant change regarding the HRI performance for this particular participant group, these data are promising enough for us to continue with the study to observe what happens as sample size increases. Ultimately, by conducting longitudinal user studies with fine-tuned machine-assisted intervention systems and reliable psychological evaluations, it will be clearer how HRI impacts children with ASD in their daily interactions.

1.4 References

[1] M. L. Ganz, “The lifetime distribution of the incremental societal costs of autism,” *Archives of Pediatrics and Adolescent Medicine, Am Med Assoc,* vol. 161, no. 4, pp. 343-349, 2007.

[2] *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the Diagnostic Criteria from DSM-IV-TR*, Fourth ed., Washington D.C.: American Psychiatric Association, 2000.

[3] a. P. I. Developmental Disabilities Monitoring Network Surveillance Year, “Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010,” *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002),* vol. 63, pp. 1, 2014.

[4] Z. Warren, M. L. McPheeters, N. Sathe, J. H. Foss-Feig, A. Glasser, and J. Veenstra-VanderWeele, “A systematic review of early intensive intervention for autism spectrum disorders,” *Pediatrics,* vol. 127, no. 5, pp. e1303-e1311, 2011.

[5] Z. Warren, A. Vehorn, E. Dohrmann, C. Newsom, and J. L. Taylor, “Brief report: Service implementation and maternal distress surrounding evaluation recommendations for young children diagnosed with autism,” *Autism*, 2012.

[6] M. Al-Qabandi, J. W. Gorter, and P. Rosenbaum, “Early autism detection: Are we ready for routine screening?,” *Pediatrics,* vol. 128, no. 1, pp. e211-e217, 2011.

[7] *Interagency Autism Coordinating Committee Strategic Plan for Autism Spectrum Disorder Research,* NIH Publication No. 09-7465, USDHHS Document, http://iacc.hhs.gov/strategic-plan/2009/, U.S. Department of Health and Human Services, 2009.

[8] E. A. Lee, and S. A. Seshia, *Introduction to embedded systems: A cyber-physical systems approach*: Lee & Seshia, 2011.

[9] J. H. Williams, A. Whiten, T. Suddendorf, and D. I. Perrett, “Imitation, mirror neurons and autism,” *Neuroscience & Biobehavioral Reviews,* vol. 25, no. 4, pp. 287-295, 2001.

[10] A. Whiten, and J. Brown, “Imitation and the reading of other minds: Perspectives from the study of autism, normal children and non-human primates,” *Intersubjective communication and emotion in early ontogeny*, pp. 260-280, 1998.

[11] B. Ingersoll, “Brief report: Effect of a focused imitation intervention on social functioning in children with autism,” *Journal of autism and developmental disorders,* vol. 42, no. 8, pp. 1768-1773, 2012.

[12] B. Ingersoll, “Pilot Randomized Controlled Trial of Reciprocal Imitation Training for Teaching Elicited and Spontaneous Imitation to Children with Autism. ,” *Journal of Autism and Developmental Disorders,* vol. 40, no. 1154-1160, 2010.

[13] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, “Early social attention impairments in autism: social orienting, joint attention, and attention to distress,” *Developmental psychology,* vol. 40, no. 2, pp. 271, 2004.

[14] G. Dawson, A. N. Meltzoff, J. Osterling, and J. Rinaldi, “Neuropsychological correlates of early symptoms of autism,” *Child development,* vol. 69, no. 5, pp. 1276-1285, 1998.

[15] A. S. Nadig, S. Ozonoff, G. S. Young, A. Rozga, M. Sigman, and S. J. Rogers, “A prospective study of response to name in infants at risk for autism,” *Archives of pediatrics & adolescent medicine,* vol. 161, no. 4, pp. 378-383, 2007.

[16] M. W. Mosconi, J. S. Reznick, G. Mesibov, and J. Piven, “The Social Orienting Continuum and Response Scale (SOC-RS): A dimensional measure for preschool-aged children,” *Journal of autism and developmental disorders,* vol. 39, no. 2, pp. 242-250, 2009.

[17] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[18] G. Dawson, “Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder,” *Development and Psychopathology, Cambridge Univ Press,* vol. 20, no. 3, pp. 775-803, 2008.

[19] K. K. Poon, Watson, L.R., Baranek, G.T., Poe, M.D., “To What Extent Do Joint Attention, Imitation, and Object Play Behaviors in Infancy Predict Later Communication and Intellectual Functioning in ASD?,” *Journal of Autism and Developmental Disorders, DOI: 10.1007/s10803-011-1349-z*, pp. 1-11, 2011.

[20] C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., “Language outcome in autism: Randomized comparison of joint attention and play interventions.,” *Journal of consulting and clinical psychology,* vol. 76, no. 1, pp. 125–137, 2008.

[21] C. Kasari, Gulsrud, A.C., Wong, C., Kwon, S., Locke, J., “Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism,” *Journal of autism and developmental disorders,* vol. 40, no. 9, pp. 1045-1056, 2010.

[22] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, “The clinical use of robots for individuals with autism spectrum disorders: A critical review,” *Research in autism spectrum disorders,* vol. 6, no. 1, pp. 249-262, 2012.

[23] B. Scassellati, H. Admoni, and M. Mataric, “Robots for use in autism research,” *Annual Review of Biomedical Engineering,* vol. 14, pp. 275-294, 2012.

[24] U. Lahiri, Bekele, E., Warren, Z., Sarkar, N., “Design of a Virtual Reality based Adaptive Response Technology for Children with Autism,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2012.

[25] E. Bekele, Z. Zheng, U. Lahiri, A. Swanson, J. Davidson, Z. Warren, and N. Sarkar, “Design of a novel virtual reality-based autism intervention system for facial emotional expressions identification,” in International Conference on Disability, Virtual Reality and Associated Technologies, 2012, pp. in press.

[26] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, “Using the Humanoid Robot KASPAR to Autonomously PlayTriadic Games and Facilitate Collaborative Play Among Children With Autism.”

[27] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder,” *International Journal of Human-Computer Studies, Elsevier,* vol. 66, no. 9, pp. 662-677, 2008.

[28] Z. Zheng, Q. Fu, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Design of a Computer-assisted System for Teaching Attentional Skills to Toddlers with Autism."

[29] J. Greczek, E. Kaszubksi, A. Atrash, and M. J. Matarić, “Graded Cueing Feedback in Robot-Mediated Imitation Practice for Children with Autism Spectrum Disorders,” *Proceedings, 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014) Edinburgh, Scotland, UK* Aug. 2014.

[30] D. Feil-Seifer, Mataric, M., “Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features,” in In Proceedings of the 6th international conference (ACM/IEEE) on Human-robot interaction, New York, NY: ACM Press, 2011, pp. 323-330.

[31] F. Amirabdollahian, B. Robins, K. Dautenhahn, and Z. Ji, "Investigating tactile event recognition in child-robot interaction for use in autism therapy." pp. 5347-5351.

[32] D. Francois, K. Dautenhahn, and D. Polani, “Using real-time recognition of human-robot interaction styles for creating adaptive robot behaviour in robot-assisted play. ,” in IEEE Symposium on Artificial Life, 2009, pp. 45-52.

[33] M. A. Goodrich, and A. C. Schultz, “Human-robot interaction: a survey,” *Foundations and trends in human-computer interaction,* vol. 1, no. 3, pp. 203-275, 2007.

[34] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-computer interaction*: Addison-Wesley Longman Ltd., 1994.

[35] R. Ballagas, J. Borchers, M. Rohs, and J. G. Sheridan, “The smart phone: a ubiquitous input device,” *Pervasive Computing, IEEE,* vol. 5, no. 1, pp. 70-77, 2006.

[36] S. Weir, and R. Emanuel, *Using LOGO to catalyse communication in an autistic child*: Department of Artificial Intelligence, University of Edinburgh, 1976.

[37] D. J. Ricks, and M. B. Colton, "Trends and considerations in robot-assisted autism therapy." pp. 4354-4359.

[38] B. Robins, K. Dautenhahn, R. Boekhorst, and A. Billard, “Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?,” *Universal Access in the Information Society,* vol. 4, no. 2, pp. 105-120, 2005.

[39] J. Diehl, Schmitt, L., Villano, M., and Crowell, C., “The clinical use of robots for individuals with Autism Spectrum Disorders: A critical review,” *Research in Autism Spectrum Disorders,* vol. 6, no. 1, pp. 249-262, 2011.

[40] K. Dautenhahn, and I. Werry, “Towards interactive robots in autism therapy: Background, motivation and challenges,” *Pragmatics & Cognition,* vol. 12, no. 1, pp. 1-35, 2004.

[41] B. Robins, K. Dautenhahn, and J. Dubowski, “Does appearance matter in the interaction of children with autism with a humanoid robot?,” *Interaction Studies,* vol. 7, no. 3, pp. 509-542, 2006.

[42] A. C. Pierno, M. Mari, D. Lusher, and U. Castiello, “Robotic movement elicits visuomotor priming in children with autism,” *Neuropsychologia,* vol. 46, no. 2, pp. 448-454, 2008.

[43] G. Bird, J. Leighton, C. Press, and C. Heyes, “Intact automatic imitation of human and robot actions in autism spectrum disorders,” *Proceedings of the Royal Society B: Biological Sciences,* vol. 274, no. 1628, pp. 3027-3031, 2007.

[44] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay, and W. Jones, “Two-year-olds with autism orient to nonsocial contingencies rather than biological motion,” *Nature,* vol. 459, no. 7244, pp. 257–261, 2009.

[45] T. Falck-Ytter, S. Bölte, and G. Gredebäck, “Eye tracking in early autism research,” *Journal of neurodevelopmental disorders,* vol. 5, no. 1, pp. 28, 2013.

[46] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, “The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents,” *International Journal of Social Robotics,* vol. 3, pp. 41-52, 2011.

[47] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, “The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains,” in The Annual Meeting of the Cognitive Science Society, COGSCI 2012, 2012.

[48] A. Ioannou, E. Andreou, and M. Christofi, “Pre-schoolers’ Interest and Caring Behaviour Around a Humanoid Robot,” *TechTrends,* vol. 59, no. 2, pp. 23-26, 2015.

[49] S. Costa, F. Soares, C. Santos, M. J. Ferreira, F. Moreira, A. P. Pereira, and F. Cunha, "An approach to promote social and communication behaviors in children with Autism Spectrum Disorders: Robot based intervention." pp. 101-106.

[50] S. M. Anzalone, E. Tilmont, S. Boucenna, J. Xavier, A.-L. Jouen, N. Bodeau, K. Maharatna, M. Chetouani, D. Cohen, and M. S. Group, “How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot,” *Research in Autism Spectrum Disorders,* vol. 8, no. 7, pp. 814-826, 2014.

[51] C. A. Costescu, B. Vanderborght, and D. O. David, “Reversal Learning Task in Children with Autism Spectrum Disorder: A Robot-Based Approach,” *Journal of autism and developmental disorders*, pp. 1-11, 2014.

[52] B. Scassellati, "Quantitative metrics of social response for autism diagnosis." pp. 585-590.

[53] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, "Autonomous robot-mediated imitation learning for children with autism." pp. 2707-2712.

[54] Z. Zheng, L. Zhang, E. Bekele, A. Swanson, J. Crittendon, Z. Warren, and N. Sarkar, "Impact of Robot-mediated Interaction System on Joint Attention Skills for Children with Autism ".

[55] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Online affect detection and robot behavior adaptation for intervention of children with autism,” *Robotics, IEEE Transactions on,* vol. 24, no. 4, pp. 883-896, 2008.

[56] C. M. Stanton, P. H. Kahn, R. L. Severson, J. H. Ruckert, and B. T. Gill, "Robotic animals might aid in the social development of children with autism." pp. 271-278.

[57] L. Dickstein-Fischer, E. Alexander, X. Yan, H. Su, K. Harrington, and G. S. Fischer, “An affordable compact humanoid robot for autism spectrum disorder interventions in children,” in Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, 2011, pp. 5319 - 5322.

[58] Kim ES, Berkovits LD, Bernier EP, Leyzberg D, Shic F, Paul R, and B. Scassellati, “Social Robots as Embedded Reinforcers of Social Behavior in Children with Autism.,” *J Autism Dev Disord.* , 2012.

[59] G. Pioggia, M. Sica, M. Ferro, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, "Human-robot interaction in autism: FACE, an android-based social therapy." pp. 605-612.

[60] M. A. Goodrich, M. A. Colton, B. Brinton, and M. Fujiki, “A case for low-dose robotics in autism therapy,” in Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on, 2011, pp. 143 - 144.

[61] H. Kozima, and H. Yano, "A robot that learns to communicate with human caregivers." pp. 47-52.

[62] I. Fujimoto, T. Matsumoto, P. R. S. De Silva, M. Kobayashi, and M. Higashi, “Mimicking and evaluating human motion to improve the imitation skill of children with autism through a robot,” *International Journal of Social Robotics,* vol. 3, no. 4, pp. 349-357, 2011.

[63] J. Wainer, K. Dautenhahn, B. Robins, and F. 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,* vol. 6, no. 1, pp. 45-65, 2014.

[64] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, “KASPAR–a minimally expressive humanoid robot for human–robot interaction research,” *Applied Bionics and Biomechanics,* vol. 6, no. 3-4, pp. 369-397, 2009.

[65] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar, “A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism,” *Neural Systems and Rehabilitation Engineering, IEEE Transactions on,* vol. 21, no. 2, pp. 289-299, 2013.

[66] N. A. Torres, N. Clark, I. Ranatunga, and D. Popa, "Implementation of interactive arm playback behaviors of social robot Zeno for autism spectrum disorder therapy." p. 21.

[67] I. Ranatunga, M. Beltran, N. A. Torres, N. Bugnariu, R. M. Patterson, C. Garver, and D. O. Popa, "Human-robot upper body gesture imitation analysis for autism spectrum disorders," *Social Robotics*, pp. 218-228: Springer, 2013.

[68] H. Kozima, C. Nakagawa, and Y. Yasuda, "Interactive robots for communication-care: A case-study in autism therapy." pp. 341-346.

[69] L. Boccanfuso, and J. M. O’Kane, “CHARLIE: An adaptive robot design with hand and face tracking for use in autism therapy,” *International Journal of Social Robotics,* vol. 3, no. 4, pp. 337-347, 2011.

[70] M. C. Chuah, D. Coombe, C. Garman, C. Guerrero, and J. Spletzer, "Lehigh Instrument for Learning Interaction (LILI): An Interactive Robot to Aid Development of Social Skills for Autistic Children." pp. 731-736.

[71] P. Ravindra, S. De Silva, K. Tadano, A. Saito, S. G. Lambacher, and M. Higashi, "Therapeutic-assisted robot for children with autism." pp. 3561-3567.

[72] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, “Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?,” *Universal Access in the Information Society,* vol. 4, no. 2, pp. 105-120, 2005.

[73] S. Costa, H. Lehmann, K. Dautenhahn, B. Robins, and F. Soares, “Using a Humanoid Robot to Elicit Body Awareness and Appropriate Physical Interaction in Children with Autism,” *International Journal of Social Robotics*, pp. 1-14, 2014.

[74] A. Duquette, Michaud, F., Mercier, H., “Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism,” *Autonomous Robots,* vol. 24, no. 2, pp. 147-157, 2008.

[75] M. A. Goodrich, M. Colton, B. Brinton, and M. Fujiki, “A case for low-dose robotics in autism therapy,” in Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on, 2011, pp. 143 - 144.

[76] D. Feil-Seifer, Matarić, M., “Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders,” in Experimental Robotics, 54, 2009, pp. 201-210.

[77] P. J. Yoder, McDuffie, A.S., "Treatment of responding to and initiating joint attention," *Social & communication in autism spectrum disorders: Early identification, diagnosis, & intervention* T. Charman, W. Stone ed., pp. 117-142, New York: Guilford, 2006.

[78] F. Amirabdollahian, B. Robins, K. Dautenhahn, and Z. Ji, “Investigating tactile event recognition in child-robot interaction for use in autism therapy,” in Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, 2011, pp. 5347 - 5351.

[79] C. Liu, Conn, K., Sarkar, N., Stone, W., “Online affect detection and robot behavior adaptation for intervention of children with autism,” *Robotics, IEEE Transactions on,* vol. 24, no. 4, pp. 883 - 896, 2008.

[80] E. Bekele, J. A. Crittendon, A. Swanson, N. Sarkar, and Z. E. Warren, “Pilot clinical application of an adaptive robotic system for young children with autism,” *Autism,* vol. 18, no. no. 5, pp. 598-608, 2014.

[81] Z. E. Warren, Z. Zheng, A. R. Swanson, E. Bekele, L. Zhang, J. A. Crittendon, A. F. Weitlauf, and N. Sarkar, “Can Robotic Interaction Improve Joint Attention Skills?,” *Journal of autism and developmental disorders*, pp. 1-9, 2013.

[82] P. J. Yoder, A. S. McDuffie, T. Charman, and W. Stone., "Treatment of responding to and initiating joint attention," *Social & communication in autism spectrum disorders: Early identification, diagnosis, & intervention* T. Charman, W. Stone ed., pp. 117-142, New York: Guilford, 2006.

[83] Z. Zheng, E. Bekele, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "The Impact of Robots on Children with Autism Spectrum Disorder (ASD)," *Autism Imaging and Devices*, M. F. Casanova, A. S. ElBaz and J. Suri, eds.: Taylor & Francis, 2016.

[84] R. A. Cooper, B. E. Dicianno, B. Brewer, E. LoPresti, D. Ding, R. Simpson, G. Grindle, and H. Wang, “A perspective on intelligent devices and environments in medical rehabilitation,” *Medical Engineering & Physics,* vol. 30, no. 10, pp. 1387-1398, 2008.

[85] S.-F. Tam, D. W.-K. Man, Y.-P. Chan, P.-C. Sze, and C.-M. Wong, “Evaluation of a Computer-Assisted, 2-D Virtual Reality System for Training People With Intellectual Disabilities on How to Shop,” *Rehabilitation Psychology,* vol. 50, no. 3, pp. 285, 2005.

[86] A. Cerasa, M. C. Gioia, P. Valentino, R. Nisticò, C. Chiriaco, D. Pirritano, F. Tomaiuolo, G. Mangone, M. Trotta, and T. Talarico, “Computer-Assisted Cognitive Rehabilitation of Attention Deficits for Multiple Sclerosis A Randomized Trial With fMRI Correlates,” *Neurorehabilitation and neural repair,* vol. 27, no. 4, pp. 284-295, 2013.

[87] S. Bonavita, R. Sacco, M. Della Corte, S. Esposito, M. Sparaco, A. d’Ambrosio, R. Docimo, A. Bisecco, L. Lavorgna, and D. Corbo, “Computer-aided cognitive rehabilitation improves cognitive performances and induces brain functional connectivity changes in relapsing remitting multiple sclerosis patients: an exploratory study,” *Journal of neurology,* vol. 262, no. 1, pp. 91-100, 2015.

[88] C. Williams, B. Wright, G. Callaghan, and B. Coughlan, “Do children with autism learn to read more readily by computer assisted instruction or traditional book methods? A pilot study,” *Autism,* vol. 6, no. 1, pp. 71-91, 2002.

[89] C. Hedbring, and C. Newsom, “Visual overselectivity: A comparison of two instructional remediation procedures with autistic children,” *Journal of autism and developmental disorders,* vol. 15, no. 1, pp. 9-22, 1985.

[90] U. Frith, “Autism: Explaining the enigma,” 1989.

[91] T. Reed, “Performance of autistic and control subjects on three cognitive perspective-taking tasks,” *Journal of autism and developmental disorders,* vol. 24, no. 1, pp. 53-66, 1994.

[92] K. M. Colby, “The rationale for computer-based treatment of language difficulties in nonspeaking autistic children,” *Journal of Autism and Childhood Schizophrenia,* vol. 3, no. 3, pp. 254-260, 1973.

[93] M. Heimann, K. E. Nelson, T. Tjus, and C. Gillberg, “Increasing reading and communication skills in children with autism through an interactive multimedia computer program,” *Journal of autism and developmental disorders,* vol. 25, no. 5, pp. 459-480, 1995.

[94] T. Hagiwara, and B. S. Myles, “A multimedia social story intervention teaching skills to children with autism,” *Focus on Autism and other developmental disabilities,* vol. 14, no. 2, pp. 82-95, 1999.

[95] O. E. Hetzroni, and J. Tannous, “Effects of a computer-based intervention program on the communicative functions of children with autism,” *Journal of autism and developmental disorders,* vol. 34, no. 2, pp. 95-113, 2004.

[96] L. M. Passerino, and L. M. C. Santarosa, “Autism and digital learning environments: Processes of interaction and mediation,” *Computers & Education,* vol. 51, no. 1, pp. 385-402, 2008.

[97] M. Silver, and P. Oakes, "Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others," *Autism,* vol. 5, no. 3, pp. 299-316, 2001.

[98] S. Cobb, S. Kerr, and T. Glover, "The AS Interactive Project: Developing virtual environments for social skills training in users with Asperger syndrome," *Robotic and Virtual Interactive Systems in Autism Therapy, Communications of the Adaptive Systems Research Group, University of Hertfordshire (Report no 364)*, 2001.

[99] S. Wallace, S. Parsons, A. Westbury, K. White, K. White, and A. Bailey, "Sense of presence and atypical social judgments in immersive virtual environments Responses of adolescents with Autism Spectrum Disorders," *Autism,* vol. 14, no. 3, pp. 199-213, 2010.

[100] B. A. Mineo, W. Ziegler, S. Gill, and D. Salkin, "Engagement with electronic screen media among students with autism spectrum disorders," *Journal of autism and developmental disorders,* vol. 39, no. 1, pp. 172-187, 2009.

[101] A. Tartaro, and J. Cassell, "Authorable virtual peers for autism spectrum disorders."

[102] S. Bernardini, K. Porayska-Pomsta, and T. J. Smith, "ECHOES: An intelligent serious game for fostering social communication in children with autism," *Information Sciences,* vol. 264, pp. 41-60, 2014.

[103] D. Moore, Y. Cheng, P. McGrath, and N. J. Powell, "Collaborative virtual environment technology for people with autism," *Focus on Autism and Other Developmental Disabilities,* vol. 20, no. 4, pp. 231-243, 2005.

[104] J. Wade, D. Bian, J. Fan, L. Zhang, A. Swanson, M. Sarkar, A. Weitlauf, Z. Warren, and N. Sarkar, "A Virtual Reality Driving Environment for Training Safe Gaze Patterns: Application in Individuals with ASD," *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being*, pp. 689-697: Springer, 2015.

[105] O. Grynszpan, P. L. T. Weiss, F. Perez-Diaz, and E. Gal, "Innovative technology-based interventions for autism spectrum disorders: a meta-analysis," *Autism,* vol. 18, no. 4, pp. 346-361, 2014.

[106] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley, "Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model," *Pediatrics,* vol. 125, no. 1, pp. e17-e23, 2010.

[107] Z. E. Warren, and W. L. Stone, "Best practices: Early diagnosis and psychological assessment," *Autism Spectrum Disorders*, David Amaral, Daniel Geschwind and G. Dawson, eds., pp. 1271-1282, New York: Oxford University Press, 2011.

[108] G. Bird, Leighton, J., Press, C., Heyes, C., "Intact automatic imitation of human and robot actions in autism spectrum disorders," *Proceedings of the Royal Society B: Biological Sciences,* vol. 274, no. 1628, pp. 3027–3031, 2007.

[109] Z. Zheng, E. Young, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Robot-mediated Imitation Skill Training for Children with Autism," *IEEE transactions on neural systems and rehabilitation engineering*, 2015.

[110] "Microsoft Kinect for Windows," http://www.microsoft.com/en-us/kinectforwindows/.

[111] Z. Zheng, E. M. Young, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Robot-mediated mixed gesture imitation skill training for young children with ASD." pp. 72-77.

[112] Z. Zheng, Q. Fu, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Design of an Autonomous Social Orienting Training System (ASOTS) for Young Children with Autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2016.

[113] Z. Zheng, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Non-invasive Autonomous Robot-mediated Joint Attention Intervention for Young Children with ASD," *IEEE Transactions on Human-Machine Systems*, (under review).

Chapter II. Robot-mediated Single Gesture Imitation Intervention

2.1 Abstract

In this chapter we present a novel robot-mediated intervention system for imitation skill learning, which is considered a core deficit area for children with ASD. The Robot-mediated Imitation Skill Intervention Architecture 1 (RISIA1) is designed in such a manner that it can operate either completely autonomously or in coordination with a human therapist depending on the intervention need. A finite state machine based gesture recognition algorithm was proposed to recognize both partially and fully completed gestures. The recognition results were used by the robot to provide individualized feedback to the participants regarding their imitation performance. Preliminary results show that this novel robotic system draws more attention from the children with ASD and teaches gestures more effectively as compared to a human therapist. While no broad generalized conclusions can be made about the effectiveness of RISIA1 based on our small user studies, initial results are encouraging and justify further exploration in the future.

Keywords—Autism Spectrum Disorder, Imitation Skill Intervention, Robot and Autism, Robot-mediated Intervention.

2.2 Introduction

In the current work, we describe the development and initial application of a non-invasive intelligent robotic intervention system capable of dynamic and individualized interaction with potential relevance to improving imitation skills for young children with ASD. Imitation involves translating from the perspective of another individual to oneself, and creating representation of this individual's primary representation of the world [1]. Although the exact reasons of the imitation impairment associated with ASD is still unclear, evidence suggests that this imitation impairment may be related with the basic ability to map actions of others onto an imitative match by oneself [2]. Imitation is a critically important social communication skill that emerges early in life and it is theorized to play an important role in the development of cognitive, language, and social skills [3]. Children with ASD show powerful impairments in imitation and such deficits have been tied to a host of associated neurodevelopmental and learning challenges over time[4].

There are only a few preliminary robotic studies reported on imitation learning for children with ASD. Duquette et al. [5] compared the impact of a mobile robot Tito with a human therapist on the imitation behavior of children with ASD. The participants paired with the robot demonstrated more shared attention and imitated more facial expression, while the participants paired with the human imitated more body movement. Bugnariu et al. [6] developed a method to quantify imitation using a robot, kinematic data and a Dynamic Time Warping algorithm. Cabibihan et al. [7] claimed that imitation skills taught with the aid of humanoid robots had potential to be generalized with people. Robins et al. [8] conducted a longitudinal study where children freely interacted with a humanoid robot Robota where the participants exhibited diverse imitation behaviors after repeated exposure. Srinivasan et al. [9] conducted a interaction study with a small humanoid robot Isobot. Results showed that the task-specific imitation and generalized praxis performance improved for a group of typically developing children and one child with ASD.

The imitation intervention literature suggests that intervention approaches are most effective when children show sustained engagement with a variety of objects, can be utilized within intrinsically motivating settings, and when careful adaptation to small gains and shifts can be incorporated and utilized over longer intervals of time [4].

The contribution of this work is two-fold: first, we present a novel autonomous Robot-mediated Imitation Skill Intervention Architecture or RISIA1 specifically designed for children with ASD including a new gesture recognition algorithm that can assess imitated gestures in real-time and provide dynamic feedback. Second, we present a preliminary user study to demonstrate the tolerability and usefulness of robotic interaction using RISIA1 with a group of both typically developing children and children with ASD. Our initial concepts and results were presented in [10].

In what follows, we first introduce the robot-mediated imitation skill intervention system architecture in Section 2.3. Section 2.4 describes the gesture representation method used in this study. Section 2.5 discusses the gesture recognition algorithm embedded in the system. The experiments and results are described in Section 2.6. Finally, Section 2.7 discusses the potential and limitations of the current study.

2.3 Robot-Mediated Imitation Skill Intervention Architecture 1 (RISIA1)

The RISIA1 for children with ASD that we present in this chapter has a humanoid robot as a task administrator, a camera for gesture recognition, a gesture recognition algorithm to assess the imitated gesture, and a feedback mechanism to encourage interaction (Fig. II.1). Existing literature [11, 12] suggests, at least in preliminary form, the ability of humanoid robotic interaction systems to capture interest of some children with ASD in a manner that could potentially be leveraged into meaningful intervention approaches. RISIA1 is designed to teach imitation skills via the robot by first making the robot demonstrate a target gesture and asking the child to imitate it, assessing the imitated gesture, and finally providing relevant feedback – all autonomously and in a closed-loop manner. An interesting feature of this system is that the robot can be replaced by a human therapist within the architecture when needed without altering the rest of the system components, such as the gesture recognition and feedback modules, allowing the system to be used for co-robotic intervention.

The RISIA1 architecture is illustrated in Fig. II.1. There are four important modules in RISIA1. The robotic gesture demonstration (RGD) module is meant for demonstrating a gesture by the robot and is implemented on the humanoid robot NAO [13]. NAO has 25 degrees of freedom, 2 flashing LED "eyes", 2 speakers, and a synthetic childlike voice. The imitated gesture sensing (IGS) module is used to sense the imitated gesture by the child and is implemented using Microsoft Kinect [14], which can track a person's skeleton with an average of accuracy of 5.6 mm in 3D space [15, 16]. The supervisory controller (SC) is the primary control mechanism for RISIA1 and is designed based on a timed automata model [17]. We use a timed automata model for SC because it fits well with the Finite State Machine (FSM) based gesture recognition method that we use for gesture recognition and the state-based predesigned robot behavior libraries that we utilize for robot gestures. The SC manages the component communications, handles the experimental logic and is embedded with a gesture recognition algorithm that can recognize a partial or completed gesture. It instructs the robot to show a target gesture to the child and once the target gesture is completed, the child is asked by the robot to imitate the gesture. The SC continuously monitors the IGS and evaluates the child's imitative performance for feedback. Based on this performance, the SC may instruct the robot either to give rewards or aid the child with reinforcement components and approximations of the gestures within their motor movements. The feedback provided by the robot were predefined by autism clinicians for every state in the FSMs and stored in the system software library. The SC continues this procedure in a closed-loop manner for a specified duration of time and collects data to evaluate the efficacy of the trials.

The Graphical User Interface (GUI) is designed in a manner that it is easily operated by an experimenter, e.g., a therapist, who may not be technologically savvy. The head pose estimation, skeleton tracking and the participant's real time video are displayed for observation.

In the trial, the NAO provides imitation prompts in the form of recorded verbal scripts, mirroring movements, and gestural movements for imitation. Given the gesture prompts, the participant's response is sensed by a Microsoft Kinect at 30 frames per second (fps). Skeleton data from Kinect are processed using a Holt double exponential smoothing filter to avoid glitch and jitter. The Kinect SDK face tracking functions fit a 3D convex mesh on the participant's head and provide 3D position and orientation of the participant's head within the Kinect frame. The head pose is then used for estimating a participant's attention on the robot or on the human therapist. If the participant moved out of the Kinect tracking zone during interaction, a signal would be triggered to suspend the robotic action and RISIA1 would not proceed with the interaction until the child returned to the appropriate region and the Kinect resumed its tracking.

In this work we chose a set of arm gestural movements for imitation skill learning, which were: 1) raising one hand, 2) waving, 3) raising two hands, and 4) reaching arms out to the side. These four gestures were intentionally selected due to the low motor skill requirements they presented to participating children. These gestures were also selected to avoid motor limitations of the humanoid robot (e.g., challenges crossing midline and adequately positioning fingers/digits). However, the capabilities of RISIA1 are not limited to these gestures alone. Rather, they represent the kinds of gestures that a therapist might use for imitation skill learning intervention.

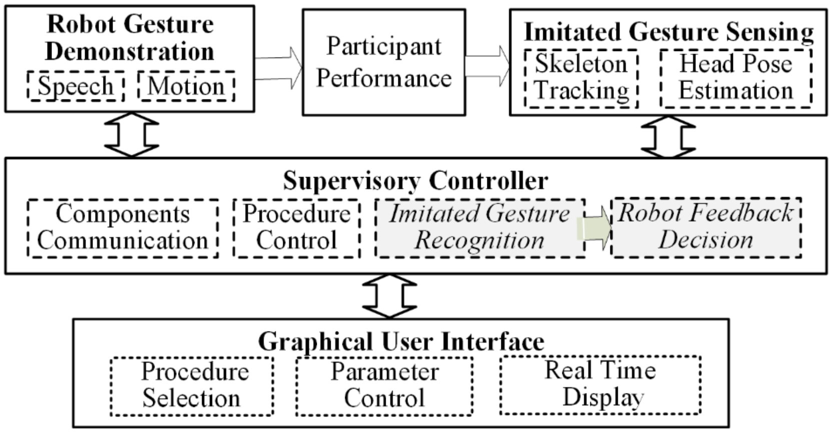


Fig. II.1 RISTA1 system architecture

2.4 Gesture Representation

2.4.1 Gesture representation for the human

We first defined a set of variables that could mathematically capture each of the four chosen gestures as described above. The gesture variables are shown in Table II.1. Fig. II.2 shows some of these gesture variables with respect to the Kinect frame. Each gesture was broken down into several salient parts where each part was represented as a state (Fig. II.3). Note that this decomposition is not unique and was designed based on common sense and with several autism clinicians' input.

2.4.2 Gesture representation for the robot

In order to implement the same gestures on the robot so that the robot can demonstrate these gestures, each gesture was carefully designed by specifying its joint angle trajectories. Finally, each gesture was stored in a library that the supervisory controller could select from and play. A part of our imitation skill intervention paradigm included gesture mirroring by the robot. In other words, sometimes the robot was also required to copy a participant's gesture. The skeleton tracking module of Kinect was used to acquire a participant's arm joint angles, which were then mapped to the corresponding joint angles of the robot. If the participant's joint angles were outside the robot's workspace, then the robot angles were set for their maximum attainable values.

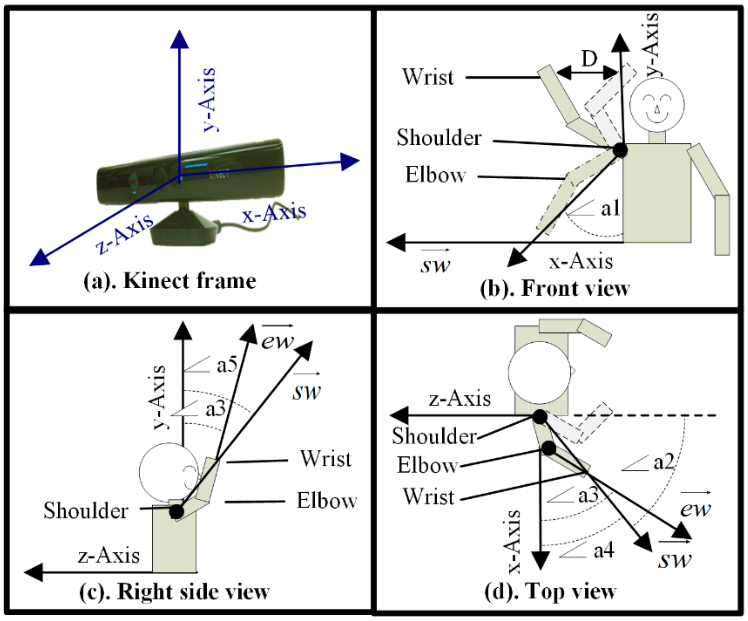


Fig. II.2 Gesture variables demonstration in Kinect frame

2.5 Gesture Recognition by the Robot

An interactive robot-mediated imitation skill intervention system must be able to dynamically provide feedback to the participants similar to the way a therapist does during intervention. The robot's feedback depended on the accuracy and speed of the gesture recognition algorithm. In addition, given the target population it was quite likely that the participants would not be able to completely imitate all the gestures. In order to scaffold participant skills, the robot needed to recognize partially completed gestures, detect what components of a gesture require attention, and provide specific feedback to improve the detected deficiency.

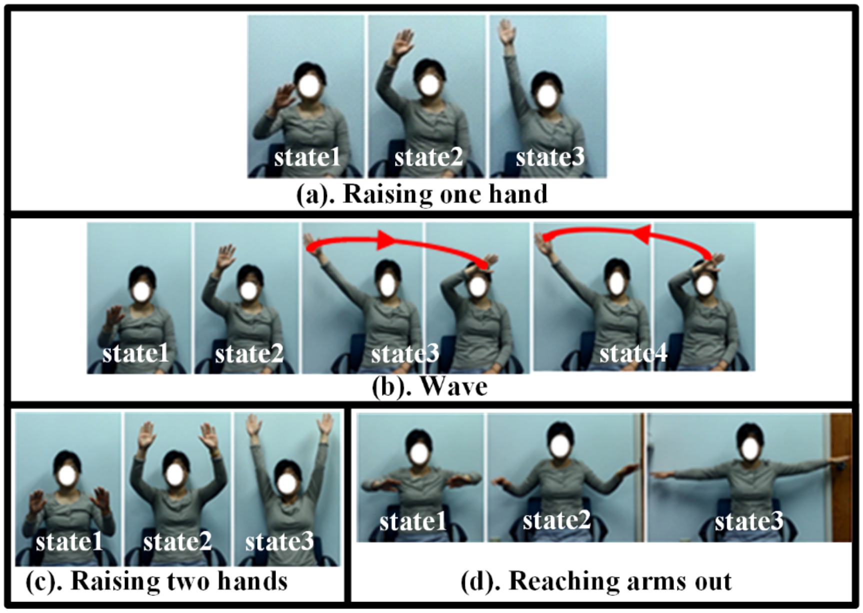
In order to achieve these goals, we designed a rule-based finite state machine (FSM) method to recognize gestures. While there are several powerful probabilistic methods for gesture recognition such as

Hidden Markov Model [18] and particle filtering [19], we chose a rule-based method to avoid the complexity of computation and the difficulty of generating a training data set due to the young age of our participants. It is difficult for a young child to repeat a standard gesture multiple times accurately to create training data. The recognition accuracy was of utmost importance in this task since the robot should not provide erroneous feedback to the children, which might confuse or frustrate them.

Table II.1 Gesture Variables of Gesture Representation and Recognition

Symbol	***Definition***
$\overrightarrow{sw}$	Vector pointing from shoulder to wrist
$\overrightarrow{ew}$	Vector pointing from elbow to wrist
w_y	y coordinate of the wrist joint
e_y	y coordinate of the elbow joint
s_y	y coordinate of the shoulder joint
$\angle a1$	Angle between $\overrightarrow{sw}$ and negative y axis
$\angle a2$	Angle between $\overrightarrow{sw}$ and yz plane, when $\overrightarrow{sw}$ in negative z direction (arm pointing forward)
$\angle a3$	Angle between $\overrightarrow{ew}$ and xy plane
$\angle a4$	Angle between $\overrightarrow{sw}$ and positive x axis for right arm, angle between $\overrightarrow{sw}$ and negative x axis for left arm.
$\angle a5$	Angle between $\overrightarrow{sw}$ and xy plane, $\overrightarrow{sw}$ with positive x direction for right arm, and with negative x direction for left arm.
$\angle wes$	Angle between the upper arm and the forearm
D	x direction movement
H	y direction raised height
T_{item}	Threshold for distances or angles

FSM has been widely used to model and recognize gestures [20]. We chose a FSM method because we can break down each gesture into a number of intermediate states, such that the recognition algorithm can precisely detect a partial gesture and thus allow the robot to provide more targeted feedback. We designed a FSM representation for each gesture and defined a region of interest (ROI) in which each FSM would be activated. These ROIs are defined in Eqn. (II-2)-(II-4). For example, a wave gesture FSM was only activated when the participant's arm was raised in front of the torso, the wrist was higher than the shoulder and the forearm was pointing upwards. The input variables to the gesture recognition FSM are computed from skeleton coordinates. Five sliding windows (1-5 seconds) were used to chop the FSM input data. Those windows were updated in every frame. In this way, we set up the maximum completion time for a gesture to be 5 seconds. Although usually a gesture lasts for 2-3 seconds, we introduce additional time for flexibility.



(a). Raising one hand

(b). Wave

(c). Raising two hands

(d). Reaching arms out

Fig. II.3 Gesture states of the four gestures in this study

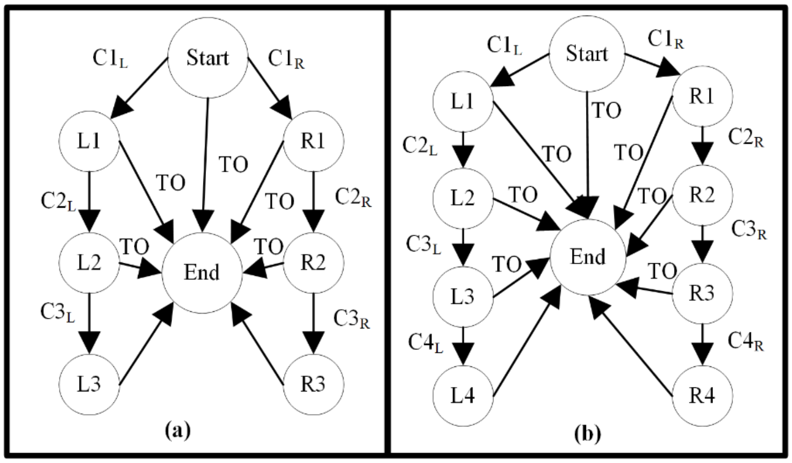


Fig. II.4 FSM model for gesture recognition

$$ROI_{Wave} = \{\angle a1 < T_{ang1}, \angle a2 < T_{ang2}, w_y > s_y, \angle a3 < T_{ang3}\}\ . \qquad \text{(II-1)}$$

$$ROI_{RaiseHand(s)} = \{\angle a1 < T_{ang3}, \angle a2 < T_{ang2}\}\ . \qquad \text{(II-2)}$$

$$ROI_{ReachArmsOut} = \{\angle a1 < T_{ang3}, \angle a4 < T_{ang3}, \angle a5 < T_{ang2}, \overrightarrow{sw}_y < 0, (\overrightarrow{sw}_x > 0 \text{ (right arm) or } \overrightarrow{sw}_x < 0 \text{ (left arm)})\}\ . \qquad \text{(II-3)}$$

We briefly discuss how the FSM works for each gesture.

2.5.1 Raising one hand

This gesture includes a) start raising a hand until the wrist is higher than the elbow; b) continue raising the hand and stretching the arm until the elbow is higher than the shoulder; and c) stretch the arm slightly further until it becomes straight (Fig. II.3(a)). Fig. II.4(a) shows its FSM representation.

L1 to L3 in Fig. II.4(a) are the corresponding states that describe the 3 stages for the left arm, and R1 to R3 are for the right arm. $C1_L$ to $C4_L$ are the guard conditions for the left arm and $C1_R$ to $C4_R$ are for the right arm. The guard condition for the same level of state is in the same form for both arms since a participant can perform this gesture with either arm. The definitions below are not repeated for two arms separately. Using the variables in Table II.1, $C1$ to $C3$ are defined as:

$$C1 = w_y > e_y \ for\ n\ continous\ frames \wedge H > T_{height1}\ , \qquad \text{(II-4)}$$

$$C2 = e_y > s_y \wedge \angle wes_y > T_{ang1}\ for\ n\ continous\ frames\ , \qquad \text{(II-5)}$$

$$C3 = \angle wes > T_{ang2} \wedge |\pi - \angle a1| < T_{ang3}\ for\ n\ continous\ frames\ . \qquad \text{(II-6)}$$

A gesture is considered successful if it is done within an appropriate time period. At the end of the time window, the guard condition “time out” (TO) is provided to terminate the recognition process if the arm has not reached the next state.

This gesture is graded in a 5-point scale according to the following rule: raising one hand should be done with only one arm. So if only one arm is raised to state 3 and the other arm is kept below state 2, it gets a score 4. If this raised hand is held still in state 3 for a certain amount of time, it is scored 5. However, there are scores for partial completion as well. If one or both arms are raised to state 1 but no further, it is scored 1. If one or both arms are raised to state 2, the score is 2, and if one arm is in state 3 and the other arm is in state 2 or 3, the score is 3.

2.5.2 *Waving one hand*

This gesture includes (Fig. II.3(b)): a) start raising a hand; b) further move the hand higher than the shoulder; c) wave the raised hand to one side; and d) wave the same hand to the other side. Fig. II.4(b) shows its FSM representation.

$C1$ to $C4$ are defined as follows,

$$C1 = w_y > e_y \ for\ n\ continous\ frames \wedge H > T_{height2}\ , \tag{II-7}$$

$$C2 = e_y > s_y \ for\ n\ continous\ frames\ , \tag{II-8}$$

$$C3 = D > T_{dis}\ in\ one\ direction\ , \tag{II-9}$$

$$C4 = D > T_{dis}\ in\ the\ other\ direction\ . \tag{II-10}$$

In this case, if one or both arms are raised to state 1, the gesture gets a score 1. Both arms reaching state 2 leads to a score 2. Waving should also be done with only one arm, so if one arm gets to state 2 to 4 while the other is below state 2, this performance gets a score between 3 and 5.

2.5.3 *Raising two hands*

This gesture is similar to raising one hand (Fig. II.3(c)), so the FSM graph is the same as Fig. II.4(a). However, it requires the raising of both hands. For either arm, if it reaches state 1 to 3, it gets a score from 1 to 3, respectively. If the hand is held still in state 3, it gets a score 4. Since this gesture should be done with both arms, the final score is the average score of the left and right arms. For example, if only one arm is fully raised and the other is not raised at all, the final score is 2.

2.5.4 *Reaching arms out*

Reaching arms out follows these steps (Fig. II.3(d)): a) raise arms up to shoulder level; b) get the raised arms sideways; and c) stretch them sideways.

Its FSM graph looks exactly as that in Fig. II.4(a), except the guard conditions, which are:

$$C1 = |\frac{\pi}{2} - \angle a1| < T_{ang4}\ for\ n\ continous\ frames \wedge H > T_{height3}\ , \tag{II-11}$$

$$C2 = |\angle a4| < T_{ang5}\ for\ n\ continous\ frames\ , \tag{II-12}$$

$$C3 = \angle wes > T_{ang6}\ for\ n\ continous\ frames\ . \tag{II-13}$$

The score for either arm equals to the states it reaches. Similar to raising two hands, the final score in this case is the average score of the left and the right arms.

We can see that all the gestures use similar FSM structure but with different guard conditions. All the parameters are adjustable for different application environments and user groups. We used the following values: $n = 10$; $T_{height1} = 20cm$; $T_{height2} = 20cm$; $T_{height3} = 10cm$; $T_{dis} = 20cm$; $T_{ang1} = \pi / 2$, $T_{ang2} = 3\pi / 4$; $T_{ang3} = \pi / 6$; $T_{ang4} = \pi / 6$; $T_{ang5} = \pi / 6$, $T_{ang6} = \pi / 4$ for the user study. These values were chosen by the clinicians and engineers involved in this project based on the ability of the participant group. It is important that the gesture recognition algorithm runs in parallel with the robot's gesture demonstration task. In this way, even if a participant finishes a gesture before the robot finishes its own gesture prompt, a reward will be given and the robot will stop prompting. If the robot were to continue prompting the child (even though the child might have completed the required gesture) and only give rewards after the prompting was over, the child might feel frustrated.

2.6 Experimental Setup

We conducted a user study to assess user acceptance and performance of RISIA1. The study was approved by the Vanderbilt Institutional Review Board (IRB). The experiment room is shown in Fig. II.5. The participant was seated about 120cm from the Kinect and 150cm from the administrator.

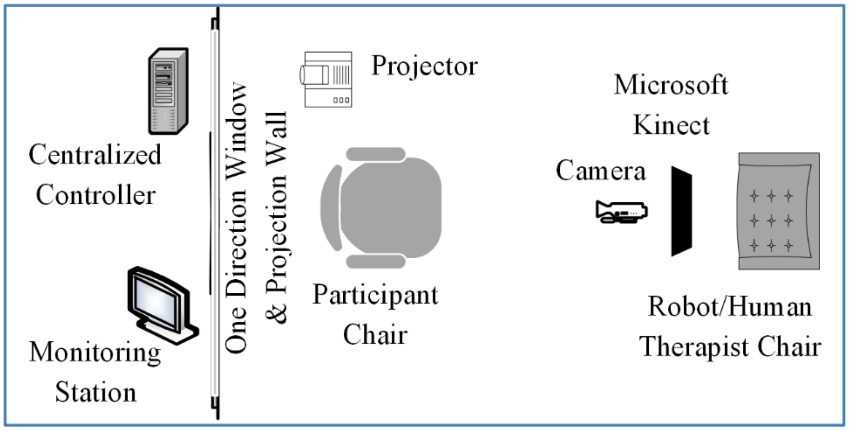


Fig. II.5 Schematic of the Experiment Room

2.6.1 Participants for the user study

Twelve children with ASD and 10 typically developing (TD) children were originally recruited to participate in this experiment. However, 4 children with ASD and 2 TD children did not complete the study. Two ASD children refused to sit in the experiment chair and thus did not start the experiment. Two other ASD children and the two TD children exhibited mild distress in the protocol and were withdrawn from the study. Group characteristics of the participants who completed the study are shown in Table II.2. The ASD group had received a clinical diagnosis of ASD based on DSM-IV-TR [21] criteria from a licensed psychologist, met the spectrum cut-off of the Autism Diagnostic Observation Schedule Comparison Score (ADOS CS) [22], and had existing data regarding cognitive abilities from the Mullen Scales of Early Learning, Early Learning Composite (MSEL) [23] in the registry. Parents of participants in both groups completed the Social Responsiveness Scale– Second Edition (SRS-2) [24], and Social Communication Questionnaire Lifetime Total Score (SCQ) [25] to index current ASD symptoms.

Table II.2 Characteristics of Participants

Mean (SD)	***ADOS CS***	***MSEL***	***SRS-2***	***SCQ***	***Age (Year)***
ASD	7.63 (1.69)	64.75 (22.11)	75.29 (12.62)	17.88 (6.58)	3.83 (0.54)
TD	NA	NA	42.75 (10.08)	3.88 (2.95)	3.61 (0.64)

2.6.2 Task and protocol

We wanted to assess how a RISIA1-based robotic system compares with human therapist-based imitation intervention. We hypothesized that the robotic system would elicit imitation performance and garner interest from children with ASD as well as a human therapist. To test this hypothesis, we conducted 2 human-administered sessions and 2 robot-administered sessions for each participant. In each group, one-half of the participants followed the order: robot session 1, human session 1, robot session 2, and human session 2 while the other half had human session 1, robot session 1, human session 2, and robot session 2. The human administrator was not present in robot sessions and the robot was not present in human administered sessions. Each session tested 2 gestures, and each gesture was tested in 2 trials. All 4 gestures were exhaustively tested in a randomized order. We compared participants' performance between the robot-administered sessions and the human-administered sessions for 1) gesture imitation performance and 2) attention towards the administrators (robot or human).

In the robot sessions, as shown in Fig. II.6, prior to the practice of each gesture, the robot initiated a mirroring interaction segment for 15 seconds with the verbal prompt, "Let's play! I will copy you!". In this segment, the robot copied a participant's arm gesture to the best of its motor capability. This was designed to maintain interest of the children on the robot and provided a break between the imitation intervention of different gestures. Following that, the child was asked to imitate the gestures of the robot in two trials.

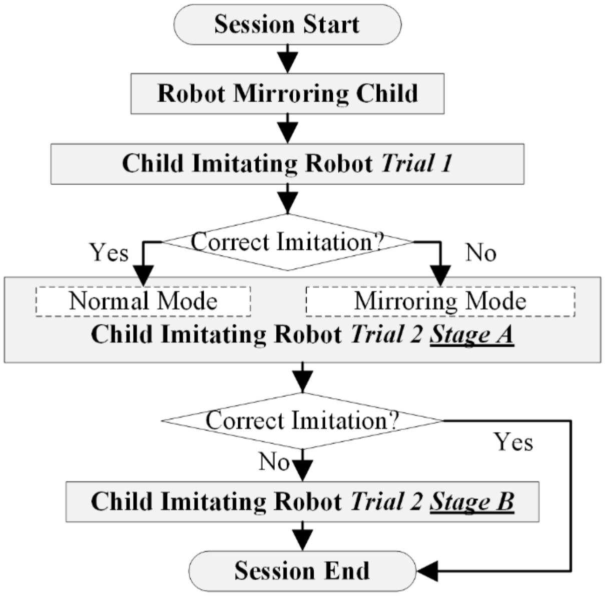


Fig. II.6 Flow of the Imitation Training Procedure for Each Gesture

In Trial 1, the robot said, “Okay! Now you copy me. Look at what I am doing!” and demonstrated the gesture twice, and prompted “You do it!”. The proposed gesture recognition was initiated immediately upon the first demonstration and ended 5 seconds following the second demonstration. As soon as the participant imitated the gesture correctly, the trial was terminated with a verbal praise, “Good job!”. The system recorded the performance score. Otherwise, the system provided feedback on the approximation if applicable, and recorded the best score the participant got. Consider the gesture “raising one arm” as an example. If the participant did not raise his/her arm high enough within the given time limit, the robot would take its (i.e., the robot’s) arm at the participant’s best raised position and gave a verbal response, “you were here”, and then would raise its (i.e., the robot’s) arm further until the desired height with the verbal response, “higher!”.

Trial 2 included Stage A and Stage B (Fig. II.6). If the participant succeeded in Trial 1, then the Trial 2 executed Stage A in Normal mode, which was the repeated procedure from Trial 1. Otherwise, Trial 2 executed Stage A in Mirroring mode, where the robot mirrored the participant’s motion after gesture demonstration. For instance, if the participant was waving, the robot would wave its arm to follow the child. This mirroring helped the children check on their own performance. If the child imitated the gesture successfully in Stage A, a verbal reward was given and Stage B was omitted. Otherwise, after the robot told the child where he/she was wrong, Stage B was presented. It provided the final two gesture demonstrations and another 2 seconds following the gesture demonstration as the final response time. Without Stage B, the child would be frustrated since no chance was left to try the gesture again. However, this procedure should not be repeated too many times since the child would lose interest in doing one gesture continuously.

In human-administered sessions, the supervisory controller computed all of the information needed for the human administrator as it would for the robot in the robot-administrated session, which included the grading of imitation performance of the participant and what to respond to the participant. These messages were projected on the wall behind the participant, and thus the human administrator could read and follow those instructions while still looking in the participant's direction. The human administrator did not make any personal judgment.

Eye gaze approximation via head pose was a coarse indicator of a person's attention. It was estimated by the Kinect tracking module. We assumed that the participant's attention was on the administrator if his/her head pose was oriented towards the attention box discussed in section VI.A.

2.7 Experimental Results

2.7.1 System validation results

In order to validate the accuracy of the proposed gesture recognition algorithm, 7 adults and 3 typically developing (TD) children were recruited. Each participant performed each of the 4 gestures 10 times under the experimental conditions. Each participant was also instructed to perform some non-specific movements during testing, and slightly shift their front facing postures between gesturing to create a naturalistic condition. The gesture recognition algorithm classified the performed gestures into one of the 4 categories or a "not recognized" category. These recognition results were compared with the subjective ratings of a therapist and got the overall accuracy of 98% (1.5% false positive and 0.5% false negative) [10]. In the few cases where it failed, it was mainly due to the tracking failure of the Kinect when the subjects quickly shifted their postures.

The system also inferred a participant's attention to the task administrator (i.e., either the robot or a human therapist) based on where he was looking. The robot height is similar to the human therapist's upper body height. A box of 85.77 cm×102.42 cm around the robot and the upper body of human therapist was set as the target attention regions. The gaze was approximated based on head pose estimation. To test the attention inference method, those same participants were asked to first look at the bounding box covering the region where an administrator would stand with their natural head pose. These head poses were reordered as the baseline data. Then the participants were asked to look away and back to the region for 10 times. Their raw head poses were normalized by their baseline values and those rectified poses were computed to see if they were oriented towards the administrator region. The results show that for 91% (2% false positive, 7% false negative) of the times, their head pose indicated the gaze towards the administrator region. In the user study, all participants' natural head poses were also calibrated in the same way.

2.7.2 Preferential attention towards the administrator

Attention to the administrator is a marker for eventual learning within intervention paradigms. On average, the ASD group paid attention to the robot and the human therapist for 55.01 (SD: 28.42) seconds and 43.32 (SD: 25.47) seconds per session, respectively. The TD group paid attention to the robot and human therapist for 61.35 (SD 28.89) seconds and 47.02 (SD: 17.30) seconds per session, respectively. The duration of a session depended on the participant's performance in that session, and each participant had different imitation abilities. The ASD group required similar amounts of time to complete the tasks across robot sessions (Avg = 105.52, SD = 24.47 seconds) and human sessions (Avg = 104.35, SD = 23.51 seconds), while the TD group required more time to complete the robot sessions (Avg = 99.69, SD = 29.76 seconds) than the human sessions (Avg = 86.67, SD = 28.64 seconds). Therefore, the ratios of the duration of attention on the administrator to the total session length was used as a normalized representation of how much attention the participants paid to the administrator, which are shown in Table II.3 for both groups.

We can see that participants in the ASD group spent 11% more time attending to the robot than the human therapist, while participants in the TD group paid similar attention to the robot and the human therapist across sessions. However, Wilcoxon signed rank test shows that the differences were not statistically significant for either group with p = 0.0663 for the ASD group and p = 0.7367 for the TD group. These results, although only approaching significant, support a part of our hypothesis in that the children with ASD paid more attention to the robot administrator than the human administrator, which indicate the potential for such a system to garner interest in imitation intervention.

Table II.3 The Ratio of the Duration of Attention on the Administrator to the Total Session Time (%)

Mean(SD)	***Robot session***	***Human session***
ASD	52.38% (24.23%)	41.38% (21.27%)
TD	63.50% (23.53%)	61.59% (29.34%)

2.7.3 Gesture imitation performance

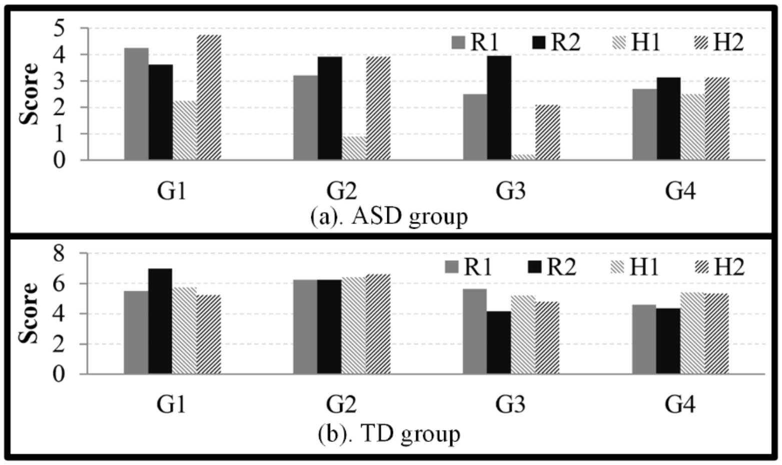


Fig. II.7 Group performance of individual gestures

Next we analyzed the demonstrated imitation skills of both groups in human and robot administered sessions. The score of each gesture in each trial was normalized to [0,10]. Table II.4(a) lists the group performance between ASD and TD across sessions. For each participant, their imitation scores for all 4 gestures in both Trial1 and Trial2 for both robot-administered and human-administered sessions were added together and presented in Table II.4(a) to show the overall performance. Further trial by trial analysis is presented in Table II.4(b). The average scores of all the robot-administrated Trial1 (R1) and Trial2 (R2) as well as human administrated Trial1 (H1) and Trial2 (H2) were computed for each group. Fig. II.7 shows the group performance on each gesture. G1 to G4 represent raising one hand, raising two hands, wave, and reaching arms out, respectively.

Consistent with previously demonstrated imitation deficits in individuals with ASD, results show that the ASD group was less successful than the TD group in general. Participants in the ASD group performed better in robot-administered sessions than in human-administered sessions, especially in Trial1 for G1 to G3. Among the 4 gestures, we can see that wave got the lowest scores in both group due to its complexity. Participants in the TD group did not demonstrate much difference across trials in all the sessions.

Table II.4(c) give the statistical *p* values of the imitation performance for each group. We can see that the ASD group's performance in the robot sessions was not statistically significantly different between Trial1 and Trial2, while that of human sessions it was significant. The result of robot-administered Trial1 was significantly better than that of human-administered Trial1, while the two Trial2s' performance were not significantly different. The TD group's performance was similar across all the trials. However, putting Trial1 together with Trial2, the statistical analysis of robot-administered session vs. human-administered session showed non-significant results for both ASD (p=0.5781) and TD (p=0.6406) groups. Table II.4(d) lists the ASD vs. TD group comparison across different trials. We can see that the main differences were on Trial1 for both robot-administered sessions and human-administered sessions.

Table II.4 Imitation Performance Results—Gesture Scores

(a). General Session Performance Results

Mean (SD)	***Robot session scores***	***Human session scores***
ASD	27.31 (32.07)	19.75 (13.64)
TD	43.75 (28.26)	44.79 (31.98)

(b). Performance Results in Trials

Mean (SD)	***R1 scores***	***R2 scores***	***H1 scores***	***H2 scores***
ASD	3.17 (4.43)	3.66 (4.31)	1.46 (2.55)	3.47 (3.67)
TD	5.49 (4.32)	5.45 (4.37)	5.70 (4.50)	5.50 (4.73)

(c). Wilcoxon Signed-Rank p Values on Trial vs. Trial Performance within Group

Trial vs. Trial	***ASD***	***TD***	***Trial vs. Trial***	***ASD***	***TD***
R1 vs. R2	0.2793	0.9893	***H1 vs. H2***	**0.0006**	0.6318
R1 vs. H1	**0.0494**	0.6946	***H2 vs. R2***	0.8669	0.9811

(d). Mann-Whitney U Test p Values on ASD vs. TD Trial Performance

Group vs. Group	***R1***	***R2***	***H1***	***H2***
ASD vs. TD	**0.0277**	0.0905	**0.0003**	0.1473

These results show that in robot-administrated sessions, children with ASD showed better imitation skills more quickly than they did with human mediated sessions. Initial interactions with the human therapist yielded significantly lower imitation scores than interactions with the robot in Trial1. In Trial2, the difference between human and robot-administrated session scores decreased, but the average scores for human sessions were still lower. There was no significant difference for the TD group across either condition throughout the whole experiment.

2.8 Discussion and Conclusion

In this chapter, we presented the design and development of a novel robot-mediated imitation skill intervention system, RISIA1, with potential relevance to core areas of deficit in young children with ASD.

RISIA1 was suitable for both a robot and a human therapist administrator. It detected the participant's imitation performance in real-time and fed this back to the administrator for adaptive intervention. Within this proof of concept experiment we also replicated previous findings demonstrating that young children with ASD paid more attention to the robot administrator and performed better in robot facilitated imitation tasks than in human-administered sessions under the same experimental protocol.

A particular strength of the RISIA1 is its use of a non-invasive configuration that does not require the participants to wear any physical sensors. This is extremely important for young children with ASD, as they can find wearable sensors uncomfortable and distracting. Another important contribution of this work relates our modeling method affording for closed-loop interaction. The FSM-based gesture recognition method that we designed allowed us to obtain real-time evaluation of participant performance and provide adaptive and individualized feedback on different levels of imitation completion. Such extension is bolstered by the fact that the FSM recognition method does not require specific training data from children to be gathered prior to participation.

In terms of performance within the system, most children with ASD and TD children were able to respond with some degree of accuracy to prompts delivered by a humanoid robot and a human administrator within the standardized protocol. Children with ASD paid more attention to the robot than the human administrator, a finding replicating previous work suggesting attentional preferences for robotic interactions over brief intervals of time. Further, some young children with ASD seemed to demonstrate enhanced performance in response to robotic prompts than those delivered by human counterparts. This suggests that robotic systems endowed with enhancements for successfully capitalizing on baseline enhancements in non-social attention preference might be utilized to meaningfully enhance skills related to core symptoms of ASD. Although this work does not demonstrate generalization beyond the experimental sessions, this documented preferential attention could potentially be harnessed to drive towards such an outcome. Future work examining more in-depth prompt and reinforcement strategies, upgrading and accommodating the system into a formal clinical study, including more gestures, and combining gesture imitation with other meaningful daily tasks would likely enhance future applications of this system.

There are several methodological limitations of the current study that are important to highlight. The small sample size examined and the limited time frame of interaction restricted our ability to realistically comment on the value and ultimate clinical utility of this system as applied to young children with ASD. Further, the brief exposure of the current paradigm, in combination with unclear baseline skills of participating children, ultimately cannot answer questions as to whether heightened attention paid to the robotic system or performance differences in conditions displayed during the study are simply the artifact of novelty or of a more characteristic pattern of preference that could be harnessed over time. We also did not explore test-retest reliability in this preliminary study. Regarding gesture recognition, Kinect has a limited range and thus puts constraints on the set of gestures that can be used for imitation tasks. Therefore, extending the range of Kinect is to be explored to improve the system capabilities. In addition, developing an optimization algorithm for autonomous selection of the parameters of the FSM based will further enhance the system. Another important technical limitation was the approximation of attention with head pose. It must be emphasized that head orientation approximating gaze or attention does not necessarily equate to actual eye gaze or by extension, attention. However, such data does provide a coarse proxy for documenting feasibility. In terms of the robot, Nao was not suitable for very fast paced motion due to its limited motor ability. It was programmed to provide intermittent verbal prompts and rewards but did not engage the participants in continuous verbal communication. There could be some benefits in engaging the children continuously through verbal communication, however, in this study we thought that continuous verbal communication might distract the participants from imitating gestures.

Despite limitations, this work is the first to our knowledge to design and empirically evaluate the usability, feasibility, and preliminary efficacy of a non-invasive closed-loop interactive robotic technology capable of modifying response based on within system measurements of performance on imitation tasks with young children. Movement in this direction introduces the possibility of realized technological intervention tools that are not simple response systems, but systems that are capable of necessary and more sophisticated adaptations. Our platform represents a move toward realistic deployment of technology capable of accelerating and priming a child for learning in key areas of deficit.

2.9 References

[1] J. H. Williams, A. Whiten, T. Suddendorf, and D. I. Perrett, “Imitation, mirror neurons and autism,” *Neuroscience & Biobehavioral Reviews,* vol. 25, no. 4, pp. 287-295, 2001.

[2] A. Whiten, and J. Brown, “Imitation and the reading of other minds: Perspectives from the study of autism, normal children and non-human primates,” *Intersubjective communication and emotion in early ontogeny*, pp. 260-280, 1998.

[3] B. Ingersoll, “Brief report: Effect of a focused imitation intervention on social functioning in children with autism,” *Journal of autism and developmental disorders,* vol. 42, no. 8, pp. 1768-1773, 2012.

[4] B. Ingersoll, “Pilot Randomized Controlled Trial of Reciprocal Imitation Training for Teaching Elicited and Spontaneous Imitation to Children with Autism. ,” *Journal of Autism and Developmental Disorders,* vol. 40, no. 1154-1160, 2010.

[5] A. Duquette, F. Michaud, and H. Mercier, “Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism,” *Autonomous Robots,* vol. 24, no. 2, pp. 147-157, 2008.

[6] I. Ranatunga, M. Beltran, N. A. Torres, N. Bugnariu, R. M. Patterson, C. Garver, and D. O. Popa, "Human-robot upper body gesture imitation analysis for autism spectrum disorders," *Social Robotics*, pp. 218-228: Springer, 2013.

[7] J.-J. Cabibihan, H. Javed, M. Ang Jr, and S. M. Aljunied, “Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism,” *International journal of social robotics,* vol. 5, no. 4, pp. 593-618, 2013.

[8] B. Robins, K. Dautenhahn, R. Boekhorst, and A. Billard, “Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?,” *Universal Access in the Information Society,* vol. 4, no. 2, pp. 105-120, 2005.

[9] S. M. Srinivasan, K. A. Lynch, D. J. Bubela, T. D. Gifford, and A. N. Bhat, “Effect of Interactions Between a Child and a Robot on the Imitation and Praxis Performance of Typically Developing children and a Child with Autism: A Preliminary Study,” *Perceptual & Motor Skills,* vol. 116, no. 3, pp. 885-904, 2013.

[10] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, "Autonomous robot-mediated imitation learning for children with autism." pp. 2707-2712.

[11] Z. Zheng, L. Zhang, E. Bekele, A. Swanson, J. Crittendon, Z. Warren, and N. Sarkar, "Impact of Robot-mediated Interaction System on Joint Attention Skills for Children with Autism ".

[12] J. Wainer, K. Dautenhahn, B. Robins, and F. 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,* vol. 6, no. 1, pp. 45-65, 2014.

[13] "Aldebaran Robotics," http://www.aldebaran-robotics.com/en/.

[14] "Microsoft Kinect for Windows," http://www.microsoft.com/en-us/kinectforwindows/.

[15] M. Livingston, J. Sebastian, Z. Ai, and J. W. Decker, "Performance measurements for the Microsoft Kinect skeleton." pp. 119-120.

[16] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, "Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population." pp. 1188-1193.

[17] E. A. Lee, and S. A. Seshia, *Introduction to embedded systems: A cyber-physical systems approach*: Lee & Seshia, 2011.

[18] A. D. Wilson, and A. F. Bobick, “Parametric hidden markov models for gesture recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 21, no. 9, pp. 884-900, 1999.

[19] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering." pp. 126-133.

[20] S. Mitra, and T. Acharya, “Gesture recognition: A survey,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 37, no. 3, pp. 311-324, 2007.

[21] *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the Diagnostic Criteria from DSM-IV-TR*, Fourth ed., Washington D.C.: American Psychiatric Association, 2000.

[22] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[23] E. M. Mullen, *Mullen scales of early learning: AGS edition*, Circle Pines, MN: American Guidance Service, 1995.

[24] J. N. Constantino, and C. P. Gruber, "The social responsiveness scale," Los Angeles: Western Psychological Services, 2002.

[25] M. Rutter, A. Bailey, and C. Lord, "The Social Communication Questionnaire," Los Angeles, CA: Western Psychological Services, 2010.

Chapter III. Robot-mediated Mixed Gestures Imitation Intervention

3.1 Abstract

In this chapter, we propose another robotic platform that mediates imitation skill intervention for young children with ASD. While a few previous works (e.g., RISIA1) have provided methods for single gesture imitation intervention, the current chapter extends the intervention to incorporate mixed gestures consisting of multiple single gestures during intervention. A preliminary user study showed that the proposed robotic system, named RISIA2, was able to stimulate mixed gesture imitation in young children with ASD with promising gesture recognition accuracy.

Keywords— robot assisted intervention; autism spectrum disorder; gesture recognition; imitation

3.2 Introduction

Several studies have shown that teaching imitation skills to children with ASD through the use of robotic technologies is feasible and has great potential [1, 2]. Dautenhahn et al. developed a humanoid robot KASPAR, which was able to interact with children with ASD using imitation games [3]. Fujimoto et al. designed techniques using wearable sensors for mimicking and evaluating human motion in real time to improve imitation skills of children with ASD [4]. Greczek et al. proposed a graded cuing mechanism to encourage the imitation behavior of children with ASD in a closed-loop "copy-cat" game [2]. Zheng et al. created a robotic system (i.e., RASIA1 discussed in Chapter II) that provided imitation intervention for children with ASD with online feedback regarding the quality of gesture accomplished [5]. While the above-mentioned studies were important in establishing the feasibility and usefulness of robot-mediated systems for imitation skills intervention, they focus on simple, single-gesture based imitation skills. In reality, a child is expected to learn more complex gestures that can be combinations of a set of simple gestures. In this chapter, we present a framework for robot-mediated imitation skill intervention for complex mixed gestures. Moreover, this current work utilizes a non-invasive setup that did not require the children to wear any physical sensors, since many children with ASD tend to reject body-attached hardware[6].

This chapter presents a new gesture recognition method capable of detecting mixed gestures, which are defined as simultaneous execution of multiple simple gestures from a participant, as well as identifying ("spotting") the start and the end of each detected gesture. A new intervention protocol was designed to test this algorithm and a preliminary user study with children with ASD and their typically developing (TD) peers was conducted to show the feasibility and potential of this robotic system.

The rest of this article is organized as follows. Section 3.3 describes the development of the robot-mediated intervention system. Section 3.4 features the experimental setup. Section 3.5 presents the experimental results, followed by the authors' conclusions of this work in Section 3.6.

3.3 System Development

3.3.1 System architecture

The proposed robot-mediated imitation skill intervention architecture 2 (RISIA2) consists of a robot module and a gesture tracking module that were operated based on the commands sent from a supervisory controller. The robot module utilized the humanoid robot NAO [7]. NAO is about 58cm high and has a childlike appearance. It is built with 25 degrees of freedom, flashing LED "eyes", speakers, multiple sensors, and a well maintained software development kit (SDK). We chose this robot due to its attractive appearance to the children, simplified but adequate motion range and patterns, as well as the stability and flexibility of its software development environment. NAO communicated with the participants using both speech and motion. Its default text-to-speech functions and voice were used to provide verbal instructions. The physical motions needed in the experiments were preprogrammed and stored in a software library, and were called whenever needed.

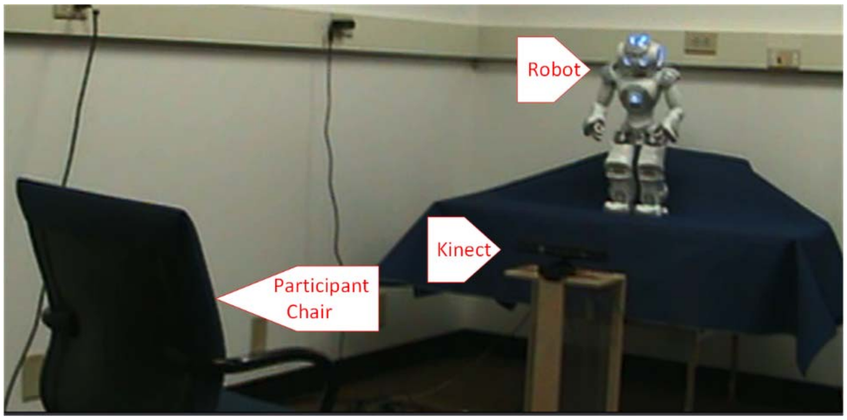


Fig. III.1 RISIA2 system picture

The gesture tracking module used Microsoft Kinect [8]. Its SDK provides robust functions for real time skeleton tracking and head pose estimation. Skeleton data were used for imitation performance evaluation, and the head pose was treated as a coarse attention indicator which revealed how much attention the participant paid to the robot. The supervisory controller was in charge of the system execution logic, communication, and data logging. The robot module and the gesture tracking module were distributed and they communicated with the supervisory controller using different threads to achieve parallel operation. In this way, the system was able to both monitor the performance of the participant and provide different prompts to the child. The collected data were logged for offline analysis.

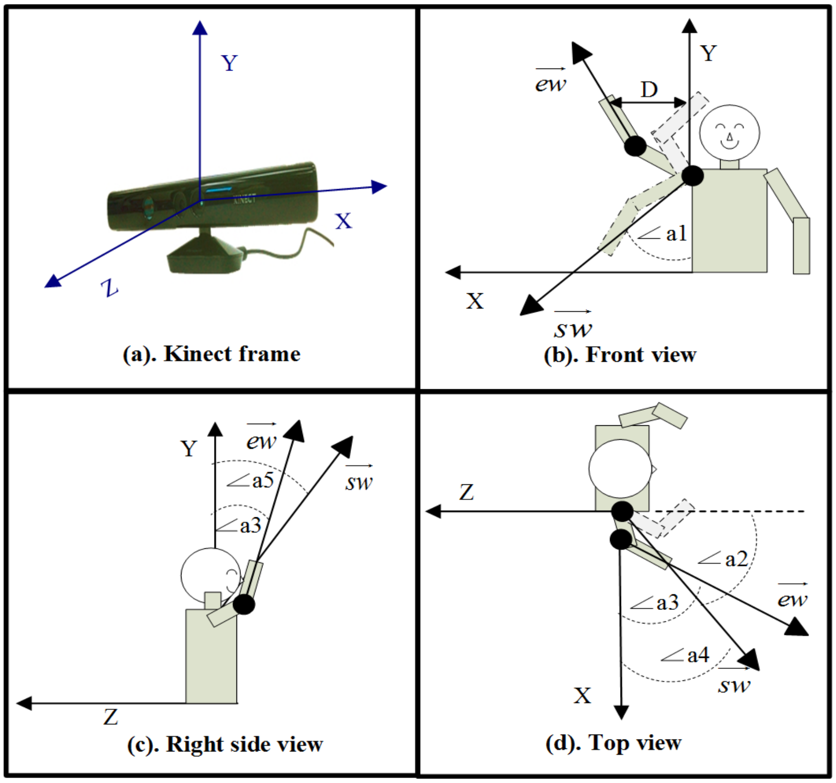
The participants were seated facing the robot about 2m away, and the Kinect was placed between the participant's chair and the robot. Fig. III.1 shows the implemented system.

3.3.2 Single gesture recognition

Table III.1 Gesture Variables

Symbol	***Definition***
$\overrightarrow{SW}$, $\overrightarrow{EW}$	Vector from shoulder to wrist, Vector from elbow to wrist
W_y, E_y, S_y	Y coordinates of wrist, elbow, and shoulder joint
$\angle A1$	Angle between $\overrightarrow{SW}$ and negative Y axis
$\angle A2$	Angle between $\overrightarrow{SW}$ and YZ plane, when arm pointing forward
$\angle A3$	Angle between $\overrightarrow{EW}$ and XY plane
$\angle A4$	Angle between $\overrightarrow{SW}$ and positive X axis (right arm) or negative X axis (left arm).
$\angle A5$	Angle between $\overrightarrow{SW}$ and XY plane,
$\angle WES$	Angle between upper arm and forearm
H	Wrist raised height in Y direction
D	Wrist movement in X direction
T_{item}	Threshold for distances or angles
$(R)_{nf}$	Condition R in the parentheses should be held for *n* consecutive frames

The single gesture recognition method (SGR) proposed by Zheng et al. [5] (i.e., RASIA1 in Chapter II) was used as a basic component of the proposed mixed gesture recognition and spotting algorithm. In the SGR, the input is a temporal sequence of gesture variables (as listed in Table III.1), which are computed from the subject's arm skeleton tracking data. Fig. III.2 (b-d) shows some of gesture variables of the right arm as examples.



(a). Kinect frame

(b). Front view

(c). Right side view

(d). Top view

Fig. III.2 Kinect frame and participant's gesture view

A correct gesture is defined by trajectory constraints (TC) under preconditions (PC). PC describes the basic regional and positional constraints of a gesture. Since a participant was instructed to follow the robot's gesture, the TC was defined as multiple gesture stages in accordance with the order in which the robot presented a gesture. The recognition of a stage is triggered by the completion of the previous stage(s). The output of SGR is the gesture stage computed based on the input data.

Four gestures were studied in the original work: raising one hand (Gesture 1), raising two hands (Gesture 2), waving (Gesture 3) and reaching arms out (Gesture 4). Gesture 1 and Gesture 3 can be accomplished by either the right or the left hand/arm. If the imitation data satisfy the first *n* TCs, the performance is graded as $n \times 10 / \text{number of } TC$. Gesture 2 and Gesture 4 need to be accomplished by both hands to receive a full score. The grades for two hands were averaged for the final score. The PCs and TCs for each gesture applied in the current study are shown Eqn. (III-1) – (III-6). These rules together with the gesture variables were preselected by experienced psychologists and engineers based on the analysis of 16 children's gesture performing data in the repository (8 children with ASD and 8 typically developing children, ages 2-5 years). This method utilizes the most important features of the selected gestures while keeping a low computational complexity of the gesture recognition module.

$$PC_{Wave} = \{\angle A1 < T_{ang1} \vee \angle A2 < T_{ang2} \vee W_y > S_y \vee \angle A3 < T_{ang3}\} \quad \text{(III-1)}$$

$$\begin{aligned} TC_{Wave} = \{ & (W_y > E_y)_{nf} \wedge \mathrm{H{>}T}_{up}, (\mathrm{W}_y > S_y)_{nf}, (\text{Only one hand})_{nf}, \\ & \mathrm{D{>}T}_{dis} \text{in one direction}, \mathrm{D{>}T}_{dis} \text{in both directions } \} \end{aligned} \quad \text{(III-2)}$$

PC_{Wave} Implies the gesture shall be started from raising an arm in front of the body. The first 3 constraints of TC_{Wave} requires that only one wrist shall be raised until higher than the shoulder. The last 2 constraints of TC_{Wave} indicate that the raised hand should be waved from side to side.

$$PC_{RaiseHand(s)} = \{\angle A1 < T_{ang3} \vee \angle A2 < T_{ang2}\} \quad \text{(III-3)}$$

$$\begin{aligned} TC_{RaiseHand(s)} = \{ & (W_y > E_y)_{nf} \wedge \mathrm{H{>}}T_{up}, (\mathrm{W}_y > S_y \wedge \angle WES > T_{ang4})_{nf}, (\angle WES > T_{ang5} \wedge (\pi - \angle A1) < T_{ang6})_{nf}, \\ & (\text{Only One Hand (raising one hand gesture)})_{nf}, (\mathrm{D{<}T}_{dis})_{nf} \} \end{aligned} \quad \text{(III-4)}$$

These conditions represent that the hand(s) shall be raised from low to high in front of the body ($PC_{RaiseHand(s)}$) until they are gradually stretched straight and held still for a while ($TC_{RaiseHand(s)}$).

$$\begin{aligned} PC_{ReachArmsOut} = \{ & \angle A1 < T_{ang3} \vee \angle A4 < T_{ang3} \vee \angle A5 < T_{ang2} \vee \overrightarrow{SW}_y < 0 \vee (\overrightarrow{SW}_x > 0 \text{ (right arm) or} \\ & \overrightarrow{SW}_x < 0 \text{ (left arm))}\} \end{aligned} \quad \text{(III-5)}$$

$$TC_{ReachArmsOut} = \{| \pi/2 - \angle A1 |< T_{ang7})_{nf},\ (|\angle A4 |< T_{ang8})_{nf}, (\angle WES > T_{ang9})_{nf}\} \quad \text{(III-6)}$$

These rules imply that the arms shall be raised from a low position at the side of the body ($PC_{ReachArmsOut}$), and then stretched out evenly on each side ($TC_{ReachArmsOut}$).

In this work $n = rL$, where $r = 0.8$ and L is the length of the sliding time window. This indicates that the continuous constraint has to satisfy as least 80% of the sliding time window. This is important for gesture spotting discussed later. In the experimental study, psychologists and engineers found that the threshold values listed in Table III.2 were suitable for the participants. Note that those values may need to be adjusted for other user studies with different participant groups.

A mixed gesture recognition algorithm is introduced within the following context. The robot demonstrates a continuous sequence of the four previously mentioned single gestures, and asks the child to imitate. The child might start and stop at any time and may or may not imitate all the demonstrated gestures. The task for the robot was to recognize when and what gestures were imitated as well as the quality of the imitation.

3.3.3 *Mixed gesture recognition and spotting (MGRS)*

The newly proposed MGRS solves two problems in the mixed gesture prompting environment: a) how to recognize different gestures in parallel from the same input data sequence; and b) how to spot the start and the end points of each detected gesture. MGRS embeds the SGR as its components in a novel framework to address these two challenges. The four gestures described previously are presented as examples here; the proposed MGRS is not limited to those 4 gestures alone.

Table III.2 Gesture Thresholds

Variable	***value***	***Variable***	***value***
T_{up}	10cm	T_{dis}	[10,15] cm
T_{ang1} , T_{ang2}	π	T_{ang3}	$\pi/2$
T_{ang4}	$[\pi/4, \pi/3]$* $\pi/2$**	T_{ang5}	$[\pi/5, \pi/3]$* $[\pi/5, \pi/4]$**
T_{ang6}	$[\pi/6, \pi/4]$* $[\pi/4, \pi/2]$**	T_{ang7}	$\pi/4$
T_{ang8}	$[\pi/5, \pi/4]$	T_{ang9}	$[\pi/4, \pi/3]$

* Raising one hand ** Raising two hands

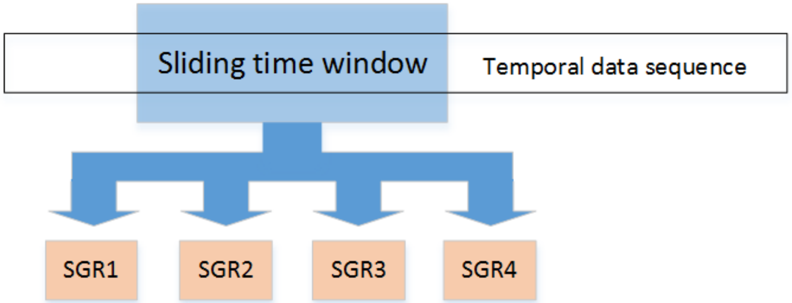


Fig. III.3 Mixed gesture recognition data flow

An initial imitation stage can evolve into different gestures. For example, both raising one hand and waving start with raising a hand from low to high. Therefore, the data subsequence extracted by a sliding time window is sent to all four SGRs for computing their stages. SGR1 to SGR4 represent the single gesture recognition algorithm for Gesture1 to Gesture4 in Fig. III.3.

A gesture is detected if a data subsequence matches with its SGR's detecting stages. The hypothesis is that the more a data subsequence matches with a gesture's stages, the better it represents the corresponding gesture. This idea is similar to correlation based template matching in computer vision [9]. The start and end of this subsequence are the start and end of the corresponding gesture, respectively. This is formally defined as searching for T_{start} and T_{end} that satisfies

$$\underset{T_{start},T_{end}}{\arg\max}\ stage = SGR(Data(T_{start},T_{end}))\ . \qquad \text{(III-7)}$$

Given a gesture's SGR, we would like to find a data subsequence $Data(T_{start},T_{end})$ which starts from T_{start} and ends at T_{end} that reaches a stage higher than the stage reached by any other subsequences overlapped with or adjacent to $Data(T_{start},T_{end})$. This local optimum can be computed by updating the sliding window's position and length to refresh the stages detected by the SGR accordingly.

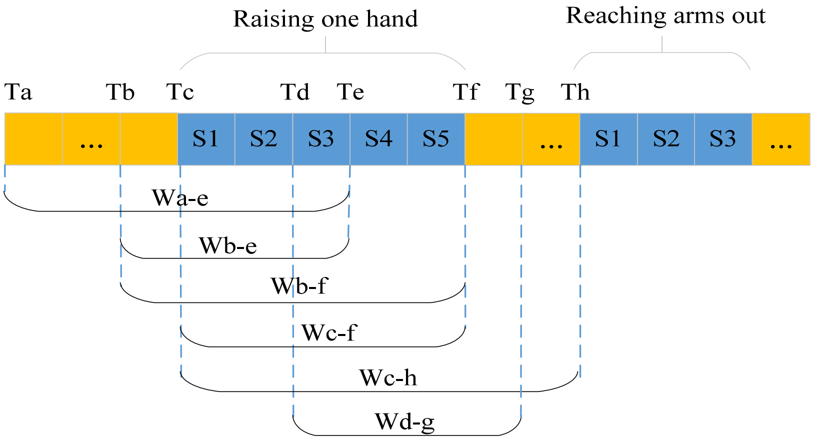


Fig. III.4 MGRS algorithm demonstration

Consider the gesture of "raising one hand" as an example. It contains five stages in its TCs. In Fig. III.4, blue blocks represent gesture imitation subsequences. The "raising one hand" gesture imitation subsequence starts from stage 1 (S1) and ends with stage 5 (S5). The yellow blocks are non-imitation data, which can be unpurposive movements, resting postures, and so on. The sliding time window is updated in two nested loops: i) shifting its start point; and ii) with the same start point, adjusting its length from 1s to 5s. A gesture can be imitated within 1s to 5s. Ta to Th represent time points along the data sequence, and Wx-y means that the sliding window's start and end time points are x and y, respectively.

The following example illustrates how the algorithm computes the correct result as “the imitation reaches S5” and “the start and end point is Tc and Tf.” The sliding time window updates from right (earlier) to left (later).

1) At Wa-e, the non-imitation data is over 20% of the window, and thus the subsequence violates the continuous constrains in the TCs. As a result, the current stage is set as 0;

2) The sliding window in then moved and at Wb-e, the window satisfies some of the TCs as only a small amount of non-imitation data is included. However, the window only contains data up to S3. So S3 is recorded as the current gesture stage, and the start and end time are noted as Tb and Te.

3) When the window length is extended to Wb-f, the window includes S5. In this case S5 is recorded as the current gesture stage, and the start and the end points are Tb and Tf.

4) When the window is moved at Wc-f, the window tightly cuts the data from S1 to S5, so the recorded state is still S5, but the start and the end points are refreshed to Tc and Tf.

5) At Wc-h, although the window includes the data from S1 to S5, but a large amount of non-imitation data after S5 prevents satisfaction of the continuous constraints. So the results of step 4 are kept.

6) At Wd-g, S1 and S2 are excluded. Because SRG does not jump previous gesture stages to reach later stages, the results of step 4 will not be refreshed.

Therefore, the final result is a gesture reaching stage 5 (S5) with a start point of Tc and end point of Tf.

This updating procedure can be executed using Algorithm1. Here *m* and *k* are time variables refreshed at each frame in the data sequence. *GestStage(m)* records the highest gesture stage that the subsequence reaches at time *m*. T_{start} (*m*) and T_{end} (*m*) mark the start and end time points of the whole gesture period where *GestStage(m)* belongs. Starting from the first frame of the data sequence, a 1 second length (LB, short for lower bound) data subsequence is used to compute an initial result using SGR. Then by appending frames onto the current sequence, 1 frame per update, SGR refreshes the results. If a higher stage is reached, it is recorded and its T_{start} and T_{end} are refreshed. This procedure is executed until the subsequence’s length reaches the 5 second upper bound (UB). At that point, the sliding time window’s start point is pushed forward by 1 frame and the above procedure is repeated. The iterations are executed until the end of the recognition period. Every gesture’s iteration is updated in parallel as a new frame’s data becomes available, and its results are recorded individually.

Algorithm1

```
GestStage(1:DataLength) = 0;
$T_{start}$ (1:DataLength) = 0;
$T_{end}$ (1:DataLength) = 0;
for k = 1 : DataLength
        [Stage1] = InitializeSGR(Data(k : k + LB));
        GestStage(k : k + LB) = Stage1;
        $T_{start}$ ( k : k + LB) = k;
        for m = k+LB+1 : k+UB
              [Stage2] = UpdateSGR(Data(m), Stage1);
              if Stage2 >= Stage1
                    GestStage(m) = Stage2;
                    $T_{start}$ (m) = k;
                    $T_{end}$ (m) = m;
              end
              Stage1 = Stage2;
        end
end
Return (GestStage, $T_{start}$ , $T_{end}$);
```

Note that this algorithm is an *example* of how to program the MGRS, but MGRS is not limited by it. Any computation procedure that reflects the goal in Eqn. (III-7) can be applied.

3.4 Experimental Setup

3.4.1 Participants

The MGRS algorithm should be able to successfully detect performed gestures as well as avoid giving false positive results when no targeted gestures are performed. Therefore, in this pilot study we selected participants with different imitation baseline levels, which helped us to collect imitation data ranging from good completion to non-completion. Two TD children (1 male and 1 female) and 4 children with ASD (3 males and 1 female) participated in this experiment. This group size is small due to the limited participant pool.

Table III.3 lists the characteristics of the participants. Those in the ASD had received a clinical diagnosis of ASD based on DSM-IV-TR [10] criteria. They met the spectrum cut-off of the Autism Diagnostic Observation Schedule Comparison Score (ADOS CS) [11], and had existing data regarding cognitive abilities from the Mullen Scales of Early Learning (MSEL) [12] in the clinical registry. Parents of participants in both groups completed the Social Responsiveness Scale– Second Edition (SRS-2) [13], and Social Communication Questionnaire Lifetime Total Score (SCQ) [14] to index current ASD symptoms.

Table III.3 Participant Characteristics

Mean (SD)	***ADOS CS***	***MSEL***	***SRS-2***	***SCQ***	***Age (Years)***
ASD	8.50 (1.73)	51.00 (4.00)	76.50 (16.60)	20.50 (7.77)	4.61(0.60)
TD	NA	NA	42.50 (3.54)	1.50 (0.71)	4.63 (0.01)

3.4.2 Task and protocol

This study was approved by the Vanderbilt Institutional Review Board (IRB). All the experiments were supervised by qualified clinicians and engineers. Videos of the experimental procedures were recorded for algorithm validation. The experiment had two steps:

Step1. Participant warmed up with the robot-administrated single gesture session "TrialA" of the previous work [5], for all 4 gestures. In this step, the robot first showed the participant a gesture, and then asked the participant to copy it. If the participant imitated the gesture correctly, the robot verbally praised the child. Otherwise, the robot provided a verbal explanation of what was wrong. This was intended to inform the participant that he/she was expected to copy the robot's gesture.

Step2. This was the mixed gesture imitation intervention part, consisting of 4 trials. Before each trial, the robot first mirrored the participant's physical motions for 15s to help the participant feel he/she was playing with the robot as a peer. Then, the robot asked the participant to now copy its gestures. The robot showed all 4 gestures twice in random order, all accompanied by background music. The "wave" gesture lasted for 4.4 seconds and the other 3 gestures lasted for 3.2 seconds. Two adjacent gestures were separated by a short transitional motion. In total, each trial lasted for about 49 seconds.

From the logged data we analyzed: 1) the accuracy of the MGRS algorithm; 2) the participant's attention on the robot; and 3) the participant's imitation performance.

3.5 Experimental Results

3.5.1 Gesture recognition and spotting results

For gestures that were successfully detected by the MGRS algorithm, the deviation between the human detected and MGRS detected start and end time was calculated. From Table III.5, we can see that the average deviation of the start and end time in all cases were smaller than 1s.

Table III.4 Comparison between MGRS Recognition and Human Coding Results

Gesture	*Human coded*	*Algorithm detected*	*False*	*Miss*
Gesture1	17	14	0	3
Gesture2	19	24	5	0
Gesture3	19	23	6	2
Gesture4	23	26	3	0
Total	78	87	14	5

Table III.5 Start and End Time Deviation for Correctly Detected Gestures in Table III.4

Mean (SD)	*Start time deviation (s)*	*End time deviation (s)*
Gesture1	0.25 (0.18)	0.50 (0.47)
Gesture2	0.46 (0.32)	0.48 (0.46)
Gesture3	0.88 (0.65)	0.97 (0.64)
Gesture4	0.33 (0.28)	0.54 (0.45)

To validate the accuracy of the MGRS algorithm, the results obtained from the MGRS were compared with an experienced therapist's ratings on the participants' imitation. From the videos recorded during the experiments, the therapist manually marked each gesture's start and end time as the ground truth. It is difficult for a therapist to identify and log different gesture stages similar to what a computer can do. Therefore, only gesture stages that were intuitively recognizable were marked. Accordingly, those marked stages were compared with those recognized by the MGRS.

We assessed two aspects of the MGRS: 1) could the algorithm identify a gesture correctly? 2) if the identification was correct, did the algorithm spot the start and end time of this gesture correctly? Table III.4 lists the number of gestures detected by the human therapist and the MGRS algorithm in the experiment. Seventy-three (5 detection misses) out of 78 human coded gestures (93.59%) were correctly detected by the MGRS, while 14 out of 87 MGRS detected gestures (16.09%) were false detections.

3.5.2 System tolerance, participants' attention on the robot, and imitation performance

All participants completed the experiments except one child with ASD who did not complete 2 trials. Thus we had data for 14 trials from the children with ASD, and 8 trials from the TD children. The small sample was not sufficient for statistical significance testing. As a result, only the mean and standard deviation values are presented here.

The attention that the children paid on the robot was closely connected to his/her imitation performance. Attention can be coarsely estimated by head pose [15]. In this study, we utilized head pose estimation to infer the degree to which a participant was attending to the robot. A box of 85.77 cm × 102.42 cm around the robot (which covered the robot's full range of motion with a small margin) was set as the attention reference region. We assumed that a participant paid attention to the robot if his/her head faced toward this defined region. Table III.4 lists the total time that each group paid attention to the robot. The ratios are the percentage of the time spent facing the robot within a trial.

Table III.6 Participant's Attention Spent on the Robot

Mean (SD)	***Time on Robot(s)***	***Ratio (%)***
ASD	19.6 (9.98)	39.97 (20.27)
TD	39.34 (3.31)	80.41 (6.76)

Due to the deficit of social communication, children with ASD usually pay significantly less attention to the social cues compared to the TD children. However, children with ASD still spent an average of 39.97% of trials facing the robot. This was less time compared to the single gesture sessions (60% in robot session, 42% in human session) as found in our previous study [5]. However, considering the increased task complexity, this result was not surprising. TD children spent a majority of trials looking at the robot since they were interested in the robot as reported by their parents and experiment supervisors.

The participants' imitation performance was analyzed based on the MGRS algorithm. The scores of each gesture were normalized to [0, 10]. On average, children with ASD got a score of 8.71 (SD: 1.59) out of 10, while TD children got 9.58 (SD: 0.93). Children with ASD took 3.79s (SD: 1.91s) to finish one gesture on average, while their TD peers took 2.8s (SD: 1.37s). The results confirmed expected results that children with ASD would have a lower gesture imitation ability compared to that of TD children of similar ages.

3.6 Discussion and Conclusion

We proposed a robotic system with the RISIA2 architecture that aims to teach imitation skills to young children with ASD. This work utilized a humanoid robot for gesture prompting and a non-invasive setup for effectively evaluating the participants' performance. This system extends the previous robot-mediated intervention from single gesture to mixed, multiple gestures. Naturally, the participant was allowed to imitate different gestures in any order, and at any time during the intervention. In order to achieve mixed gesture imitation recognition, we developed a novel algorithm, the MGRS algorithm, which not only detects the imitated gestures, but also spots the start and end times of the performed gestures.

A preliminary user study showed that the MGRS algorithm achieved high accuracy in both gesture recognition and spotting. The RISIA2 system was well tolerated by the young children, attracted their attention, and showed great potential for extending the intervention of imitation skills for children with ASD.

There were some limitations in this study. The proposed MGRS algorithm was tested with only 4 gestures. Extensive tests with more gesture categories are necessary in the future to examine MGRS's scalability and robustness. Although the SGRs embedded in the MGRS are rule-based, the framework of MGRS is not limited to rule-based components. In fact, any single gesture detection algorithm can be embedded in this framework. If a large number of gestures are needed to be detected in parallel, then the correlation and similarity between those gestures can be used for pruning to avoid repeated computation. Finally, since the user group was small, any conclusion drawn based on this study requires further validation with larger samples and more pervasive analyses. Yet the current work may provide a beneficial preliminary framework for developing and evaluating multi-gesture imitation intervention for young children with ASD.

3.7 References

[1] Z. Warren, Z. Zheng, S. Das, E. M. Young, A. Swanson, A. Weitlauf, and N. Sarkar, "Brief Report: Development of a Robotic Intervention Platform for Young Children with ASD," *Journal of autism and developmental disorders*, pp. 1-7, 2014.

[2] J. Greczek, E. Kaszubksi, A. Atrash, and M. J. Matarić, "Graded Cueing Feedback in Robot-Mediated Imitation Practice for Children with Autism Spectrum Disorders," *Proceedings, 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014) Edinburgh, Scotland, UK* Aug. 2014.

[3] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "KASPAR–a minimally expressive humanoid robot for human–robot interaction research," *Applied Bionics and Biomechanics,* vol. 6, no. 3-4, pp. 369-397, 2009.

[4] I. Fujimoto, T. Matsumoto, P. R. S. De Silva, M. Kobayashi, and M. Higashi, "Mimicking and evaluating human motion to improve the imitation skill of children with autism through a robot," *International Journal of Social Robotics,* vol. 3, no. 4, pp. 349-357, 2011.

[5] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, "Autonomous robot-mediated imitation learning for children with autism." pp. 2707-2712.

[6] E. Bekele, U. Lahiri, A. Swanson, Julie A. Crittendon, Zachary Warren, and N. Sarkar, "A Step towards Developing Adaptive Robot-mediated Intervention Architecture (ARIA) for Children with Autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 289-299, 2013

[7] "Aldebaran Robotics," http://www.aldebaran-robotics.com/en/.

[8] "Microsoft Kinect for Windows," http://www.microsoft.com/en-us/kinectforwindows/.

[9] J. P. Lewis, "Fast template matching." pp. 15-19.

[10] *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the Diagnostic Criteria from DSM-IV-TR*, Fourth ed., Washington D.C.: American Psychiatric Association, 2000.

[11] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[12] E. M. Mullen, *Mullen scales of early learning: AGS edition*, Circle Pines, MN: American Guidance Service, 1995.

[13] J. N. Constantino, and C. P. Gruber, "The social responsiveness scale," Los Angeles: Western Psychological Services, 2002.

[14] M. Rutter, A. Bailey, and C. Lord, "The Social Communication Questionnaire," Los Angeles, CA: Western Psychological Services, 2010.

[15] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar, “A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism,” *Neural Systems and Rehabilitation Engineering, IEEE Transactions on,* vol. 21, no. 2, pp. 289-299, 2013.

Chapter IV. Computer-Mediated Autonomous Social Orienting Intervention

4.1 Abstract

Social communication is among the core areas of impairment for children with Autism Spectrum Disorders (ASD). The training of social orientation is believed to be important for improving social communication of children with ASD. In this chapter, we propose a fully closed-loop autonomous computer system, named ASOTS, for training social orientation skills to very young children with ASD. A fast large range non-invasive gaze tracking algorithm was designed to detect and track a child's attention in response to social orientation bids. An intelligent attention tracking and prompting mechanism was developed to help the child towards appropriate social orientation when needed. Response to name, an important social orientation skill, was used to demonstrate the functionality of the proposed system. Ten toddlers with ASD participated in a pilot user study to show whether the system could be used on young children who have been diagnosed with ASD. Another pilot user study with 10 TD infants tested whether this system has a potential to be applied for early detection for infants who were younger than the age when ASD diagnoses can be done. This was done intentionally to separately demonstrate utility and functionality for the clinical population of interest and to demonstrate functionality beyond current clinical identification capacity (i.e., infants). The results showed that the proposed system and the protocol were well tolerated by both groups, successfully captured young children's attention, and elicited the desired behavior.

Keywords—computer-assisted system for young children with ASD, social skill training system, response to name, toddlers with ASD, TD infants

4.2 Introduction

The primary objective of this chapter is to present a novel autonomous social orienting training system (ASOTS) that could be useful in ASD intervention in the future. Among multiple aspects of social communication development, social orienting is one the most fundamental and critical skills that naturally develops in children [1]. Social orienting indicates spontaneous orientation to naturally occurring social stimuli in one's environment [2], which is closely related to other important social communication skills such as joint attention. Unfortunately, children with ASD usually show powerful deficits in this development. While ASOTS is designed for social orientation skill training that can be adapted for various paradigms, in this chapter, we focus on an important social orientation skill, Response to Name (RTN), to demonstrate the usefulness of this novel system. RTN, as the name suggests, is a task that assesses how a child responds when his (since the ratio of individuals with ASD is estimated at 4-5 to 1 in terms of male to female, we consistently utilize male pronouns for individual specific description in this chapter) name is called. A decreased tendency to RTN is one of the most sensitive and specific predictors of whether an infant will later be diagnosed with ASD when he is old enough for a definitive diagnosis (typically 24 months or later) [3]. RTN is also a key measurement of the standard ASD diagnostic assessment such as the Autism Diagnostic Observation Schedule (ADOS) [4]. As a result, RTN intervention during a period when the brain is still highly malleable and prior to the full manifestation of behavioral impairments of the disorder is extremely important. The ultimate goal of the RTN training is to have the child successfully respond to caregivers' attempts to garner attention by calling his/her name from a variety of locations within the learning environment.

Although several machine-assisted interventions have been designed for children with ASD, few technologies have been reported to assess and train the RTN skill. In a RTN training, a child needs to shift his attention and turn his head to respond to the name call. Previous studies illustrated that visual and audio attention attractors are capable of drawing subjects' attention and shifting it from one position to another [5, 6]. Leblanc et al. [7] pointed out that voluntary shifts of attention were usually driven by the goals of the individual, whereas involuntary shifts occurred in response to the characteristics of the stimuli of which the most salient stimuli attracting attention were the exogenous ones. In this study, we designed salient visual and audio name prompts and attention attractors to help children learn RTN skills.

In an autonomous RTN training system where name prompts occur from different directions, it is important to investigate how to detect the response from children autonomously. A recent work by Bidwell et al. [8] used computer vision algorithms to achieve large range head pose estimation and used this information to infer the visual attention of the participant in a human-administrated RTN training. While this work was an important step towards RTN training, it was not an autonomous system that provided autonomous name prompting or an attention guiding mechanism.

In this chapter, we present the ASOTS, which enables computer-based name calling from a wide range of angles around a participant by providing a distributed display mechanism, allows real-time attention inference of the participant through gaze tracking using a distributed array of cameras and offers an adaptive attention guiding mechanism to shape his response. This new system was first tested via a user study with toddlers with ASD to show whether ASOTS has potential for helping a young child with ASD learn RTN skills. Another user study with typically developing (TD) infants tested whether ASOTS could elicit and access the RTN behaviors of an infant before he is old enough for ASD diagnosis. Early concepts of ASOTS was presented in [9]. The current chapter significantly expands our preliminary work in terms of details of technical development, system validation results, comprehensive user study analysis, and a thorough discussion.

The remainder of this chapter is organized as follows. Section 4.3 presents the design and development of the ASOTS. Section 4.4 describes the validation of ASOTS. The user study design and the results are discussed in Sections 4.5 and 4.6, respectively. Finally, Section 4.7 summarizes the contributions of the chapter and highlights future research directions.

4.3 System Development

4.3.1 Task Definition

We created a system as shown in Fig. IV.1(a), where a child sat on a chair, and was surrounded by a few computer monitors in different locations. A video of a person could be displayed from any of these monitors who would call the child's name. We named this monitor as the target. If the child looked at the target within a limited time, a reward would be given. Otherwise, an attention attractor would be shown to catch and guide the child's attention towards the target. To detect the attention of the child, an array of cameras were used for real-time gaze tracking. In Fig. IV.1(a), the target monitor was Monitor N. However, the child initially looked at Monitor 2. Thus an attention attractor was activated to shift the child's attention from Monitor 2 to Monitor N.

A task is formally defined as a 4-tuple (P , A , Ψ , Φ), where:

P : The name prompt, which represents the name calling displayed on one of the monitors of the set $\{$ Monitor 1, Monitor 2, …, Monitor N$\}$.

A : The attention attractor, which guides the attention of the participant towards the target monitor. This is a continuous process activated by discrete events generated in Φ .

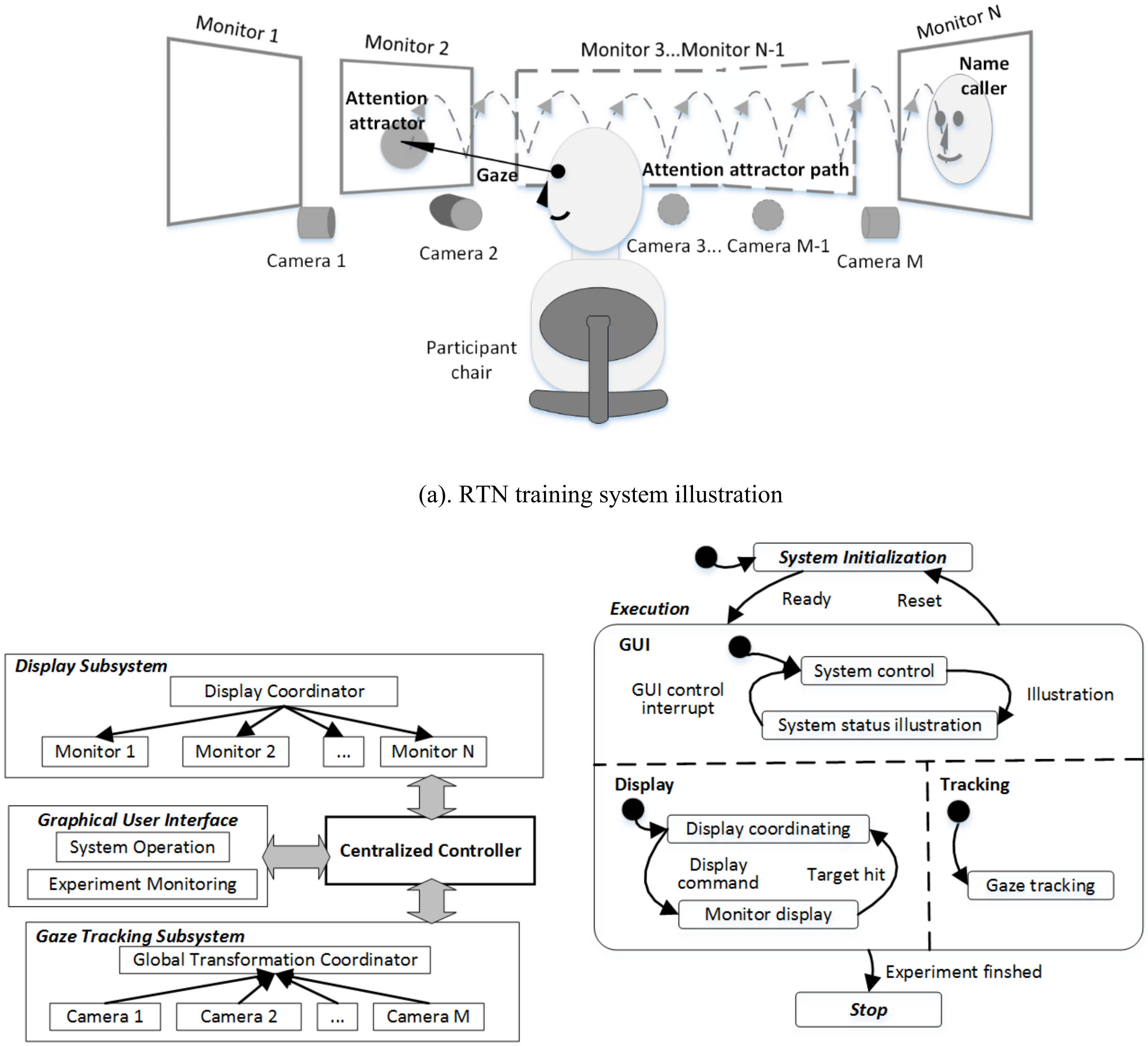
Ψ : Tracking the participant's gaze using Camera 1 to Camera M.

Φ : The autonomous closed-loop RTN interaction protocol, which coordinates P , A , and Ψ .

Let us now introduce the ASOTS system architecture as an overall view of the whole design.

4.3.2 ASOTS Architecture

As shown in Fig. IV.1(b), the main components of the ASOTS architecture are: 1) The display subsystem (implemented by P and A in Section II. C); 2) The gaze tracking subsystem (implemented by Ψ in Section II. D); 3) The centralized controller (CC), which managed the RTN interaction protocol (implemented by Φ in Section III. C); and 4) A Graphical User Interface (GUI).



(a). RTN training system illustration

(b). ASOTS components

(c). ASOTS Statechart model

Fig. IV.1 RTN training system—ASOTS

In order to realize a smooth real-time interaction, the ASOTS was modeled and built as a concurrent system. As shown in Fig. IV.1(c), its execution was modeled using the Harel Statechart model [10], which is an extended state machine capable of modeling hierarchical and concurrent system states. In a Harel Statechart, rounded rectangles denotes system states, *S*. *E* represents a set of events that trigger the system state transitions. When an event happens, a state transition takes place indicated by a directed arrow. The solid rectangle marks exclusive-or (XOR) states, and the dotted line marks AND states. Encapsulation represents the states hierarchy. In the same hierarchy (encapsulated by the same rectangle), the system can only be in one XOR state, while it must be in all of its AND states. In Fig. 1(c), the large rectangle *Execution* encapsulates 3 smaller dotted rectangles *GUI*, *Display*, and *Tracking*. Therefore, when the system was in the *Execution* state, it concurrently ran the *GUI*, *Display*, and *Tracking* sub-states. The *GUI* rectangle contains 2 solid rectangles, *System control* and *System status illustration*. As a result, when the system was in the state *GUI*, it must be in one of its sub-states, *System control* or *System status illustration*.

The lowest level XOR states *S1* and events *E1* are:

$$S1 = \{\textit{System Initialization, Execution, Stop}\} \tag{VI-1}$$

$$E1 = \{\textit{Ready, Reset, Experiment finished}\} \tag{VI-2}$$

In the *System Initialization* state, the hardware and software were initialized and the system component communication were set up. Then the event *Ready* was generated to transfer the state of the system to *Execution*, where an experimental protocol was run. At any time during the execution, if a *Reset* was needed (such as the participant needed to restart), the system was reinitialized. After the protocol was completed, an *Experiment finished* event was generated to stop the system. During the execution, the centralized controller ran in the background to control the interaction logics, generated state transition events, controlled the communication between different system components, and logged the data.

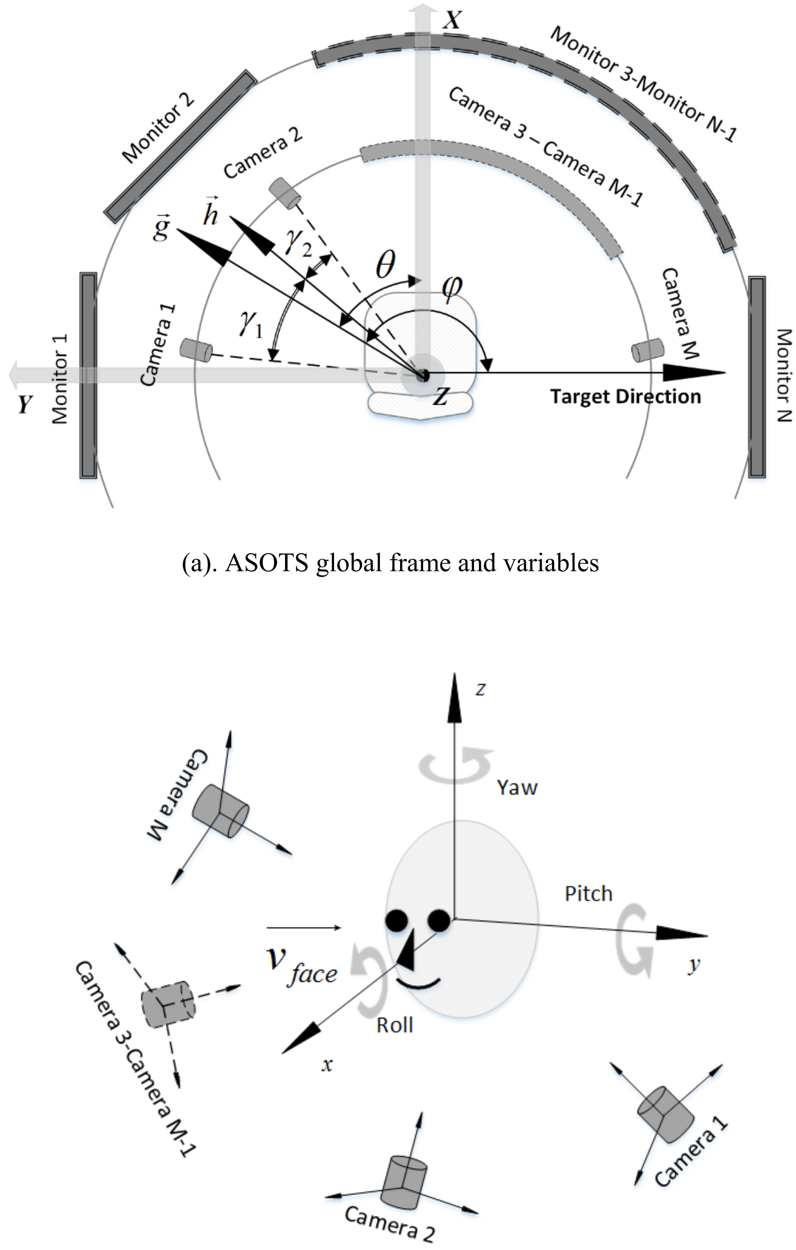
The *S1* state *Execution* encapsulates 3 AND sub-states in set *S2* that described the three main concurrent processes of the system. Each of the AND states also contained its own XOR sub-states in the next hierarchy.

$$S2 = \{\textit{GUI, Display, Tracking}\} \tag{VI-3}$$

The events *E2* that triggered the state transitions were

$$\text{E2} = \{\textit{GUI control interrupt, Illustration, Display command, Target hit}\}\,. \tag{VI-4}$$

The default sub-state of the *GUI* state was *system control*, where the buttons for starting the RTN interaction, pausing the execution, and reset were shown. Once the experimenter pressed the start button, the event *Illustration* was generated to make the GUI show the system status that included the real-time gaze direction, target location, and attention attractor location. If the pause button had been clicked, event *GUI control interrupt* was generated to suspend the system until the start button was clicked again to retrieve. If the reset button was clicked, the system would leave the *Execution* state and go back to the *System Initialization* state.



(a). ASOTS global frame and variables

(b). Different cameras viewing the head of a participant

Fig. IV.2 ASOTS variables, global and local frames

The default sub-state in *Display* was *Display coordinating*, where the display coordinator computed the target monitor index, the label of video or audio that was going to be shown, and the trajectory and effect of the attention attractor. Then, a *Display command* event was generated to trigger the display on monitors accordingly in the *Monitor Display* sub-state. The *Tracking* state only had a default transition to its unique sub-state *Gaze tracking*. When the participant looked at the target monitor, a *Target hit* event was generated, and thus the system went back to *Display coordinating* state to re-compute the parameters based on the experimental protocol.

Fig. IV.2(a) shows the details of the display subsystem and the gaze tracking subsystem as discussed in the following sections. The global frame of ASOTS was a Cartesian coordinate system, with the X-axis pointed forward, the Y-axis pointed to the left, and the Z-axis pointed upwards. The monitors and the cameras were placed on two concentric arcs. The center of the arcs was the origin of the global frame, which was the participant's head position (head frame origin in Fig. IV.2(b)) when he was seated.

4.3.3 Display Subsystem

As shown in Fig. IV.1(b) and Fig. IV.2(a), the display subsystem consisted of N monitors to cover a wide range around the *Z*-axis ($N= 4$, and each monitor was 70cm × 43cm in size in the user studies). The display coordinator (DC) worked as a server and controlled the N monitors as clients using asynchronous socket communication. The monitor clients were embedded with a library of video/audio clips and attention attractor animations. The display subsystem was developed in C# using the Unity Game Engine [11] and a 5.1 surround sound system. Each monitor had a speaker to create sound localization consistent with the video display. Based on an RTN interaction protocol, the CC sent information to the DC (Fig. IV.1(b)), such as the interaction stages, trial numbers, and prompt levels. The DC then computed when and which video/audio clip to display and the effect and trajectory of the attention attractor. These pieces of information were sent to the monitor clients for appropriate display.

Name prompts

The name prompts were implemented as P in the 4-tuple described in Section 4.3.1. An experienced therapist recorded name calling video/audio clips with a soft neutral tone for each participant. When needed, a video/audio clip was displayed on the target monitor for the participant to look at.

Attention Attractor

The attention attractor was implemented as A in the 4-tuple described in Section 4.3.1. When the participant did not attend to a name call, the attention attractor was introduced to help the participant shift attention towards the target monitor. It was a red ball embedded with a bouncing sound that bounced from the current attention location of the participant to the target monitor through the intermediate monitors.

The attention attractor bounced in a periodic parabolic path, which approximated a natural bouncing motion of a ball. In Fig. IV.2(a), $\bar{g}$denotes the projection of the participant's gaze direction on the *XY*-

plane when the attractor was started. θ is the angle between $\bar{g}$ and X-axis, and $\varphi$is the angle between $\bar{g}$ and the target direction (center of the target monitor). $\omega(t)$ is the angular velocity of the attractor with respect to Z-axis. h, α, and c are parameters that adjust the shape of the parabola. T is the period of the repeated parabolic path, and r is the radius of the monitor arc. The trajectory of the attention attractor was:

$$\begin{cases} x(t) = \cos(\theta + \omega(t)) \\ y(t) = \sin(\theta + \omega(t)) & , 0 \ \leq t \leq \ |\varphi / \omega(t)|. \\ z(t) = h - \alpha(r\omega(t \bmod T) - c)^2 \end{cases} \tag{VI-5}$$

We set $\omega(t) = \pm 20^{\circ}/\mathrm{s}$ (+: counter-clockwise; -: clockwise), $h = 30\,\mathrm{cm}$, $r = 110\,\mathrm{cm}$, $c = 385\,\mathrm{cm}$, $\alpha = 1.35^{-4}$ cm^{-1}. Intuitively, the attractor bounced between 10cm to 30cm in height on the monitors.

4.3.4 Gaze Tracking Subsystem

The gaze tracking subsystem was implemented by Ψ in the 4-tuple described in Section 4.3.1. The gaze tracking algorithm was implemented with MATLAB and OpenCV. In the user study, we applied 4 Logitech C930e webcams with a resolution of 720p. The head frame was a Cartesian coordinate system as shown in Fig. IV.2(b). By modeling the head as an ellipsoid, the origin of the head frame was set as the center of this ellipsoid. The origin of the head frame coincided with the origin of the global frame when a participant was seated. The $x$direction pointed from the frontal face, and the unit vector along the positive x-axis, $\overrightarrow{v_{face}}$, indicated the frontal face orientation. In what follows, we discuss the steps required to compute gaze direction $\overrightarrow{v_{gaze}}$ from $\overrightarrow{v_{face}}$.

Step 1. Estimate the head orientation from the cameras. The head orientation was represented by 3 Euler angles (roll, pitch, yaw) as shown in Fig. IV.2(b). The head orientation estimation can be treated as solving a nonlinear least squares problem:

$$\min_{e} \|P(e, M) - I\|_2^2 , \tag{VI-6}$$

where $M \in \Re^{3 \times n}$ is a 3D face model. $I \in \Re^{2 \times n}$ is the projection of Mon an image. $e$is a vector of head orientation Euler angles viewed in the camera frame. P is the projection function which projects Mon the image given e resulting in I. Here we applied the inverse supervised descent method (SDM) [12, 13] to solve this problem. While there are other methods [14, 15] on head orientation estimation from a single camera, we chose SDM due to its robustness, high precision and real-time computational ability.

A software tool IntraFace [12, 13], which implemented SDM, was used in our system. Based on the 3 Euler angles estimated, $\overrightarrow{v_{face}}$ can be computed in the camera frame, denoted as $\overrightarrow{v_{face}^{camera}}$.

In our application, a participant was expected to turn his head by a large yaw angle to respond to name prompts, and thus might exceed the range of a single camera to capture the frontal face image. Therefore, multiple cameras were combined to extend the detection range. By arranging the cameras along a circular arc (Fig. IV.2(a)), a participant's frontal face could be captured by at least one of the cameras when he faced to any part of the display subsystem. Similar strategies have been used by a few other studies in different contexts [16-18]. We attached each camera with its own reference frame (Cartesian coordinate system) as illustrated in Fig. IV.2(b). The optical axes of the cameras were calibrated to intersect at the origin of the global frame. All the cameras ran in parallel. In this way, each camera could see the participant's head in the center of its view concurrently.

In Fig. IV.2(a), $\vec{h}$denotes the projection of $\overrightarrow{v_{face}}$ on *XY*-plane, and γ_i represents the angle between the optical axis of Camera i and $\vec{h}$. In our preliminary testing with children aged 1-2 years, we found that the SDM head orientation estimation was reliable when $\gamma \leq 40°$ under our experiment room illumination. Smallerγ resulted more accurate estimation. In some cased the head orientation could be estimated by more than one camera. For example, in Fig. IV.2(a), if both γ_1 and γ_2 were smaller than $40°$, then both Camera 1 and Camera 2 could estimate the head orientation. In this case, $\overrightarrow{v_{face,i}^{camera}}$ computed from a camera with the smaller γ_i was likely to produce the better estimation.

Step 2. Transform the head orientation from a camera's frame to the global frame. Based on the geometric distribution of the system discussed in Section 4.3.2 and 4.3.3, the transformation matrix R_{camera}^{global} from each camera frame to the global frame was precomputed by camera calibration. If Camera i was chosen in step 1, then its transformation matrix, $R_{camera,i}^{global}$, was used to computed $\overrightarrow{v_{face}} = R_{camera,i}^{global} \overrightarrow{v_{face,i}^{camera}}$.

Step 3. Compute $\overrightarrow{v_{gaze}}$ from $\overrightarrow{v_{face}}$. We conducted a small experiment where three adults were seated in the participant chair to look at a marker (indicated ground truth value of $\overrightarrow{v_{gaze}}$) moving back and forth between their left side (θ= 107°) and right side (θ= -107°). At the same time, the values of $\overrightarrow{v_{face}}$ computed by the gaze tracking subsystem were recorded and compared with the values of $\overrightarrow{v_{gaze}}$. 3000 data pairs were collected and the comparison showed that while y and z components were approximately the same, the x component of $\overrightarrow{v_{face}}$ needed to be amplified by 120% to obtain the x component of $\overrightarrow{v_{gaze}}$ (gaze in horizontal), i.e., $\overrightarrow{v_{gaze}} = (1.2x, y, z)_{\overrightarrow{v_{face}}}$.

4.4 System Validation

4.4.1 Experimental system setup

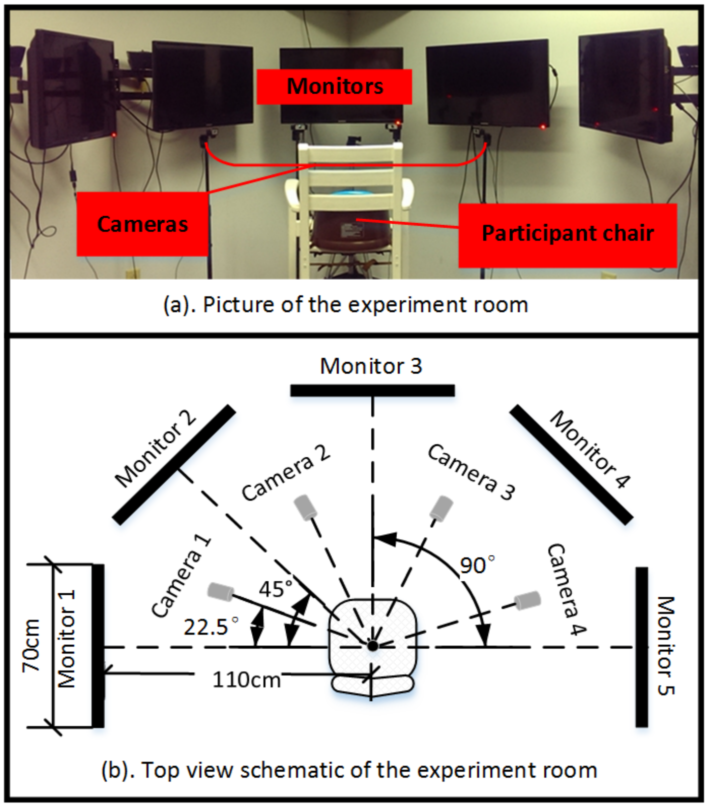


Fig. IV.3 Experimental system setup

The experiment room was arranged as shown in Fig. VI.3, where (a) shows the picture of the room, and (b) is the top view. The five monitors were 110 cm away from the origin of the global frame. The angle between the two adjacent cameras' optical axes was 45°, and the angle between Camera 1 and the positive Y-axis was 22.5°. The bottom of each monitor was at a height of 120 cm from the floor, which was also the height of the cameras. For each monitor, there was a range of gaze yaw angles (θ, as marked on Fig. 2(c)) that indicated the participant was looking at that monitor. The ranges of θ for Monitor 1 to Monitor 5 were [107°, 73°], [62°, 28°], [17°, -17°], [-28°, -62°], and [-73°, -107°], respectively. Based on our experience, when the participant looked at the monitors, their gaze pitch angle (approximated by head pose pitch angle) was in the range of [-30°, 21°]. Therefore, even when the participant's gaze yaw angle was within the range of a monitor but his gaze pitch angle was out of this threshold, this was not considered looking at that monitor. This system setup was validated for its precision and real-time execution. We used the same setup to conduct two pilot user studies that are discussed in section 4.5 and section 4.6.

4.4.2 Head Orientation Estimation

Since the gaze detection was approximated from head pose estimation, we recruited three adults and one 18 month-old child to test the accuracy of the head orientation estimation. The InertiaCube4 by InterSense [19] was mounted on top of the participant's head to provide the ground truth of the head orientation with respect to the global frame in roll, pitch, and yaw Euler angles. The IneriaCube4 offers full 360° angular range with the accuracy of 1° in yaw, and 0.25° in pitch and roll angles. The participant was seated in the chair as shown in Fig. IV.3. At the start of the testing, the results from the InertiaCube4 and the camera array were calibrated as 0. During the test, the readings of the InertiaCube4 and the camera array were synchronized and recorded as time sequences $IC(n)$ and $CA(n)$, respectively. The accuracy of the head orientation estimation by using the camera array was defined as

$$\sum_{n=1}^{n=N} \left| CA(n) - IC(n) \right| \Big/ N \text{ (} N = \text{the length of } CA\text{).} \tag{VI-7}$$

During the test, the adult participants were asked to perform 3 types of head rotation: i) free head rotation as fast as they could; ii) slow head rotation such as that might occur while viewing pictures in a museum; and iii) normal head rotation that might happen when looking at a monitor and then switching to another randomly similar to what the participants would do in the user studies. 6581 data points were collected from the adult participants and the average accuracy was 6.47°, 3.42°, 4.53° in yaw, pitch, and roll angles, respectively. The child participant was too young to perform these types of head rotation, and therefore his head rotation was stimulated by displaying videos randomly on the five monitors. There were 1101 data points recorded from the child, and the accuracy computed was 6.74°, 4.68°, 2.68° in yaw, pitch, and roll angles, respectively. These results were acceptable for the user studies since we were interested in locating a participant's gaze on the large computer monitors without needing to isolate a precise point location. A typical data plot is shown in Fig. IV.4.

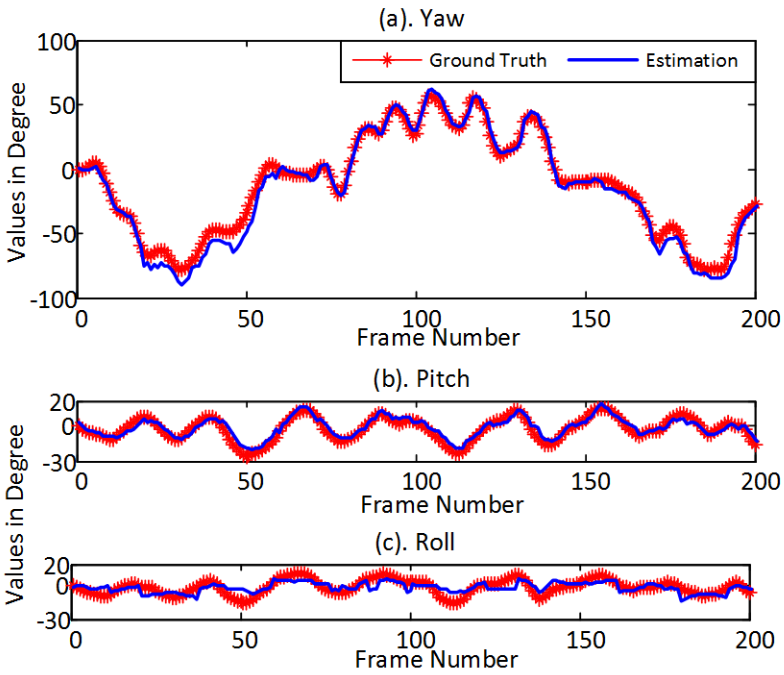


Fig. IV.4 Head orientation estimation validation result example

4.4.3 *Target Hit Recognition*

Given the head orientation estimation, the gaze direction was approximated as stated in Section 4.3.3 in Steps 2 and 3. The accuracy of this approximation was validated by 3 adults and 7 TD children aged 1-2 years. They were guided to look at a video or an image displayed on each of the monitors randomly for 10 times. 95% target hits were correctly recognized, 4% of them were false negative (i.e., the participants looked at the target monitor but the system did not recognize it) and 1% was false positive (i.e., the participant did not look at the target monitor but the system recognized this as looking at the correct monitor). While 95% accuracy was deemed sufficient for our tasks, we further investigated the cause for this 5% error. We found that this was mainly due to activities that occluded part of the participant's face such as finger sucking and drinking with a sippy cup.

4.4.4 *Real-time Execution*

A high communication speed between the gaze tracing subsystem and the display subsystem was essential to guarantee that the system responded to the participant in real-time. In order to test the communication speed, a sequence of signals were sent from the gaze tracking subsystem through the CC to the display subsystem, and vice versa. Results showed that the signal transmission between these two subsystems took 25 ms on average. The gaze tracking subsystem was refreshed at a rate of 15 fps (67 ms/frame), and the display subsystem ran at a rate of 50 fps (20ms/frame). When there was no communication needed between the two subsystems, they run in parallel independently, which was the most common case. The most time consuming scenario was when the gaze direction detected by the gaze tracking subsystem had to be sent to the display subsystem through CC to generate a visual/audio stimulus. In this case, it took about 112 milliseconds on average. However, this procedure was only needed a few times (e.g., response to a target hit event and initialization of the attention attractor's starting position) per trial. Since human visual system takes about 150ms to process a familiar object and scene [20], and takes about 330ms for gaze fixation and saccade in a scene perception [21], ASOTS was fast enough for real-time interaction.

4.5 Experimental Setup

4.5.1 Purposes and Participants

Two pilot user studies were conducted to validate the impact of ASOTS. Note that the user studies were not designed as formal psychological/clinical empirical experiments, but only for validating young children's tolerance of and response to ASOTS. Participants were recruited from a research registry of the Vanderbilt Kennedy Center, and this study was approved created by the Vanderbilt University Institutional Review Board.

The first user study tested whether ASOTS could engage young children with ASD and elicit RTN behaviors from them. This user study included 10 toddlers with ASD (Age: Avg = 2.29; SD = 0.32 years). They had confirmed diagnoses given by a clinician based on DSM-IV-TR [22] criteria. They met the spectrum cut-off on the Autism Diagnostic Observation Schedule [4] (Raw score: Avg = 22.40; SD = 3.24. Imagination/Creativity score: Avg = 2.80; SD = 0.35.), and had existing Intelligence Quotient data (Avg = 55.60; SD = 11.50) regarding cognitive abilities in the registry. Parents of toddlers with ASD also completed the Social Responsiveness Scale–Second Edition [23] (Raw score: Avg = 80.50; SD = 25.73. T score: Avg = 66.70; SD = 10.01) to index current ASD symptoms.

Although usually a definitive diagnosis of ASD cannot be made before 24 months [3], RTN is also important for predicting potential risks of ASD for younger infants [2]. Therefore, we conducted another user study with 10 TD infants (Age: Avg = 1.35; SD = 0.40 years.). These children tested whether ASOTS can be accepted by infants, and whether this system could elicit RTN behaviors from them. Thus this user study suggested whether ASOSTS has the potential to be used for early screening and training for infants.

Note that the TD infants were not the control group for the toddlers with ASD, since the two user studies were designed for different purposes and the two groups of children were not in the same age range.

4.5.2 Task and Protocol

The experimental protocol was the implementation of Φ in the 4-tuple described in Section 4.3.1. Each participant took part in two sessions, one video-based session and one audio-based session. In the video-based session, a prerecorded video clip showing a therapist calling the participant's name was displayed on one of the monitors, which simulated the condition where the caller was both visible and audible. In our initial proof of concept study [9], we found that this was an easy setup to quickly get children involved in the interaction, and generate good RTN performance. Thus we further designed the audio-based session, where visual display was omitted and only the audio portion extracted from the name calling video was emitted from one of the monitors. The audio-based session simulated scenarios such as the parents calling a child from another room, or playing hide-and-seek games. By comparing the results between the audio- and the video-based sessions, we were able to assess how much of a difference eliminating visual information made in RTN performance.

At the beginning of each session, a prerecorded welcome video from the therapist was displayed on Monitor 3. Then a Sesame Street video clip was displayed on each monitor for 8.6 seconds in the following order: Monitor 3, 4, 2, 5, 1, 5, 2, 4, and 3, respectively, for a total of 77 seconds to help the participant get familiarized with the environment.

For both the video- and audio-based sessions, participants completed 10 RTN trials. From trial 1 to trial 5, the target monitor was Monitor 5, 1, 4, 2, and 3, respectively. With this setup, the participant needed to turn his head from a target monitor on one side to another target monitor on the other side three times, and then went back to the central monitor. In each trial, a participant's name was called from the target monitor repeatedly. Each name call lasted approximately 2 seconds. A 4 level name call prompt hierarchy is shown in Table IV.1. If the participant did not look at the target monitor in one prompt level, a higher level of prompt would be presented.

Prompt 1 was the baseline prompt of 2 name calls from the target monitor. For Prompt 2, the attention attractor was activated on the monitor closest to the participant's gaze direction. This attractor then moved across all monitors and towards the target monitor in an attempt to guide the participant's gaze while, at the same time, the name call prompt was repeated. In Prompt 3, the ball first bounced in the gaze direction for 2 seconds to help the participant notice it, and then bounced toward the target monitor. Both Prompt 2 and Prompt 3 were enhanced with additional audio that resembled a Tennis ball hitting a wooden floor (normal bouncing sound). In Prompt 4, the motion of the ball was the same as that in Prompt 3, but the sound of the ball was enhanced to resemble a rubber ball hitting a gong (special bouncing sound). At any time during the prompts, once the participant looked at the target monitor, the prompting was terminated and a reward video clip was displayed on the target monitor where the therapist praised the participant (pre-recorded) by saying "Good Job! You found me!" followed by a firework animation.

Table IV.1 Prompt Levels

Prompt level	Prompting element list
Prompt 1	1
Prompt 2	Repeated 1 in parallel with 3 and 4
Prompt 3	Repeated 1in parallel with 2, 3, and 4
Prompt 4	Repeated 1 in parallel with 2, 3, and 5
Prompt Element: 1. Name calling video/audio displayed on the target monitor; 2. Attractor bouncing in the gaze direction for 2 seconds; 3. Attractor bouncing from the gaze direction to the target monitor; 4. Normal bouncing sound; 5. Special bouncing sound.	

After the first 5 RTN trials, another Sesame Street video clip of 76 seconds was displayed to give participants a short break. This clip was displayed in a similar way as the first "fun" video clip, except that the display on each monitor lasted for about 8.4 seconds and the display switching followed the order of Monitor 3, 2, 4, 1, 5, 1, 4, 2, and 3, respectively. Another 5 RTN trials followed the break. The target monitor for trial 6 to trial 10 was Monitor 1, 5, 2, 4, and 3, respectively. This reversed display order of the 2 funny video clips and the two sets of RTN trials helped reduce the chance of participant habituation. The very last part of the interaction was a "Good-bye" video displayed on Monitor 3. All of the videos except the Sesame Street video clips were recorded with the same therapist to provide a homogeneous and comparable environment for all the participants across all sessions and trials.

4.5.3 Measurements

For each group, we compared the participants' attention and performance in the video- and audio-based sessions to investigate the differences produced by increasing the task difficulty. Since we used a large scale interaction environment, we calculated the attention and performance including the whole environment in a global evaluation. We then computed the attention and performance associated with each monitor to determine whether the direction of a target influenced the interaction.

First, we evaluated the participant's attention towards the interaction environment, which reflected their engagement. We hypothesized that the more time they spent looking at the display region, the higher their engagement was during the interaction, which was related to the RTN performance and important for eventual learning and success within the interaction paradigms. Since Sesame Street videos are popular for young children in general, we can use their gaze on the display region during the "fun video" display as a baseline to assess their engagement in the RTN. We calculated the duration for which they were looking at the display region, which included the 5 monitors and the gaps between the monitors. We also calculated the duration of time spent on looking at each monitor. We hypothesized that the larger the target yaw angle was from the participants' frontal head orientation, the less attention the participant would pay to this target.

Second, we evaluated the participants' RTN performance, which reflected whether the interaction protocol was within an appropriate difficulty range to elicit the RTN behavior of young children. We computed: i) the number of trials where the participants hit the target eventually; ii) in each trial, the prompt levels they needed and how much time they spent trying to hit the target; and iii) the distribution of ii) on each monitor. Similar measurements have been widely used in psychological studies regarding RTN skills [24]. In general, we anticipated that the participant would need lower level of prompts to hit a target in the video-based session, since the name caller was visible. In other words, participants in the audio-based session would need more help from the attention attractor.

4.6 Experimental Results

Since the toddlers with ASD and the TD infants were recruited for different purposes in two separate user studies as discussed in section 4.5.1, the results of each group are analyzed separately as follows.

4.6.1 Results of Toddlers with ASD

Eleven toddlers with ASD were recruited initially, and 10 of them completed both video- and audio-based sessions. One child did not participate in the audio-based session. Fig. IV.5(a) shows the percentage of time during the whole session that the toddlers with ASD spent looking at the display region, for both video- and audio-based sessions. Fig. IV.5 (b) shows the average duration the toddlers with ASD spent looking at each monitor (M1 to M5 represent Monitor 1 to Monitor 5). The "Fun video" indicates the Sesame Street video clips display periods, and the "RTN" represents the 10 RTN trials. We can see that in both cases the toddlers with ASD looked at the display region for most of the sessions (>82.77%). Their attention towards the display region was even higher in the RTN trials in both the video- and audio-based sessions. This indicated their interest towards the RTN interaction. In general, the toddlers with ASD looked at the frontal monitor the most, and the farthest side monitors (Monitor 1 and Monitor 5) the least.

This result was consistent with our hypothesis that the farther the monitor was angled from 0° (frontal direction of the participant), the less attention was paid to that monitor.

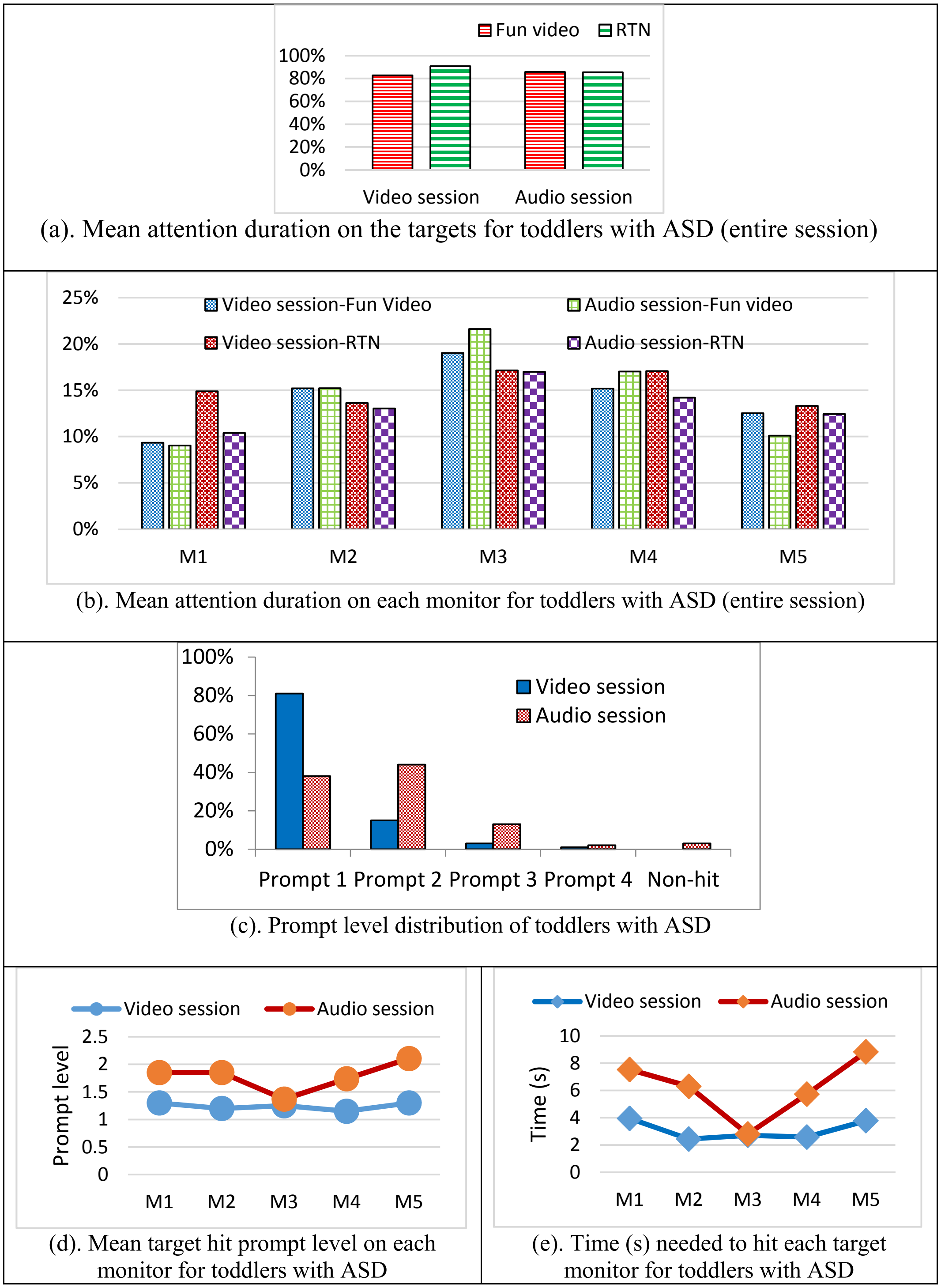
Fig. IV.5(c) presents the prompt level distribution of the toddlers with ASD. 100% and 97% trials ended up with a target hit in the video- and audio-based sessions, respectively. In the video-based session, the toddlers hit the target on Prompt 1 (no attention attractor) for 81% of trials, and needed the attention attractor (Prompt 2 to Prompt 4) for 19% of trials. In the audio-based session, they hit the target on Prompt 1 for 38% of trials, and on Prompt 2 to Prompt 4 for 59% of trials. On average, the toddlers with ASD hit the target at prompt level 1.24 (SD = 0.55) and 1.78 (SD = 0.75) in the video- and audio-based sessions, respectively. Fig. IV.5(d) shows the participants' average performance on each monitor when it was the target. We found that there was no apparent change on the average target hit prompt levels across different monitors in the video-based session. However, in the audio-based sessions, the farther away a target was, the worse the RTN performance was, which was consistent with the attention preference of toddlers with ASD on different monitors.

The toddlers with ASD needed, on average, 3.09 (SD = 3.23) seconds and 6.25 (SD = 4.15) seconds to hit the target in the video- and audio-based sessions, respectively. Since each name call lasted for about 2 seconds, the result showed that toddlers with ASD turned to their names at the second and third name calls in the video- and audio-based sessions, respectively. Similar to Fig. IV.5(d), Fig. IV.5(e) shows the average time that these participants needed to hit a target on each monitor. We can see that the pattern on Fig. IV.5(e) was consistent with that of Fig. IV.5(d). This was expected since the longer a participant needed to hit a target, the higher the target prompt level would be.

In summary, these results showed that the system was well tolerated by toddlers with ASD, and successfully elicited RTN behaviors from them. The attention attractor was helpful in both sessions, especially in the audio-based session with a more difficult RTN task. Thus, we believe that ASOTS has a great potential to be used for teaching RTN skills to young children who are diagnosed with ASD.

4.6.2 Results of TD infants

All 10 TD infants recruited completed the study. Fig. IV.6(a) shows that although they were only 16 months old on average, they still paid considerable attention on the monitors in both fun video display periods and RTN trials (>88.50%). In general, the TD infants spent comparable time on looking at the display region in both the RTN trials and the "fun video" display periods in video- and audio-based sessions. This suggested that the RTN trials were as attractive as the Sesame Street video clips to the TD infants. Fig. IV.6 (b) shows the average duration the TD infants spent on looking at each monitor. They looked at the frontal monitor (Monitor 3) mostly, and looked at the side monitors (Monitor 1, 2, 4 and Monitor 5) relatively evenly.



(a). Mean attention duration on the targets for toddlers with ASD (entire session)

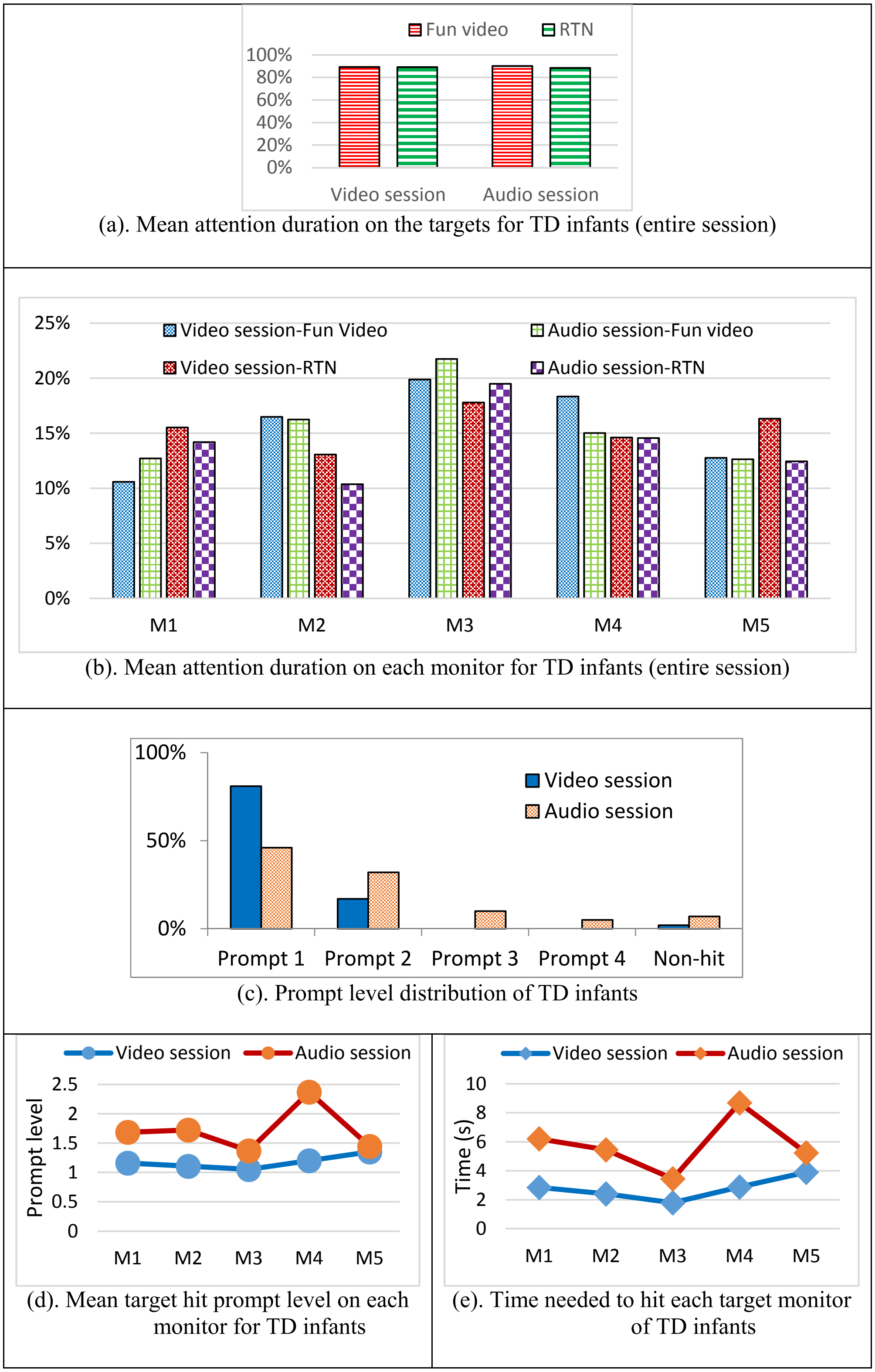
(b). Mean attention duration on each monitor for toddlers with ASD (entire session)

(c). Prompt level distribution of toddlers with ASD

(d). Mean target hit prompt level on each monitor for toddlers with ASD

(e). Time (s) needed to hit each target monitor for toddlers with ASD

Fig. IV.5 Pilot user study results of toddlers with ASD



(a). Mean attention duration on the targets for TD infants (entire session)

(b). Mean attention duration on each monitor for TD infants (entire session)

(c). Prompt level distribution of TD infants

(d). Mean target hit prompt level on each monitor for TD infants

(e). Time needed to hit each target monitor of TD infants

Fig. IV.6 Pilot user study results of TD infants

TD infants hit the target in 98% and 93% trials in the video- and audio-based sessions, respectively. Fig. IV.6(c) presents the prompt level distribution. In the video-based session, they hit the target on Prompt 1 for 81.00% of trials, and on Prompt 2 to Prompt 4 for 17% of trials. In the audio-based session, they hit the target on Prompt 1 for 46% of trials, and on Prompt 2 to Prompt 4 for 47% of trials. On average, the TD infants hit the target at prompt level 1.17 (SD = 0.38) and 1.72 (SD = 0.86), as well as required 2.77 (SD = 2.28) seconds and 5.81 (SD = 4.97) seconds before the target hit in the video-and audio-based sessions, respectively. Fig. IV.6(d) shows the TD infants' average performance associated with each target monitor. We found that the lowest target hit prompt level happened on Monitor 2, and the highest one happened on Monitor 5 in the video-based session. In the audio-base session, the lowest target hit prompt level happened on Monitor 5, and the highest one happened on Monitor 4. In general, in both sessions, the frontal monitors were still slightly easier to hit than the side monitors. The values in Fig. IV.6(d) are consistent with the duration these TD infants needed to hit a target, as shown in Fig. IV6(e).

In summary, ASOTS was well tolerated by these TD infants and elicited their RTN behaviors successfully. Therefore, in the future, ASOTS has a potential to be upgraded to conduct early screening for at risk infants (e.g., siblings of children with ASD) who are too young to be diagnosed with ASD.

4.6.3 Effect of the attention attractor

Fig. IV.7 shows a typical trial in the experiment, where the gaze of a participant was guided by the attention attractor. In this trial, the target monitor was Monitor 5. Fig. IV.7 (a) shows the path of the participant's gaze and the attractor. The horizontal axis represents the distance along the monitors, with the origin at the center of Monitor 3. M1 to M5 mark the regions of Monitor 1 to Monitor 5. The participant's gaze was at first on M2, and then shifted to M3. When the name call was finished, the participant's gaze was around the upper edge of M3, therefore, the ball showed up in M3 and bounced towards M5. The participant's gaze followed the attractor and was guided to M5 eventually. Fig. IV.7 (b) and (c) show the trajectory of the participant's gaze in horizontal and vertical directions, respectively. We can see that since second 4 (end of Prompt 1), the gaze of the participant was shifted along with the position of the attractor. At around second 9, the attractor and the gaze reached the center of M5, which meant the participant hit the target. Then the attention attractor disappeared and a reward was displayed on M5. The participant's attention was on M5 until the end of the reward (also the end of the trial).

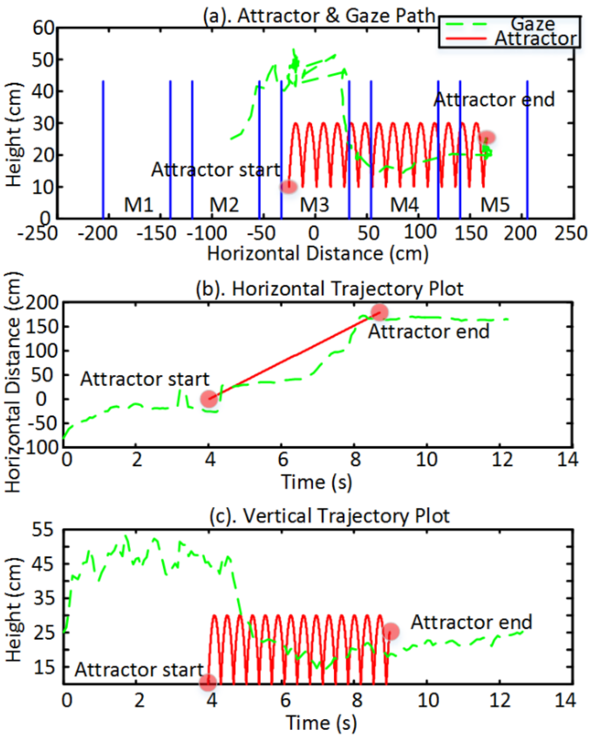


Fig. IV.7 Demonstration of attention attractor

4.7 Discussion and Conclusion

In this chapter, we have presented an autonomous system, ASOTS, to help young children with ASD learn social orientation skills. Our selected response to name (RTN) skill was targeted in that it is seen as an early red-flag of ASD. As such RTN provides a specific example to demonstrate the potential effect of the proposed system and the interaction protocol. ASOTS consisted of a distributed display subsystem which provided a wide range of name prompts. Accordingly, a distributed gaze tracking subsystem was designed to monitor the response of the participant. ASOTS is an adaptive closed-loop autonomous system, where the behaviors of the system adapts in real-time depending on the performance of the participant. The implementation and the validation of ASOTS system was discussed in detail. The validation results showed that the gaze tracking was accurate, and the system was fast enough for real-time RTN interaction.

An interaction protocol was proposed to assess the functionality of ASOTS. If a participant could not attend to a target monitor within a given time period of a prompt, a higher level of prompt was provided with the aid of an attention attractor. The attention attractor was effective in guiding the participant's attention towards the target. Two pilot user studies were conducted to test the system. Ten toddlers with

ASD tested the feasibility of ASOTS for affected children (ASD). Ten TD infants validated the potential future use of ASOTS in a much younger prodromic sample. The results demonstrated that ASOTS were well tolerated by both groups, and successfully elicited expected RTN behaviors. However, these two pilot user studies were not designed as clinical efficacy studies of ASOTS, which is beyond the scope of the current work, and thus the results of this current work should be seen as feasibility and tolerability result that indicate promise of ASOTS in future intervention.

In this context, it is important to highlight several limitations of the current study. First, ASOTS needs to be further upgraded to fit in formal clinical empirical studies in the future. The pilot user studies only involved a few participants within a limited time frame of interaction. Therefore, recruiting larger user groups and conducting longitudinal experiments will be needed in the future to answer the ultimate value of ASOTS in formal clinical empirical studies. For the toddlers with ASD, a TD control group would be needed to access the difference between toddler with ASD and their TD peers. For TD infants, a group of at risk infants (e.g., siblings of children with ASD) will be needed to test whether the RTN behaviors detected by ASOTS can contribute to the prediction of ASD in a longitudinal empirical study. While we did assess promising attentional response within system, we did not systematically compare such improvements in other methods nor did we see if such training generalized to other interactions.

Despite these limitations, this work is the first to our knowledge to design and empirically evaluate the usability and feasibility of an autonomous closed-loop social orientation training system capable of modifying prompts based on within system measurements of attention. The preliminary results for RTN presented in the chapter are promising. Note that this system is not limited to RTN protocol alone. The ASOTS architecture and system components can be adapted to address other core deficits in social orientation (e.g., joint attention skills). Importantly, we do not propose this technology as a replacement for existing necessary comprehensive behavioral intervention and care for young children with ASD. Instead, this platform represents a move toward realistic deployment of technology capable of accelerating and priming a child for learning in key areas of deficit.

4.8 References

[1] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, “Early social attention impairments in autism: social orienting, joint attention, and attention to distress,” *Developmental psychology,* vol. 40, no. 2, pp. 271, 2004.

[2] G. Dawson, A. N. Meltzoff, J. Osterling, and J. Rinaldi, “Neuropsychological correlates of early symptoms of autism,” *Child development,* vol. 69, no. 5, pp. 1276-1285, 1998.

[3] C. P. Johnson, and S. M. Myers, “Identification and evaluation of children with autism spectrum disorders,” *Pediatrics,* vol. 120, no. 5, pp. 1183-1215, 2007.

[4] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[5] F. Grani, F. Argelaguet, V. Gouranton, M. Badawi, R. Gaugne, S. Serafin, and A. Lecuyer, "Audio-visual attractors for capturing attention to the screens when walking in CAVE systems." pp. 75-76.

[6] M. Waldner, M. Le Muzic, M. Bernhard, W. Purgathofer, and I. Viola, "Attractive Flicker: Guiding Attention in Dynamic Narrative Visualizations," *IEEE Transactions on Visualization and Computer Graphics,* vol. Vol.20, no. 12, pp. 2456-2465, 2014.

[7] É. Leblanc, D. J. Prime, and P. Jolicoeur, "Tracking the location of visuospatial attention in a contingent capture paradigm," *Journal of Cognitive Neuroscience,* vol. 20, no. 4, pp. 657-671, 2008.

[8] J. Bidwell, I. A. Essa, A. Rozga, and G. D. Abowd, "Measuring Child Visual Attention using Markerless Head Tracking from Color and Depth Sensing Cameras." pp. 447-454.

[9] Z. Zheng, Q. Fu, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Design of a Computer-assisted System for Teaching Attentional Skills to Toddlers with Autism."

[10] D. Harel, "Statecharts: A visual formalism for complex systems," *Science of computer programming,* vol. 8, no. 3, pp. 231-274, 1987.

[11] U. G. Engine, "Unity Game Engine-Official Site," *Online][Cited: October 9, 2008.] http://unity3d. com.*

[12] X. Xiong, and F. De la Torre, "Supervised descent method and its applications to face alignment." pp. 532-539.

[13] X. Xiong, and F. De la Torre, "Supervised Descent Method for Solving Nonlinear Least Squares Problems in Computer Vision," *arXiv preprint arXiv:1405.0601*, 2014.

[14] E. Murphy-Chutorian, and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, no. 4, pp. 607-626, 2009.

[15] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests." pp. 617-624.

[16] S. O. Ba, and J.-M. Odobez, "Probabilistic head pose tracking evaluation in single and multiple camera setups," *Multimodal Technologies for Perception of Humans*, pp. 276-286: Springer, 2008.

[17] Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, "Real time head pose tracking from multiple cameras with a generic model." pp. 25-32.

[18] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Predicting driver maneuvers by learning holistic features." pp. 719-724.

[19] "InertiaCube4," http://www.intersense.com/pages/18/234/.

[20] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *nature,* vol. 381, no. 6582, pp. 520-522, 1996.

[21] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological bulletin,* vol. 124, no. 3, pp. 372, 1998.

[22] *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the Diagnostic Criteria from DSM-IV-TR*, Fourth ed., Washington D.C.: American Psychiatric Association, 2000.

[23] J. N. Constantino, and C. P. Gruber, "The social responsiveness scale," Los Angeles: Western Psychological Services, 2002.

[24] A. S. Nadig, S. Ozonoff, G. S. Young, A. Rozga, M. Sigman, and S. J. Rogers, "A prospective study of response to name in infants at risk for autism," *Archives of pediatrics & adolescent medicine,* vol. 161, no. 4, pp. 378-383, 2007.

Chapter V. Semi-Autonomous Robot-Mediated Joint Attention Intervention

5.1 Abstract

This chapter describes development and application of a novel semi-autonomous adaptive robot-mediated interaction technology for teaching early joint attention skills to children with ASD. The system is composed of a humanoid robot endowed with a prompt decision hierarchy to alter behavior in concert with reinforcing stimuli within an intervention environment to promote joint attention skills. Results of implementation of this system over time, including specific analyses of attentional bias and performance enhancement, with 6 young children with ASD are presented.

Keywords—Robot-mediated intervention system, children with autism, joint attention

5.2 Introduction

This research is motivated by this highlighted potential of robotic technology and designs and tests a potentially transformative co-robotic technological paradigm for future ASD intervention. In particular, it focuses on developing a co-robotic intervention platform and environment specifically designed to accelerate improvements in early joint attention skills [1, 2]. Joint attention skills are thought to be fundamental, or pivotal, social communication building blocks that are central to etiology and treatment of ASD [3, 4]. At a basic level, joint attention refers to the development of specific skills that involve sharing attention with others (e.g., pointing, showing objects, and coordinating gaze). These exchanges enable young children to socially coordinate their attention with other people to more effectively learn from their environments. Fundamental differences in early joint attention skills have been demonstrated to underlie the deleterious neurodevelopmental cascade of the disorder and successful treatment of these deficits has been demonstrated to substantially improve numerous developmental skills across settings [1, 2, 4].

The present work is built upon a previous work [5, 6] where the authors developed and piloted a robot-mediated autism intervention architecture called ARIA (Adaptive Robot-mediated Intervention Architecture) and demonstrated three significant findings (refer II. B for details): i) children with ASD demonstrated an attentional bias toward the robot as opposed to a human therapist; ii) it was possible to develop a closed-loop autonomous robot-mediated joint-attention intervention system that could dynamically adapt interaction based on the performance of the child; and iii) this system performed as well as a therapist on a small sample of children with ASD over a very limited time course (1 session). In this present work, we expand upon our previous work to test two important questions: i) whether repeated interactions with the robotic system would impact performance regarding early joint attention skills and ii) whether the initial attentional bias and preference to the robot would hold over time. These questions are extremely important to test the ultimate value of robotic interactions in children with ASD, as if the initial attentional bias quickly habituates or if repeated exposure is unable to utilize initial attentional preferences to promote skills robotic interactions may be of more limited value.

The rest of the chapter is organized as follows. In Section 5.3 we first discuss relevant literature and our own previous work that provides the motivation for the current work. We present the system architecture in Section 5.4. The experimental investigation is discussed in Section 5.5 and 5.6. Finally we summarize our contributions and discuss its impact and future directions in Section 5.7.

5.3 An Earlier Work on Robot-mediated Joint Attention

This work explicitly focuses on realizing a co-robotic interaction architecture capable of measuring behavior and adapting performance in a way that addresses a fundamental early impairment of ASD (i.e., joint attention skills).

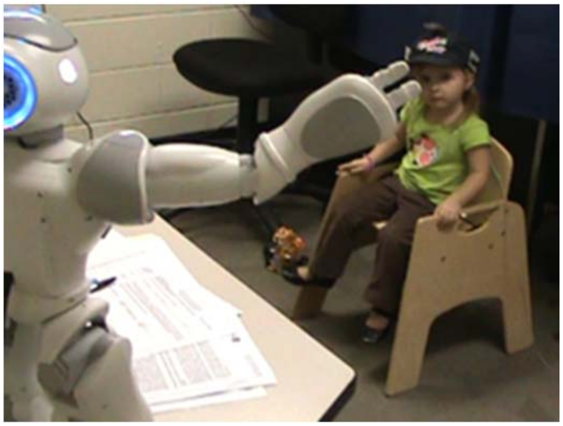


Fig. V.1 Robot-mediated joint attention study (with permission)

In order to determine the feasibility and potential value of adaptive robotic intervention system for younger children, Bekele et al. developed the prototype ARIA system capable of administering joint attention tasks to young children with ASD (Fig. V.1) [5-8]. In this work, the authors developed a test-bed that consisted of a humanoid robot NAO, 3 Infrared (IR) cameras mounted in the test room, and a series of 23 inch networked computer monitors capable of displaying relevant recorded task stimuli. This work also instrumented a baseball hat with arrays of IR LEDS and designed a gaze inference algorithm based on real-time image processing of the camera images obtained from LEDS arrays. The algorithm could detect gaze with both head pitch and yaw angles with validation detection data with a laser pointer yielding an error bounding box to be within 2.6 cm X 1.5 cm from 1.2 meters distance.[7] This study performed an initial feasibility study comparing performance and gaze detection for a sample of 6 typically developing children and children with clinically confirmed ASD diagnosis (ages 3-5; IQ range = 49-102) and variable baseline skills regarding response to joint attention (ADOS RJA Item range: 0-3, mean = 1.2 (1.1)). Within the system a series of joint attention prompts were administered via either a human administrator (x2) or the humanoid robot (x2) with randomized presentation to control order effect. The child sat in a chair across from the robot or interventionist for the trial block and was instructed through a hierarchy of prompts (i.e., head/gaze shifts, pointing, target activation) to look to a target.

The system registered gaze across all trials and provided reinforcement for looking through a simple reinforcement protocol (e.g., praise and target activation). Available data suggest that children with ASD spent approximately 27% more time looking toward the robot administrator than the human administrator, that they did not fixate on either robot or target, and ultimately directed gaze correctly to the target for 95.83% of the total 48 trials, a rate equal to TD success. Further, children successfully oriented to robotic prompting, meaning they responded to robot prompts prior to target activation, at very high levels (i.e., ASD = 77.08% success; TD = 93.75%). However, note that this study was a single session study and as such did not provide any indication whether the children will respond similarly over multiple sessions. In terms of tolerability, we anticipated a fairly large fail rate across both the ASD and TD samples in terms of willingness to wear the LED cap. Out of 10 ASD and 8 TD children, 6 ASD and 6 TD children completed the study. The completion rates of 60% (ASD) and 80% (TD) were promising, but ultimately specifically highlighted the need for the development of a non-invasive system for realistic extension to a young ASD population commonly demonstrating sensory vulnerabilities.

5.4 System architecture

In this chapter we wanted to investigate the impact of such a robot-mediated joint attention intervention on attention and performance. As a result, we have modified the ARIA architecture described above in two important ways. First, we wanted to monitor the eye gaze of a child with ASD on the robot. We hypothesized that if the child gets bored with repeated exposure to the robot then he/she will look less at the robot over multiple trials. In order to capture the eye gaze we introduce an eye tracker into the architecture (Fig. V.2 and Fig. V.3) that monitored the gaze of the child on and around the robot. Second, in order to address the previously observed sensory vulnerabilities due to the wearing of the hat, we included a human therapist in the loop to replace the hat and the camera system who determined when and how the child responded to robotic prompts (Fig. V.3).

The robot-mediated intervention system (Fig. V.2) is developed around the robot NAO (Fig. V.1). The child with ASD is seated in a booster chair. The robot NAO [9], which is a humanoid robot produced by Aldebaran Robotics, stands on a raised platform in front of the child. NAO is the size of a young child (height = 58 cm, weight = approximately 4.3 kg) and is suitable to a robot-child interaction study. Its body is made of plastic and it has 25 degrees of freedom which allow the user to control its head, fingers and feet independently. Its software modules provide convenient programming and encourage distributed processing. NAO is capable of displaying complex social communication behaviors that are often missing or under-developed in children with ASD. An Eye Tracker, Tobii X120 [10], is calibrated around NAO such that it can capture the child's gaze when he looks at the robot. The calibration for the eye tracker is done by projecting the calibration image on a screen at the robot's position. Small cartoon characters were displayed in each calibration point to attract participant's eye gaze. After calibration, the screen is removed and the robot is placed at that position. In order to display where the participant is looking on or around the robot in real-time, we use a camera to record the robot motion and then superimpose this video with gaze data. This video is displayed in the monitoring station in another room for parents and other researchers to intuitively understand the participant's attention on robot.

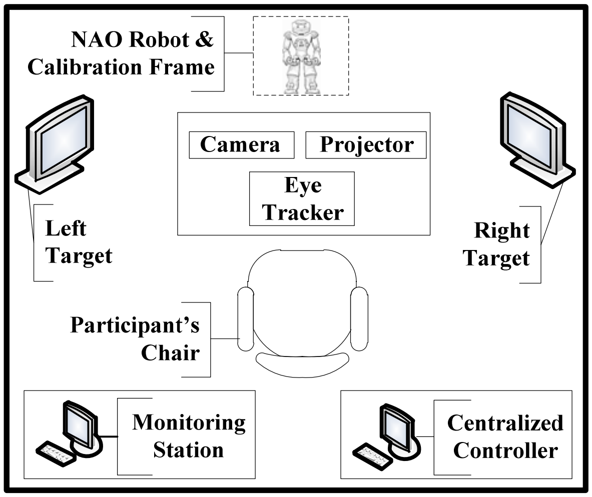


Fig. V.2 Experiment Room Sketch

There are two computer display monitors, one at the left and one at the right of the child, where joint attention stimuli are presented. The robot presents joint attention bids, which are discussed in Section 5.5, and a therapist observing the child's response indicates correct or incorrect response by pressing a button that is connected to the robot controller. Based on the child's response the robot either presents a new joint attention task if the previous task was successful, or increases the prompt level to get the child to look at the stimuli. The trial continues for a specified duration. The system interconnection is shown in Fig. V.3. The robot action and the stimuli presentation are coordinated by the Centralized Controller (CC), which performs the following tasks: i) initiate a session by sending messages to the eye tracker to initiate calibration and record time-stamped eye gaze data; ii) send message to the robot to initiate joint attention bid; iii) activate the display monitor with appropriate stimuli; iv) continuously monitor signals from the human therapist to determine whether a trial has been successful; v) based on performance continue with the trial as discussed above; and vi) end the session by sending messages to the robot, eye tracker and the display monitors. The CC is designed as an event-based system and communication between the CC and the different modules of the system has been implemented using socket communication.

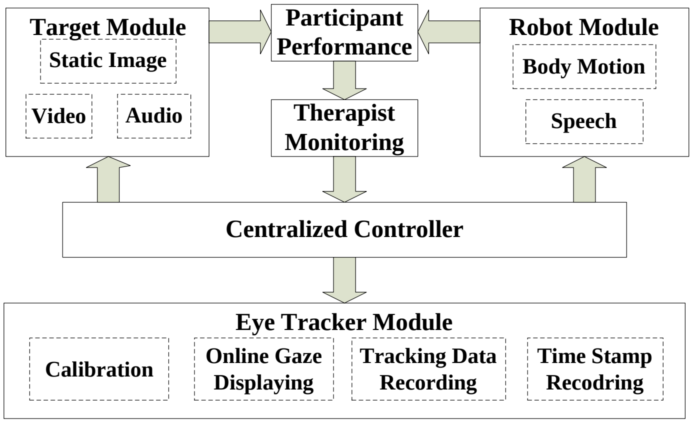


Fig. V.3 System Architecture

5.5 Experimental Setup

5.5.1 Participants

A total of 6 boys completed the tasks with their parents' consent. The details of the participants are given in Table V.1. It is a well-documented finding that males are more commonly affected with ASD than girls at a rate of almost 5:1. As such, this predominantly male sample was not atypical for this population. There was no drop-out in this study. Given the primary aims of the study regarding documenting attention/habituation and performance change over time within ASD group we did not include a comparison sample of typically developing controls.

Tables V.1 shows the specific age, baseline cognitive skills, and ratings of autism symptoms for participating children. All participants with ASD were recruited through existing clinical research programs at Vanderbilt University and had an established clinical diagnosis of ASD. The study was approved by the Vanderbilt Institutional Review Board (IRB). To be eligible for the study, participants had to be between 2–4 years of age and had to have an established diagnosis of ASD based on the gold standard in autism assessment, the Autism Diagnostic Observation Schedule (ADOS) [11]. The parents completed ASD screening/symptom measurements: the Social Responsiveness Scale (SRS) [12] and the Social Communication Questionnaire (SCQ) [13].

Table V.1 Diagnoses of the Participants with ASD

	ADOS Raw Score	*ADOS Severity Score*	*SRS-2 Raw Score*	*SRS-2 T score*	*SCQ Lifetime Total Score*	*IQ*	*Age*
P1	20	9	132	85	24	81	4.38
P2	14	5	65	59	11	69	2.52
P3	24	9	107	75	16	107	2.75
P4	18	10	92	69	18	78	3.48
P5	20	9	106	75	24	58	3.48
P6	29	10	75	63	9	49	4.13
Mean	20.83	8.67	96.17	71.00	17.00	73.67	3.46
Std	5.15	1.86	24.23	9.38	6.32	20.29	0.73

5.5.2 Joint attention stimuli, robot prompts, and experimental procedures

The stimuli presented via the monitors were pictures of interest (e.g., characters/objects, pictures of caregivers/children/animals), videos of similar content, and discrete audio and visual events relevant to the pictures. These stimuli were adaptively changed in form or content based on participant's response in order to provide additional levels of prompts toward target and to ensure that they function as reinforcing objects of interest. To give an example, the pictures, audio, and video clips were carefully selected from children's TV programs (e.g., Bob the Builder, Dora the Explorer, etc.). Segmented clips of these shows were selected wherein a dance, performance, or other actions were carried out by the character such that the clip could be easily initiated and ended without abrupt start or end. The clips were also selected based on consultant review that the particular segments were developmentally appropriate and potentially reinforcing to our ASD population. Within the study, these clips were selected to be both part of the prompt and feedback structure, utilized to draw attention as needed and reinforce correct looking.

During joint attention tasks, a hierarchy of prompts was presented by the robot. We choose a least-to-most prompt (LTM) hierarchy [14], a common convention in ASD intervention, which essentially provides support to the learner only when needed. The method allows for independence at the outset of the task, ensures opportunities for successful performance and reinforcement at baseline, and only provides increasing support when the child has been given an opportunity to display independent skills. Table V.2 explicitly demarcates the prompt hierarchy utilized in our preliminary studies of robot assisted joint attention platforms which emphasizes utilizing the least level of prompting to achieve success as a methodology for shaping and improving performance over time.

Table V.2 Prompt Hierarchy for Child Named Jim

Prompt Level	***Robot Speech***	***Robot Motion***	***Target Display***
1	"Jim, look!"	Turn head	Static picture
2	"Jim, look!"	Turn head	Static picture
3	"Jim, look over there!"	Turn head and point	Static picture
4	"Jim, look over there!"	Turn head and point	Static picture
5	"Jim, look over there!"	Turn head and point	Audio display (3 sec)
6	"Jim, look over there!"	Turn head and point	Video display (10sec)

Each participant attended 4 sessions, scheduled on different days to assess the cumulative effect of the experimental sessions on participants' attention and performance. During each session, participants were told they would be "playing with Mr. Robot Nao." Instructions were, "If you follow Mr. Robot's words, he will reward you!" To heighten children's engagement across sessions, we used different video sets as rewards but kept the main procedure in every session identical. Each session included 8 trials, described below for a total of 32 trials across all sessions.

In each trial, there are 6 potential prompt levels. After prompt 2, the robot NAO engages in successively more attention-directing Robot Speech or Robot Motion. The Target Display also becomes more attention-getting as more prompts are required. For each of the 8 trials, the system randomly put the target on the left or right monitor. The target direction remained the same within each trial. The robot turned its head or turned while pointing to the corresponding target. After the start of each prompt, a 7 second response time window was set. “Target hit” was defined as the participant responding to, i.e., turning to look at the correct target within this 7 seconds. Regardless of the participant response, the robot turned back to the starting neutral posture. If the participant failed to hit the target, the robot proceeded to the next prompt level. If the participant hit the target, the reward video was shown and then next trial started. Note that during the first 4 prompts, only one static cartoon picture is presented on both left and right targets. If the participant failed to hit the target in the first 4 prompts, audio and then video were provided to further draw the participant’s attention to the target. If the participant successfully hit the target before prompts 6 they were rewarded with a 10s video on the target display immediately (for prompt 6, the video was incorporated into the prompt.) After each successful hit, NAO also gave the verbal reward.

5.6 Experimental Results

To evaluate the system’s effect on participants’ responses to the robot, we analyzed 1) target hit response rates and 2) eye gaze data. The former evaluated whether participants followed the robot’s prompt and looked at the target. The latter evaluated participants’ attention toward the robot during the interaction.

5.6.1 Target hit response performance

Across all sessions and participants, 99.48% of the 32 trials ended with a target hit. This illustrates that our robot-mediated joint attention intervention system successfully caught and transferred the attention of children with ASD. The average prompt level before participants looked at the target is shown in Fig. V.4.

Fig. V.4 displays how participants’ performance, as measured by number of prompts until target hit, improved from session 1 to session 4. In session 1, the average target hit prompt level is 2.17. As children completed sessions and became more familiar with the “game”, the target hit prompt level went lower, falling to 1.44 by session 4. A two-sided Wilcoxon rank-sum test indicated that the median difference between session 1 and session 4 is statistically significant (p = .0029). *In other words, with more exposure, the performance improvement was found to be statistically significant.* For prompt1 to prompt4, the target was only a static picture (no sound or animation). Therefore, if the participant hit the target, one can assume it was because the participant understood the robot’s instruction and accompanying look/gesture. Fig. V.5 shows the average percentage of target hits at prompt1 and with prompt 1-4. The trends for the two cases are all increasing. In the first session, 52.8% of trials ended with a target hit on the first prompt, while in session 4 participants achieved 81.25% of target hits on the first prompt. Participants hit the target within the first 4 prompts 87.5% of the time in session 1, and that increased to 95.83% in session 4.

Therefore, by session 4, the participants hit the target by following the robot's gesture and instruction alone, without additional attraction from target itself.

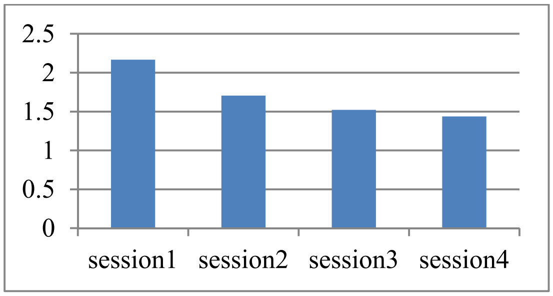


Fig. V.4 Average prompt level needed for target hit

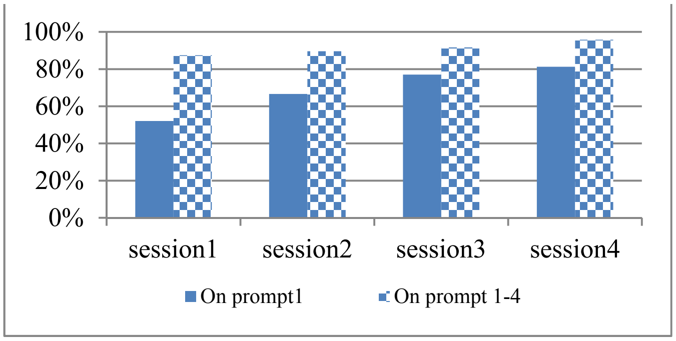


Fig. V.5 Average target hit at prompt level1 (solid) and below prompt level 5 (chessboard pattern)

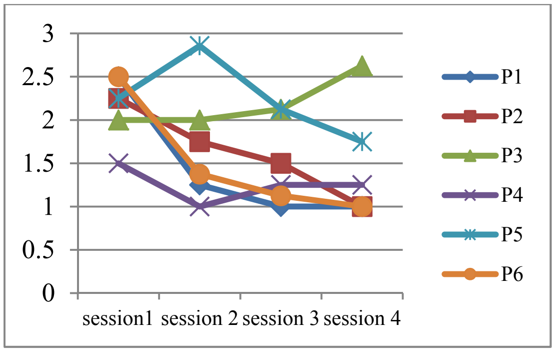


Fig. V.6 Individual performance of the 6 participants

Fig. V.6 shows the individual performance for each of the 6 participants, with the average number of prompts needed for each trial depicted on the y-axis. From Fig. V.6, we can see that four out of the six participants' performance improved, one fluctuated and one decreased.

5.6.2 Eye gaze analysis

We defined the robot attention gaze region as a box of 76cm× 58 cm which covered the body and arm movement of NAO. Given the distance from the participant to the calibration screen/robot, the accuracy of gaze detection if the participant moved his or her head was about 5cm in both the horizontal and vertical directions. The following analysis explains participants' attention to the robot in terms of their eye gaze pattern.

The gaze pattern is analyzed in two ways: 1) The whole session (from the start of the first prompt to the end of the session) and 2) Within the 7 second response time window for each prompt in a trial. For example, if the participant hit the target on the 5th prompt in a trial, then the total time looking at the robot region for that trial is the sum of looking time across all five 7 second time windows. Examining looking times across all participants and sessions, the average time that participants looked at the robot was 14.75% of total experiment time. Within the 7 second window across all participants and sessions, the average time that the participants looked at the robot was 24.80% (Fig. V.7). From session 1 to session 4, participants' average time looking at the robot were 14.88%, 15.17%, 17.94%, and 11.02% for the whole session, and 22.15%, 26.52%, 28.14%, and 22.41% for the 7 second response window.

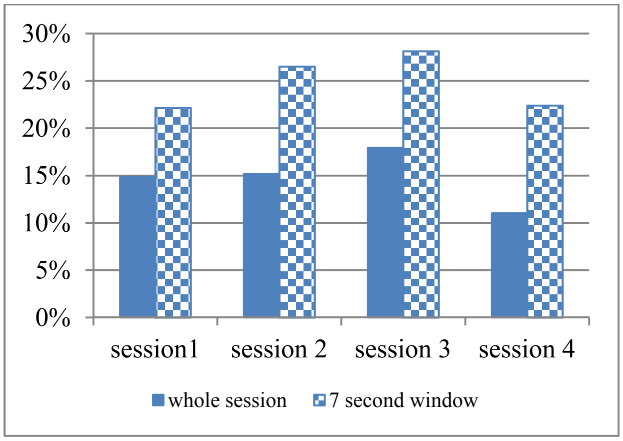


Fig. V.7 Average looking at robot time for whole session and 7 second respond window

A two-sided Wilcoxon rank-sum test showed that the median difference in looking time between all sessions was not statistically significant, with p-values range from 0.8850 to 0.1797 between different sessions. This indicates that the looking patterns of participants did not change statistically significantly across sessions. *In other words, the initial interest that the participants showed towards the robot did not statistically significantly change with repeated exposure.*

Qualitatively from the therapist's analysis we found that children's attention in session 1 was initially described as focused on the robot itself. At first, they were described as attracted to this unusual "playmate's" special appearance, accent and flashing LED eyes. Sometimes they were described as seemingly distracted by these aspects and ignored the robot's instructions. Beginning in session 2, participants were described as focusing on the robot's instructions with increasingly better performance, as previously discussed. In session 4, the participants were quite familiar with the "game". They stared at the robot less and responded to the target quickly once the robot gave out the instruction, waiting for the reward. Because of this, the average robot looking time was lower in session 4.

Statistical analyses showed that participants looked at the robot more in the response window than after, which explains why the percentage of looking time for the response window is larger than the one for whole session. We also found that in the video task (prompt 6 and video reward), the children usually looked at the monitor instead of the robot. This suggests that the video effectively captured children's attention for the final prompt as well as the reward condition.

5.7 Discussion and Conclusions

In this work, we studied the development and application of an innovative adaptive robotic system with potential relevance to core areas of deficit in young children with ASD. The ultimate objective of this study was to empirically test the attention and performance of a robotic system capable of administering and altering a joint attention hierarchy based on performance.

Children with ASD documented sustained interest with the humanoid robot over several sessions and demonstrated improved performance within system regarding joint attention skills. These findings together are promising in both supporting system capabilities and potential relevance of application. Robotic systems endowed with enhancements for successfully pushing toward correct orientation to target either, with systematically faded prompting or potentially embedding coordinated action with human-partners, might be further capable of taking advantage of baseline enhancements in non-social attention preference [15, 16] in order to meaningfully enhance skills related to coordinated attention.

There are several methodological limitations of the current study that are important to highlight. The small sample size examined is the most powerful limits of the current study. As such, while we are left with data suggesting the potential of the application, the utilized methodology, potently restricts our ability to realistically comment on the value and ultimate clinical utility of this system as applied to young children with ASD.

Another important technical limitation was the utilization of a human confederate within the robotic system loop. While this modification from our original closed-loop system resulted in dramatic improvement in terms of tolerability (all children completed the protocol), such wizard-of-oz paradigms carry additional human resource burdens to accomplish. This highlights the need to develop a non-contact remote eye gaze tracker capable of integration into a closed-loop system.

Despite limitations, this work was the first to our knowledge to design and empirically evaluate the usability, feasibility, and preliminary efficacy of an adaptive interactive robotic technology capable of modifying performance regarding joint attention skills for young children with ASD. Few other existing robotic systems [17, 18] for other tasks have specifically addressed how to detect and flexibly respond to individually derived, socially and disorder relevant behavioral cues within an intelligent adaptive robotic paradigm for young children with ASD. Movement in this direction introduces the possibility of realized technological intervention tools that are not simple response systems, but systems that are capable of necessary and more sophisticated adaptations. Systems capable of such adaptation may ultimately be utilized to promote meaningful change related to the complex and important social communication impairments of the disorder itself.

Ultimately, questions of generalization of skills remain perhaps the most important ones to answer for the expanding field of robotic applications for ASD. While we are hopeful that future sophisticated clinical applications of adaptive robotic technologies may demonstrate meaningful improvements for young children with ASD, it is important to note that it is both unrealistic and unlikely that such technology will constitute a sufficient intervention paradigm addressing all areas of impairment for all individuals with the disorder. However, if we are able to discern measurable and modifiable aspects of adaptive robotic intervention with meaningful effects on skills seen as tremendously important to neurodevelopment, or tremendously important to caregivers, we may realize transformative accelerant robotic technologies with pragmatic real-world application of import.

5.8 References

[1] C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., “Language outcome in autism: Randomized comparison of joint attention and play interventions.,” *Journal of consulting and clinical psychology,* vol. 76, no. 1, pp. 125–137, 2008.

[2] C. Kasari, Gulsrud, A.C., Wong, C., Kwon, S., Locke, J., “Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism,” *Journal of autism and developmental disorders,* vol. 40, no. 9, pp. 1045-1056, 2010.

[3] G. Dawson, “Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder,” *Development and Psychopathology, Cambridge Univ Press,* vol. 20, no. 3, pp. 775-803, 2008.

[4] K. K. Poon, Watson, L.R., Baranek, G.T., Poe, M.D., “To What Extent Do Joint Attention, Imitation, and Object Play Behaviors in Infancy Predict Later Communication and Intellectual Functioning in ASD?,” *Journal of Autism and Developmental Disorders, DOI: 10.1007/s10803-011-1349-z*, pp. 1-11, 2011.

[5] E. Bekele, U. Lahiri, A. Swanson, Julie A. Crittendon, Zachary Warren, and N. Sarkar, “A Step towards Developing Adaptive Robot-mediated Intervention Architecture (ARIA) for Children with Autism,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2012 (in press, DOI:10.1109/TNSRE.2012.2230188).

[6] E. Bekele, Swanson, A., Crittendon, J., Sarkar, N. and Warren, Z, “Pilot Clinical Application of an Adaptive Robotic System for Young Children with Autism Spectrum Disorder.,” *Journal of Autism and Developmental Disorders (Autism) (accepted for publication)*, 2012.

[7] E. Bekele, Lahiri, U., Davidson, J., Warren, Z., Sarkar, N., “Development of a novel robot-mediated adaptive response system for joint attention task for children with autism,” in IEEE International Symposium on Robot Man Interaction (Ro-Man), Atlanta, GA, 2011, pp. 276-281.

[8] E. T. Bekele, A. Swanson, A. C. Vehorn, J. A. Crittendon, Z. Warren, and N. Sarkar., “Robot-Mediated Adaptive Response System in Joint Attention Task for Children with Autism Spectrum Disorders.,” in 11th annual International Meeting for Autism Research (IMFAR), Toronto, Canada, 2012.

[9] E. Bekele, Z. Zheng, U. Lahiri, A. Swanson, J. Davidson, Z. Warren, and N. Sarkar, “Design of a novel virtual reality-based autism intervention system for facial emotional expressions identification,” in International Conference on Disability, Virtual Reality and Associated Technologies, 2012, pp. in press.

[10] D. Annaz, R. Campbell, M. Coleman, E. Milne, and J. Swettenham, “Young children with autism spectrum disorder do not preferentially attend to biological motion,” *Journal of autism and developmental disorders,* vol. 42, no. 3, pp. 401-408, 2012.

[11] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, “The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism,” *Journal of Autism and Developmental Disorders,* vol. 30, no. 3, pp. 205-223, 2000.

[12] J. Constantino, and C. Gruber, “The social responsiveness scale,” *Los Angeles: Western Psychological Services*, 2002.

[13] M. Rutter, A. Bailey, C. Lord, and S. Berument, “Social communication questionnaire,” *Los Angeles, CA: Western Psychological Services*, 2003.

[14] M. Demchak, “Response prompting and fading methods: a review.,” *American Journal of Mental Retardation,* vol. 94, no. 6, pp. 603-615, 1990.

[15] A. Klin, Lin, D.J., Gorrindo, P., Ramsay, G., Jones, W., “Two-year-olds with autism orient to nonsocial contingencies rather than biological motion,” *Nature,* vol. 459, no. 7244, pp. 257–261, 2009.

[16] D. Annaz, R. Campbell, M. Coleman, E. Milne, and J. Swettenham, “Young Children with Autism Spectrum Disorder Do Not Preferrentially Attend to Biological Motion.,” *J Autism Dev Disord.,* vol. 42, no. 3, pp. 401-408., 2012.

[17] C. Liu, Conn, K., Sarkar, N., Stone, W., “Online affect detection and robot behavior adaptation for intervention of children with autism,” *Robotics, IEEE Transactions on,* vol. 24, no. 4, pp. 883 - 896, 2008.

[18] D. Feil-Seifer, Mataric, M., “Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features,” in In Proceedings of the 6th international conference (ACM/IEEE) on Human-robot interaction, New York, NY: ACM Press, 2011, pp. 323-330.

Chapter VI. Autonomous robot-mediated joint attention intervention

6.1 Abstract

In this chapter, we propose a novel joint attention intervention system for children with ASD that overcomes several existing limitations in this domain such as the need to use body-worn sensors, non-autonomous robot operation requiring human involvement and lack of a formal model for robot-mediated joint attention interaction. We present a fully autonomous robotic system, called Norris, that can infer attention through a distributed non-contact gaze inference mechanism with an embedded Least-to-Most (LTM) robot-mediated interaction model to address the current limitations. The system was tested in a longitudinal user study with 14 young children with ASD. The results showed that participants' joint attention skills improved significantly, their interest in the robot remained consistent throughout the sessions, and the LTM interaction model was effective in promoting the children's performance.

keywords—Robot-mediated intervention, joint attention, children with ASD

6.2 Introduction

Primarily animal-like robots [1, 2] and small humanoid robots [3, 4] have been used for studies with children with ASD. Kozima et al.[5] designed a small creature-like robot called "Keepon", which successfully elicited positive social interaction behaviors in children with ASD. A humanoid robot called "KASPER" [6, 7], has been successfully used to facilitate collaborative play and tactile interaction with children with ASD. Feil-Seifer and Mataric [8] found that contingent activation of a robot during interactions yielded immediate short-term improvement in social interactions. While important to demonstrate the potential of HRI in ASD intervention, most of these earlier studies chose free play as the mode of interaction instead of focusing on the core deficits of ASD. However, studies in ASD intervention have shown that interventions are most effective when the intervention is focused on the core deficits of ASD [9]. In addition, most of these previous HRI systems were open-loop systems and thus were not responsive to the dynamic interaction cues from the participants to be able to adapt and individualize intervention. The primary goal of the current work is to design a fully autonomous closed-loop robotic system that can target core deficits of ASD.

We introduce a new closed-loop autonomous robotic system, named Norris (short for Non-contact Responsive Robot-mediated Intervention System), to help children with ASD learn joint attention skills. Joint attention is the process of sharing attention and socially coordinating attention with others to effectively learn from the environment [10]. Joint attention skills underlie the neurodevelopmental cascade of ASD, and successful intervention targeted on joint attention is essential to improve numerous developmental skills in children with ASD [11].

The current work improved previous work [12, 13] in a number of important ways. While [12] presented a novel HRI architecture for ASD intervention, ARIA, and developed an effective Least-to-Most (LTM) protocol for joint attention intervention with promising results, it required participants to wear an instrumented hat for gaze inference. Since many young children with ASD are sensitive to unfamiliar touch [14], close to 40% children did not want to wear the hat and thus could not take part in the intervention. In order to solve this problem, Zheng et al. [13] developed another robotic system that inherited the LTM protocol from ARIA but used a Wizard of Oz [15] strategy for gaze detection to eliminate the need for the instrumented hat. While it enabled 100% participation, the system became semi-autonomous and needed human involvement for gaze inference. Additionally, both [12] and [13] used LTM protocol for joint attention but did not provide a generalizable mathematical model for LTM interaction. In LTM, the teacher allows the learner an opportunity to respond independently on each training stage and delivers the least intrusive prompt first. If necessary, more intrusive prompts, usually upgraded based on the previous prompts, are then delivered to the learner to complete each training procedure [16]. Essentially, LTM provides support to the learner only when needed. LTM has been widely applied in diagnostic and screening tools for children with ASD [17, 18]. However, to our knowledge, no mathematical model of LTM has been presented in the literature such that LTM based interaction can be generalized for multiple skill training.

The contributions of the current study are two-fold: 1) development of a new fully autonomous closed-loop robot-mediated intervention system that can infer gaze non-invasively and is capable of administering LTM protocol based on a general mathematical model; and 2) results from a feasibility joint attention intervention user study that tested the newly developed system in a multi-session longitudinal study.

The remainder of this chapter is organized as follows: Section II describes the architecture and components of Norris. Section III introduces the mathematical model for LTM as the interaction logic which is followed by the design of the feasibility longitudinal user study to validate the Norris system in Section IV. Sections V and VI present the results of this user study and the summary of contributions and limitations of the chapter, respectively.

6.3 Norris System Architecture and Components

6.3.1 System Architecture

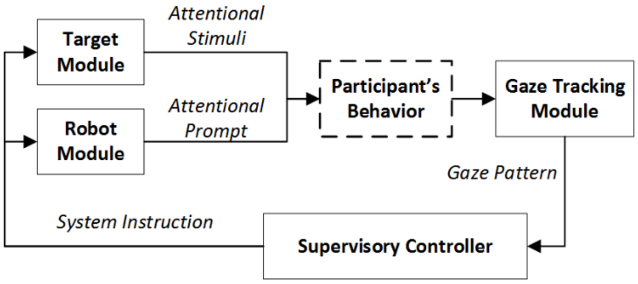


Fig. VI.1 Norris system Architecture

HRI using Norris is designed to work as follows. A child with ASD will be seated in a room in front of a humanoid robot. The room will be equipped with a set of spatially distributed computer monitors or TVs where audio-visual stimuli will be presented. The robot will administer a LTM based joint attention prompting protocol to the child and the child's response in terms of gaze direction will be inferred by a set of distributed cameras. Based on whether the child shares attention or not the robot will provide appropriate feedback and move on to the next prompt. As shown in Fig. VI.1, the Norris has 4 main components: 1) the robot module controls robot actions; 2) the target module controls environmental factors; 3) the gaze tracking module provides interaction cue sensing; and 4) the supervisory controller controls the interaction logic. The supervisory controller is the "brain" of Norris that sends commands to the robot and the target module to present directional prompts to the participant. For example, the robot can turn its head to a monitor displaying a picture, and ask the participant to look at that monitor. The participant may or may not look at the monitor, and this looking behavior is sensed by the gaze tracking module. The tracking module further computes whether the direction of the participant's gaze falls on the monitor, and sends this message back to the supervisory controller. Then the supervisory controller sends commands to the robot and target module again telling what to show next, based on an interaction protocol. Therefore, Norris provides a fully autonomous closed-loop interaction between the system and the participant.

6.3.2 System Components

Robot module

A humanoid robot NAO by Aldebaran Robotics [19] was embedded in Norris. NAO has been widely applied for children with ASD [3, 12] due to its attractive childlike appearance and high controllability.

We designed a new controller for NAO that communicates with the supervisory controller. The robot controller was embedded with a built-in library storing all the necessary motions (e.g., turning its head to a monitor) and speeches (e.g., asking the participant to look in a direction) needed for the interaction. The robot's actions are detailed in Section III.B along with the interaction protocol.

Target module

Two flat TVs (width: 70cm; height: 43 cm), one to the right and one to the left of the participant, were used as attentional targets. The robot would point to one of the monitors at a time and ask the participant to look at what was being shown in that monitor. The two TVs were controlled individually by two target controllers that received commands from the supervisory controller. A library of pictures, audios, and videos were embedded in the target controller. Based on the commands sent from the supervisory controller, static pictures, audios, or videos of children's interest were displayed. The set of target actions are detailed in the interaction protocol (Section III.B).

Gaze tracking module

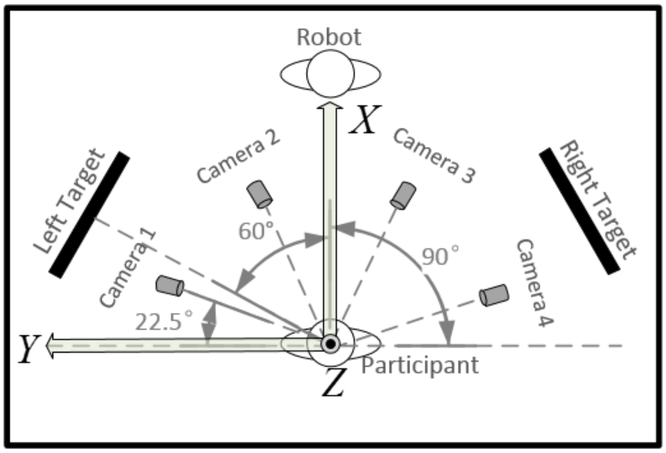


Fig. VI.2 Top view of Norris and the global frame

The gaze tracking module detected the participants' looking behavior. The direction of a participant's gaze was computed based on the orientation of his/her head as detected by a set of cameras as shown in Fig. VI.2 and Fig. VI.3. Fig. VI.2 illustrates the top view of Norris in the global reference frame. The center of the participant's head was the origin of the global frame. The *X*-axis and the *Y*-axis pointed forward and to the left of the participant, respectively, and the *Z*-axis pointed upwards out of the plane.

Fig. VI.3 shows both the body-attached head frame of the participant and the global reference frame that share the same origin. If the participant did not perform yaw (around the *Z*-axis), pitch (around the *Y*-axis), or roll rotations (around the *X*-axis), the head frame was aligned with the global frame. The unit vector along the positive *x*-axis of the head frame, $\overrightarrow{V_f}$, represents the frontal head orientation, and was used to derive the gaze direction. Four cameras were employed for gaze detection, each with its own coordinate system. The gaze tracking method has 3 steps as discussed below. It is to be noted that while Step1 and Step2 were inherited from our previous work [20], Step3 was newly developed in the current study.

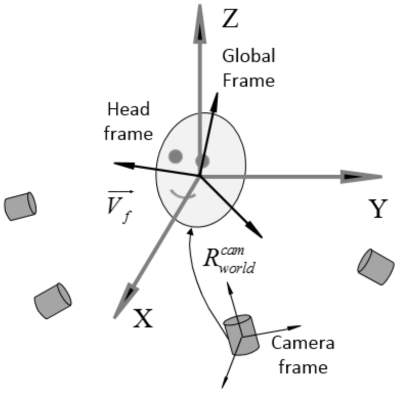


Fig. VI.3 Coordinate systems for gaze tracking

Step1: Detect head orientation from a camera. The SDM [21] method was applied to each camera to achieve fast and robust head orientation estimation with respect to the camera's frame. The image of the participant's frontal face is needed for this estimation, and thus a given camera can only detect head orientation when the frontal face is visible to it. The detected head orientation is represented by $\overrightarrow{V_f}$ in the camera's frame. However, in the current joint attention study, we wanted to detect a larger head yaw angle (about 180°) than what can be detected by one camera (about 80°) for realistic tasks. Therefore, we developed a distributed head orientation estimation algorithm for an array of 4 cameras (as shown in Fig. VI.2) around the participant with partially overlapping views to extend the detection range. This design guaranteed that no matter which part of the interaction environment the participant was looking at, at least one camera could capture his/her frontal face in order to conduct head orientation estimation.

Step2: Transform the head orientation estimation from a camera's frame to the global frame. Each camera was calibrated to get the transformation matrix, R^{cam}_{world} , between the camera's frame and the global frame. As shown in Fig. VI.3, R_{world}^{cam} transform $\overrightarrow{V_f}$ from the camera's frame to the global frame.

In order to correlate head orientation with gaze direction we conducted a small study with 10 adults where the volunteers were asked to look at a marker in front of them that was moved from left to right 5 times followed by a right to left marker movement for another 5 times.

In the horizontal direction, we trained a mapping function to derive α_g from α_f. α_g was identified as the angle of the moving marker (between -90° (left) and 90° (right)). Simultaneously, the volunteers' horizontal head orientation, α_f, was estimated by the camera array. A total of 22,236 data pairs were collected. We used polynomial fitting to reflect the relation between α_g and α_f. In Fig. VI.4(b), the blue points indicate the pairs (α_f, α_g). The red curve with a sigmoid shape is the curve that maps α_f. The equation of the red curve is

$$\alpha_g = 5.88^{-5}\alpha_f{}^3 - 8.94^{-4}\alpha_f{}^2 + 0.97\alpha_f + 1.50 \cdot \tag{VI-1}$$

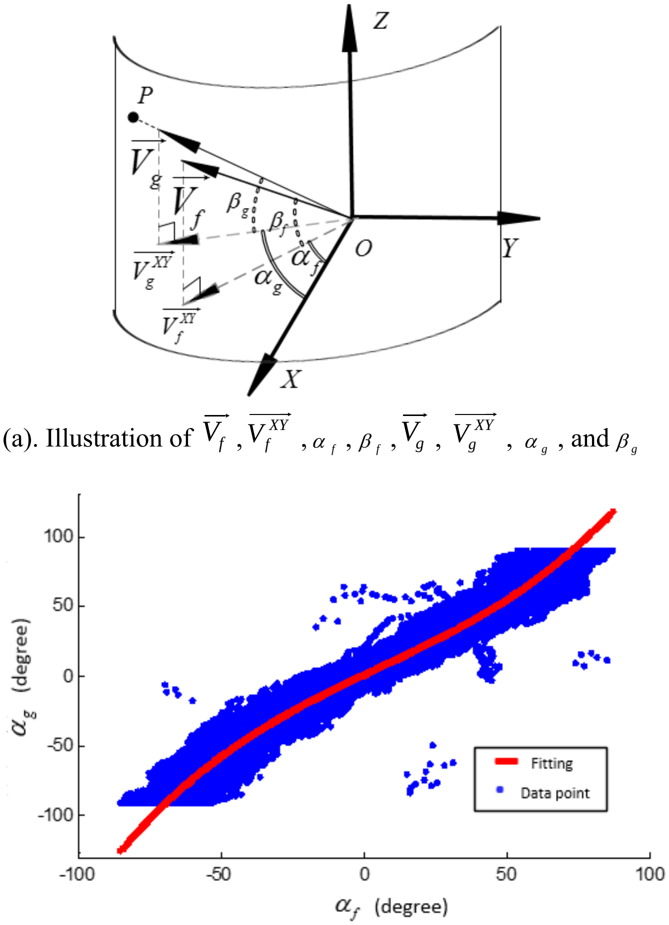
The mean distance from the data points to the red curve is 9.12°. We can see that, in general, a larger α_f leads to a larger $\|\alpha_g - \alpha_f\|$. Intuitively, the more the participant's head turned to the side, the larger the deviation between the gaze direction and the frontal head orientation in the horizontal direction.

The vertical gaze direction was approximated by $\beta_g = \beta_f - \beta_{baseline}$. Here $\beta_{baseline}$ is an offset angle which was calibrated for each participant. During the calibration, the participant was prompted to look along the *X*-axis, and β_f at this moment was recorded as $\beta_{baseline}$. In the user study presented in Section IV, $\beta_{baseline}$ ranges from -11.57° to 8.14°.

The range of both monitors can be represented by the values α_g and β_g. The α_g range of the left and the right monitors were [-70°, -50°] and [50°, 70°], respectively. However, in order to accommodate for mapping error as well as encouraging young children with ASD to continue with the intervention, we relaxed the range by 15° on each side. Similarly, the range of β_g was [-26.3°, 12.5°] from top to bottom, which covered an additional 43 cm (the height of the monitor) beyond the monitor's top and bottom edges. Therefore, if $\alpha_g \in [-85°, -35°]$, and $\beta_g \in [-26.3°, 12.5°]$, the system would infer that the participant responded to the left monitor. Similar ranges were applied for the right monitor. On average, the whole gaze tracking module refreshed at a speed of 15 fps.

Supervisory controller

The supervisory controller communicated with different system components and controlled the global logic of the interaction. The communication was implemented with TCP/IP Socket Communication method. The average communication time from sending a message to receiving the message between the supervisory controller and a system component was about 25ms, which guaranteed real-time closed-loop interaction. The global interaction logic, which we call the interaction protocol, is discussed in detail in Section III.



(a). Illustration of $\overrightarrow{V_f}$, $\overrightarrow{V_f^{XY}}$, α_f , β_f , $\overrightarrow{V_g}$, $\overrightarrow{V_g^{XY}}$, α_g , and β_g

(b). Mapping from horizontal head orientation to horizontal gaze direction

Fig. VI.4 Gaze direction computation in the global frame

6.4 LTM Interaction Protocol

The Least-to-Most (LTM) hierarchy was applied in Norris to form the interaction protocol. LTM has been widely applied in diagnostic and screening tools for children with ASD [17, 18]. In LTM, the teacher allows the learner an opportunity to respond independently on each training stage and delivers the least intrusive prompt first. If necessary, more intrusive prompts, usually upgraded based on the previous prompts, are then delivered to the learner to complete each training procedure [16]. Essentially, LTM provides support to the learner only when needed.

LTM has been applied in a few important robot-mediated intervention systems for children with ASD. Feil-Seifer et al. [22] and Greczek et al. [3] introduced the graded cueing feedback mechanism to teach imitation skills to children with ASD. In this mechanism, higher prompts were upgraded based on the initial prompt by adding additional verbal and gestural hints to help children copy gestures. Huskens et al. [23] used a robot to prompt question-asking behaviors in children with ASD. The robot used open-question prompt initially. If the participants did not respond correctly, the robot would add more hints (e.g., adding part of the correct response) in the following prompts. Zheng et al. [24] designed a robot-mediated imitation learning system using a prompting protocol to help address an incorrect imitation. The robot first showed a gesture to the participant and asked him/her to copy it. If the child could not do it correctly, the robot would point out where to improve in the following prompts. Bekele et al. [12] developed the ARIA system to teach joint attention skills to children with ASD. If the participants did not respond to simple directional prompts given by a robot, higher levels of prompts with additional visual and verbal directional hints were provided. Kim et al. [25] designed a robot-assisted pivotal response training platform, where the higher levels of prompts were built by adding target responses hints on lower level of prompts.

From these examples, we can see that LTM is not limited to a specific skill, but can be used as a general guidance mechanism for robot-mediated intervention. However, to our knowledge, no mathematical model has been proposed to create a general LTM framework. In this work, we attempt to develop a general LTM-based Robot-mediated Intervention (LTM-RI) model. Such a model can be used to teach different skills to children with ASD as well as adapt the prompts for a specific skill. We expand the model for joint attention intervention, which is used for the user study.

6.4.1 LTM-RI model

An intervention system uses prompts to teach a skill to children with ASD. These prompts may consist of robot actions (e.g., motions and speeches), and may also include environmental factors that coordinate with the robot's actions (e.g., the attentional target that the robot may refer to). Suppose we have libraries of different robot actions $\{RA\}$ and environmental factors $\{EF\}$, then we can combine different RA_i and EF_j to form different prompts. These combinations may have different strength in eliciting the expected response (e.g., child looking at the target monitor), $ExpResp$, from a child. For example, in the current joint attention study, the robot turning its head (*RA*) to a static picture display on the TV (*EF*) might have a weaker impact on the children than pointing (*RA*) to a cartoon video displayed on the monitor (*RA*). Here we arrange the order of the elements in $\{RA_i\}$ and $\{EF_j\}$ as follows: RA_a is stronger (includes more instructive information) than RA_b if $a>b$; and EF_c is stronger than EF_d if $c>d$. These orders can be determined based on common sense and clinical experiences.

LTM-RI starts from presenting the weakest combination of *RA* and *EF* to form the least intrusive prompt. If this cannot elicit $ExpResp$, stronger *RAs* and *EFs* will be provided iteratively to form more instructive prompts, until the end of an intervention trial. We formally define this iterative procedure as follows.

LTM-RI Model

```
Step 1: Initial prompt (prompt level = 1).
        $Behavior(1)=BF\left(RobAction(1),EnviFactor(1)\right)$
        $Resp(1)=ICD\left(Behavior(1)\right)$
        If $Resp(1)=ExpResp$
            Reward
            Go to Step 3
Step 2: Iterative prompting loop.
    For prompt level n =2: IN
        $\left[RobAction(n),EnviFactor(n)\right]=PF\left(Resp(n-1)\right)$
        $Behavior(n)=BF\left(RobAction(n),EnviFactor(n)\right)$
        $Resp(n)=ICD\left(Behavior(n)\right)$
        If $Resp(n)=ExpResp$
            Reward
            Break
        $n=n+1$
Step 3: Termination.
    Robot naturally stops the interaction
```

Here BF is an implicit function that describes the participant's behavior (e.g., participant's gaze direction) given *RA* (*RobAction)* and *EF (EnviFactor).* This behavior is sensed by the interactive cue detection function (*ICD*) to determine whether the behavior is *ExpResp*. In the current study, *ICD* is the gaze tracking module). In the simplest scenario, we can categorize $Perf(n)=ExpResp$ and $Perf(n)\neq ExpResp$. *PF* is the prompting function which decides what *RA* and *EF* to present, if $Perf(n)\neq ExpResp$. Therefore, *PF* is a sorted list of prompt levels following the LTM heirarchy. We want to identify the lowest level of support needed by the participant to perform $ExpResp$. If $Perf(n)\neq ExpPerf$, given $RobAction(n\text{-}1)=RA_i$, and $EnviFactor(n\text{-}1)=EF_j$, we choose $RobAction(n)=RA_l$ ($l\geq i$) from $\{RA\}$ and $EnviFactor(n)=EF_k$ ($k\geq j$) from $\{EF\}$, so the the next prompt repeats the last prompt or provides a more instructive prompt. LTM-RI steps works are as follows:

In Step 1, the participant's baseline behavior $Perf(1)$ is evaluated by prompt level 1, which consists of the weakest *RA* ($RobAction(1)$) and *EF* ($EnviFactor(1)$). If the participant's response, $Resp(1)$, is $ExpResp$, then higher prompts are not needed. The system gives rewards and then excutes Step 3 to terminate the intervention. Otherwise, Step 2 is executed.

In Step 2, prompt level 2 is given first. If the participant cannot perform $ExpResp$, the higher prompts are presented one by one until level *IN*. During this iteration, the next level of prompt ($RobAction(n)$ and $EnviFactor(n)$) is formed based on the current performance of the participant $Perf(n-1)$, according to the *PF*. If $Perf(n) = ExpResp$, the system gives a reward to the participant and goes to Step 3.

We can see that if $ExpPerf$ happened on prompt level *n*, it means that level 1 to n-1 have been executed but failed to elicit $ExpResp$. Suppose ER_n means $Resp(n)=ExpResp$, and PT_x represents prompt level *x* had been executed but was not successful. Then $P(ER_n \mid PT_1,...,PT_{n-1})$ represents the probability that $ExpResp$ happens on prompt level *n*. In order to measure the impact of LTM-RI, we define an intensity function I_n as:

$$I_n = P(ER_1) + P(ER_2 \mid PT_1) + ... + P(ER_n \mid PT_{n-1},...PT_1) \text{.} \qquad \text{(VI-2)}$$

I_n represents the probability of $ExpResp$ at or before prompt level *n*. LTM-RI procedure has two goals:

Goal 1: $I_m > I_n$, given $m > n$. This means adding prompt levels increases the probability of $ExpResp$.

Goal 2: $I_{IN} = 1 - \varepsilon$, $\varepsilon = 0$ or ε is a small positive number. This means that eventually, at the highest prompt level, the system can elicit $ExpResp$ with high probability.

We can see that the LTM-RI is a general model that is not limited to one particular skill. What behaviors of the participants that the model tracks depends on the design of *ICD*. The number of prompt levels and the content of the prompts can be easily adjusted within this framework by changing the detail of *PF*. In the current work, LTM-RI is itemized for joint attention intervention.

6.4.2 LTM-RI trial in the current study

We designed the intervention trial of Norris based on LTM-RI. The *RAs* and *EFs* applied are shown in Table VI.1, which is the *PF* in LTM-RI. Apparently, the larger the subscript, the stronger the directional information was provided.

In Step 1 of LTM-RI (prompt level 1), the robot turned its head to the target monitor, saying "Look!"

(RA_1). At the same time, the monitor displayed a static picture (EF_1).

Table VI.1 Prompt Levels

Prompt level	***Prompting element list***
1 and 2	$RA_1 + EF_1$
3 and 4	$RA_2 + EF_1$
5	$RA_2 + EF_2$
6	$RA_2 + EF_3$
Prompt Elements (TR means Target Monitor): RA_1 : Robot turned its head to the TM, saying "Look!"; RA_2 : Robot turned its head and pointed its arm to the TM, saying "Look over there!"; EF_1 : TM displayed a static picture; EF_2 : TM displayed an audio clip; EF_3 : TM displayed a video clip.	

In Step 2 of LTM-RI, IN = 6. Prompt level 2 was the same as prompt level 1. In prompt level 3 and 4, the robot not only turned its head, but also pointed its arm to the target monitor, saying "Look over there!" (RA_2). At the same time, the monitor still displayed a static picture (EF_1). In prompt level 5 and 6, the robot action was kept as RA_2, but the monitor displayed an audio clip (EF_2) and an video clip (EF_3), respectively. At any time during a trial, if the participant looked at the target (*ExpResp* happened), the robot would say "Good job!" and the target monitor would display cartoon video for 10 seconds as rewards. Otherwise, the prompt level would be presented one by one until prompt level 6 was completed. Finally, Step 3 of LTM-RI was executed, where the robot returned to its standing position, thanked the participant, and said Goodbye.

In order to implement LTM-RI within Norris, we interpreted the LTM-RI trial with the standard Harel Statechart model [26], which is an extended state machine capable of modeling hierarchical and concurrent system states. As shown in Fig. VI.5, rectangles denote states. When an event happens, a state transition takes place, which is indicated by a directed arrow. Solid rectangles mark exclusive-or (XOR) states, and the dotted lines mark AND states. Encapsulation represents the hierarchy of the states. In the same hierarchy (encapsulated by the same rectangle), the system must be in only one of its XOR states, while in all of its AND states. Therefore, the AND states represent parallel processes in the system.

Variable: p: { 1,…,7}
Input: TO, Target hit: event

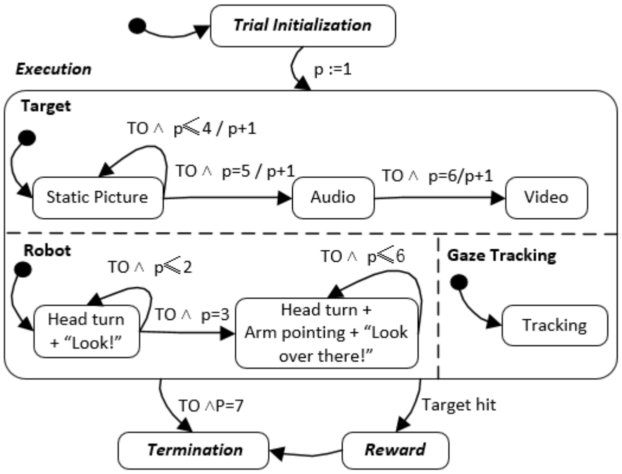


Fig. VI.5 The Harel Statechart model of the NorrisI LTM-RI trial

The first hierarchy includes 4 XOR states:

$$S1=\{\mathit{Initialization},\ \mathit{Execution},\ \mathit{Reward},\ \mathit{Termination}\}$$

At the beginning of a trial, the system is in the *Initialization* state, where the robot stands straight facing the participant. Then, the system transits to the *Execution* state and initializes variable $p=1$.

The *Execution* state is the second hierarchy, which includes 3 AND states, showing target, robot, and gaze tracking modules running in parallel.

$$\mathit{Execution}=\{\mathit{Target},\ \mathit{Robot},\ \mathit{Gaze\ tracking}\} \tag{VI-3}$$

The third hierarchy controls the prompts:

$$\mathit{Target}=\{\mathit{Static\ Picture},\ \mathit{Audio},\ \mathit{Video}\} \tag{VI-4}$$

$$\mathit{Robot}=\{\mathit{Head\ turn}+\text{"}\mathit{Look!}\text{"},\ \mathit{Head\ turn}+\mathit{Arm\ pointing}+\text{"}\mathit{Look\ over\ there!}\text{"}\} \tag{VI-5}$$

The *Target* state includes 3 *EF*s, and the *Robot* state includes 2 *RA*s. *p* is used to select *RAs* and *EFs* to form different prompts. The *Gaze tracking* state controls the *Tracking* function only, which represents the gaze tracking module.

Two pure signals "Time out (TO)" and "Target hit" are used to change the prompts and terminate the

LTM-RI trial. A pure signal either absents (no event), or presents (an event happens) any time $t \in \mathbb{R}$ [27]. If the gaze tracking module detects gaze direction towards the target monitor within 7 seconds from the beginning of each prompt, a "Target hit" event is generated, which triggers the state transition to *Reward*. If no target hit is detected, TO event is generated. TO is combined with *p* to guide the transition in the AND substates of *Execution*. Once the state transition is done, *p* is increased by 1 to mark the next level of prompt. If prompt level 6 is completed without "Target hit", the system transits to *Termination*.

6.5 Experimental User Study

Fig. VI.6 shows the experiment room. The participant was seated in a wooden chair. The robot was placed in front of the participant, standing on a platform 32cm above the floor. When the participant was seated, his/her eyes were approximately as high as the robot's face. The two monitors and the robot were all 2 meters away from the participant.

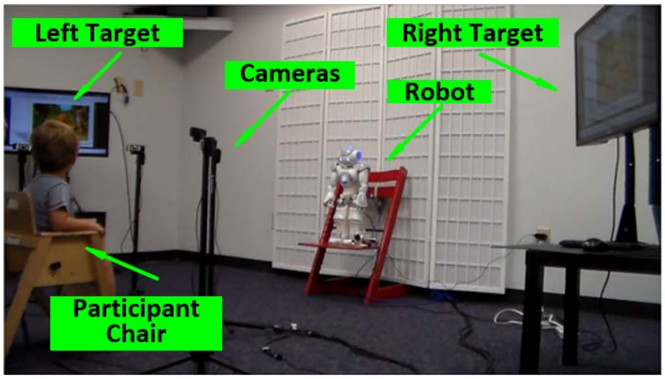


Fig. VI.6 Experiment room configuration

6.5.1 Participants

The Norris was tested by 14 children (12 males, 2 females) with ASD. They were recruited from a research registry of the Vanderbilt Kennedy Center, and this study was approved by the Vanderbilt University Institutional Review Board. The characteristics of the participants are shown in Table VI.2. They had confirmed diagnoses by a clinician based on the DSM [28] criteria. They met the spectrum cut-off on the Autism Diagnostic Observation Schedule [17], and had existing Intelligent Quotient data regarding cognitive abilities in the registry. Parents of these children also completed the Social Responsiveness Scale–Second Edition [29] and Social Communication Questionnaire Lifetime Total Score (SCQ) [30] to index current ASD symptoms.

Table VI.2 Participant Characteristics

	ADOS Raw Score	***IQ***	***SCQ***	***SRS-2 T score***	***Age (Years)***
Avg	21.29	54.71	14.86	63.36	**2.78**
SD	4.61	8.17	5.56	8.63	**0.65**

6.5.2 Experimental procedure and measurements

Four sessions were arranged on different dates for each participant. Each session involved 8 repeated LTM-RI trials as introduced in section III.B. The left or the right monitor was randomly assigned as the target for each trial.

Preferential attention

First, we evaluated the participants' attention on the robot and the monitors, a measure which reflected their engagement. A region of interest was defined for each object that covered this object with a margin of 20cm around it. We analyzed how their attention was distributed among the robot and the 2 monitors. We anticipated that participants would: 1) pay significant attention to the robot because the robot was the main interactive agent; and 2) pay more attention to the target monitor than the non-target monitor, because the target monitor was referred to by the robot and displayed visual stimuli. We also tracked the change in the participants' attention on the robot over the 4 sessions. We anticipated that if the participants' interest in the robot was sustained over the sessions, then their time spent looking at the robot would not change significantly.

Joint attention performance

Second, we evaluated the participants' joint attention performance, which reflected the effectiveness of the system. For each session, we computed: i) the number of trials in which the participants hit the target successfully; and ii) the average prompt levels the participants needed in order to hit the target. We anticipated that the participant's performance would improve significantly if the robotic intervention was effective. In addition, we computed the intensity (defined in equation (2)) of each prompt level within the sessions. If the participants' joint attention skill improved, we would see higher intensity values in low prompt levels than in high prompt levels.

6.5.3 User Study Results

All 14 participants completed the 4 sessions, and thus the completion rate was 100%. This result is very promising when compared with other technology-assisted studies [12, 31]. We used the Wilcoxon-signed rank test for statistical analysis.

Preferential attention

On average, in sessions 1 through 4,the participants spent 54.84%, 52.93%, 47.97%, and 51.58% of the session duration looking at the main objects (i.e., the robot, the target monitor, and the non-target monitor), respectively. Fig. VI.7 shows the percentage of session durations that the participants spent looking at each of the main objects across the sessions. On average, in sessions 1 to 4,the participants spent 28.76%, 29.17%, 28.23%, and 23.69% of the session duration looking at the robot, respectively. In sessions 1 to 4, they spent 19.04%, 18.85%, 14.91%, and 21.31% of the session duration looking at the target monitor, respectively. We can see that the participants looked at the robot more than the monitors in every session. As expected, the participants looked at the target monitor much more than the non-target monitor. The main reasons were: 1) the target monitor was referred to by the robot, and the participants responded more to the referred direction; 2) the target monitor displayed visual stimuli during prompts and rewards, which caught and held the participants' attention. Results showed that the participants spent very small portions of the session duration looking at the non-target monitor (7.05%, 4.92%, 4.83%, and 6.58% in sessions 1 to 4, respectively).

We compared the time that the participants spent looking at the robot across all sessions, and found no significant change (p = .9515 to .1937). This result suggests that the participants' interest in the robot held over the course of the sessions. The change in attention duration on the target monitor was not significant, except between sessions 2 and 3 (p = .0203), and between sessions 3 and 4 (p = .0009). In each session, different sets of static pictures, audio clips and video clips were presented in the prompts. We noticed that the participants had different preferences for certain stimuli (e.g., one participant liked "Scooby Doo" more than "Dora"). Therefore, the fluctuation in attention time on the target monitor might be attributable to the change of stimuli. In addition, the attention time on the target monitor was significantly higher than the non-target monitor (p ranges from .0001 to .0006) in all 4 sessions. In summary, these results indicated: 1) among the three main objects, the participants paid most of their attention to the robot; 2) The participants' initial interest in the robot were maintained across the sessions; 3) They paid significantly more attention to the target monitor than the non-target monitor. Due to the heterogeneous development trajectory and behavior pattern of children with ASD, the participants had quite different attention patterns and joint attention capabilities. Therefore, Fig. VI.7 shows large standard deviations in all cases. The standard deviation of the looking time on the robot decreased from session 1 to session 4, which showed that the participants' looking towards the robot tended to be stable after a few sessions' intervention. However, this pattern was not shown for the two monitors.

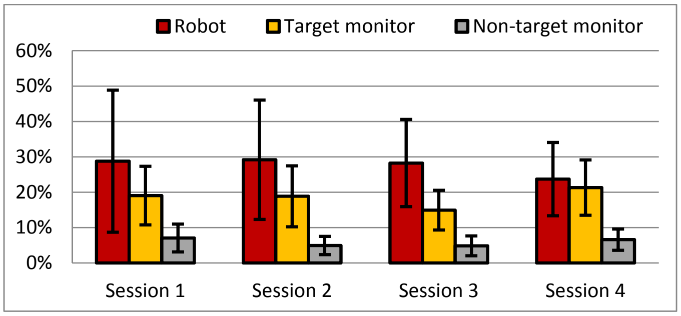


Fig. VI.7 Percentage of the session time participants spent looking at the robot, target monitor, and the non-target monitor

Joint attention performance

Fig. VI.8 shows the average prompt levels participants needed to hit the target. Note that the lower the prompt level needed by participants, the better their performance was. We observe that from session 1 to session 4, the average target hit prompt level decreased from 2.31 to 1.71 monotonously. A Wilcoxon-signed rank test showed that the decrease in prompt level from session 1 to session 4 was significant (p = 0.0115). This indicated that the participants' performance improved significantly.

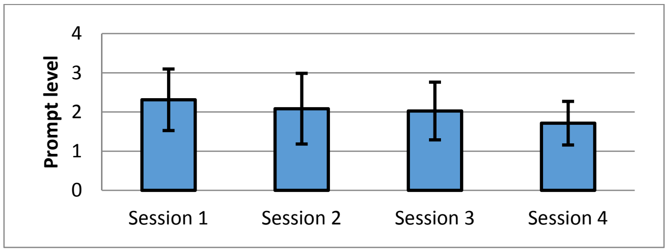


Fig. VI.8 Average target hit prompt levels in all sessions

We further evaluated how the incremented LTM prompt levels elicited target hit behavior, i.e., whether the two goals discussed in section III.A were achieved. The computation can be performed as follows:

$$P\left(ER_n \mid PT_{n-1},\ldots PT_1\right) = \frac{number\ of\ trials\ ended\ with\ target\ hit\ on\ prompt\ level\ n}{total\ number\ of\ trials} \cdot \tag{VI-6}$$

Then the intensity of prompt level *n* can be computed according to equation (2). Here we mark the intensity of prompt level *n* in session *x* as I_n^x . Fig. VI.9 shows the values of I in sessions 1 to 4.

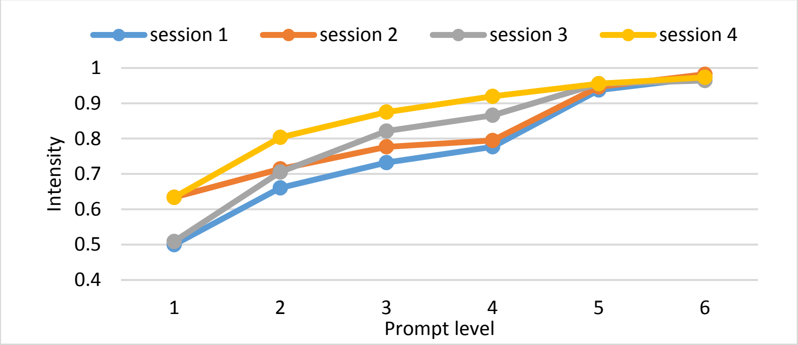


Fig.VI.9 Intensity of prompt levels across 4 sessions

We observe that, $I_a^x > I_b^x$ (given $a>b$) in all 4 sessions. This indicated that the adding more instructive prompt levels on top of low prompt level elicited more target hits. Therefore, **Goal 1 was achieved.** Fig. VI.9 also shows that for the same prompt level, $I_a^x > I_a^y$ or $I_a^x \approx I_a^y$ (given $x>y$) in most of the cases. The only exception is that I_1^2 is apparently higher than I_1^3 . This means that, in general, as the participants had received more interventions in later sessions, their chances of a target hit on the same prompt level increased with occasional fluctuations.

I_6^1 to I_6^4 were 0.97, 0.98, 0.96, and 0.97, respectively. This leads us to conclude that the LTM-RI trial could eventually help participants hit the target in almost all of the trials. Thus the prompt level content and the number of the levels (IN = 6) were properly designed. Therefore, **Goal 2 was also achieved.**

6.6 Discussion and Conclusion

In this chapter, we introduced a new non-invasive autonomous robot-mediated joint attention intervention system, Norris. A humanoid robot was embedded as the intervention administrator. The looking behavior of the participants in response to robot prompts was detected using a new non-contact gaze tracking method, which could track the participants' real-time gaze direction in a large range. The prompts were designed and implemented based on the presented Least-to-Most Robot-mediated

Intervention prompting hierarchy, LTM-RI, which is a general model of robot-mediated intervention for children with ASD.

Norris was validated through a 4-session longitudinal study. Fourteen children with ASD were recruited and all of them successfully participated in all the sessions. We measured their preferential attention towards the robot, target monitor, and the non-target monitor. Results showed that the participants looked at the robot longer than other objects and this interest did not change significantly over the sessions. As expected, the participants paid significantly more attention towards the target monitor than the non-target monitor in all the sessions. We also evaluated their joint attention performance. Results showed that the participants' performance improved significantly after the 4 intervention sessions. The results also proved the effectiveness of the LTM-RI model: i.e., the higher the prompt level, the higher the probability that target hit was achieved by the participants, and the participants could hit the target eventually in almost all the trails.

Therefore, we conclude that this study has three major contributions: 1) the design and development of a new joint attention intervention system; 2) the introduction of the LTM-RI model and a successful instantiation of LTM-RI in a joint attention study; and 3) a longitudinal user study which validated the effectiveness of Norris and LTM-RI.

However, it is important to notice that the current study also had limitations that need to be addressed in the future. First, Norris was only validated by a small group of children with ASD. In order to thoroughly evaluate the efficacy of Norris, it needs to be tested in formal clinical studies in the future. Second, while we did assess promising joint attention skills within the system, we did not systematically compare such improvements with other methods nor did we see if such training can be generalized to other interactions (e.g., human-human interaction). Third, the current study repeated a straightforward LTM-RI procedure in a limited number of sessions. However, using the same interaction content repeatedly will eventually cause the ceiling effect (e.g., participants hit the target on the first prompt in most of the trials) and/or loss of interest (e.g., participants feeling tired of doing the same intervention again and again) after a large number of sessions. Therefore, once these issues are detected, new interaction content under the same interaction protocol or a completely new interaction protocol need to be adapted by the system to update and reinforce the training procedure. Finally, LTM-RI was proposed as a general interaction model to implement robot-mediated intervention for children with ASD. Although the current study successfully validated LTM-RI for joint attention, we did not test it thoroughly in other types of training. Therefore, the eventual value of LTM-RI will need to be verified through other interventions.

Despite these limitations, this work is the first to our knowledge to design and empirically evaluate the longitudinal usability and feasibility of a non-invasive autonomous robot-mediated joint attention intervention system. The preliminary results of this work are promising. Note that the Norris system architecture, components, and the LTM-RI protocol can be adapted to address other core deficits in ASD (e.g., social orienting, response to name). Thus this work provided an example of how to design and implement an effective robot-mediated intervention system in general. It is important to note that we do not propose this technology as a replacement for existing necessary comprehensive behavioral intervention and care for young children with ASD. Instead, this platform represents a meaningful step towards realistic deployment of technology capable of accelerating and priming a child for learning in key areas of deficits.

6.7 References

[1] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon," *International Journal of Social Robotics,* vol. 1, no. 1, pp. 3-18, 2009.

[2] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of autism and developmental disorders,* vol. 43, no. 5, pp. 1038-1049, 2013.

[3] J. Greczek, E. Kaszubksi, A. Atrash, and M. J. Matarić, "Graded Cueing Feedback in Robot-Mediated Imitation Practice for Children with Autism Spectrum Disorders," *Proceedings, 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014) Edinburgh, Scotland, UK* Aug. 2014.

[4] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "KASPAR–a minimally expressive humanoid robot for human–robot interaction research," *Applied Bionics and Biomechanics,* vol. 6, no. 3-4, pp. 369-397, 2009.

[5] H. Kozima, C. Nakagawa, and Y. Yasuda, "Children–robot interaction: a pilot study in autism therapy," *Progress in Brain Research,* vol. 164, pp. 385-400, 2007.

[6] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism," *Autonomous Mental Development, IEEE Transactions on,* vol. 6, no. 3, pp. 183-199, 2014.

[7] B. Robins, and K. Dautenhahn, "Developing play scenarios for tactile interaction with a humanoid robot: a case study exploration with children with autism," *Social Robotics*, pp. 243-252: Springer, 2010.

[8] D. Feil-Seifer, and M. J. Matarić, "Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders." pp. 201-210, 2009.

[9] Z. E. Warren, and W. L. Stone, "Best practices: Early diagnosis and psychological assessment," *Autism Spectrum Disorders*, David Amaral, Daniel Geschwind and G. Dawson, eds., pp. 1271-1282, New York: Oxford University Press, 2011.

[10] P. Mundy, Block, J., Delgado, C., Pomares, Y., Van Hecke, A.V., Parlade, M.V., "Individual differences and the development of joint attention in infancy," *Child development,* vol. 78, no. 3, pp. 938-954, 2007.

[11] C. Kasari, A. C. Gulsrud, C. Wong, S. Kwon, and J. Locke, "Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism," *Journal of autism and developmental disorders,* vol. 40, no. 9, pp. 1045-1056, 2010.

[12] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar, “A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism,” *Neural Systems and Rehabilitation Engineering, IEEE Transactions on,* vol. 21, no. 2, pp. 289-299, 2013.

[13] Z. Zheng, L. Zhang, E. Bekele, A. Swanson, J. Crittendon, Z. Warren, and N. Sarkar, "Impact of Robot-mediated Interaction System on Joint Attention Skills for Children with Autism ".

[14] S. R. Leekam, C. Nieto, S. J. Libby, L. Wing, and J. Gould, “Describing the sensory abnormalities of children and adults with autism,” *Journal of autism and developmental disorders,* vol. 37, no. 5, pp. 894-910, 2007.

[15] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: simulating the human for interaction research." pp. 101-107.

[16] M. E. Libby12, J. S. Weiss, S. Bancroft12, and W. H. Ahearn12, “A comparison of most-to-least and least-to-most prompting on the acquisition of solitary play skills,” 2008.

[17] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[18] P. Mundy, Hogan, A., & Doelring, P, *A preliminary manual for the abridged Early Social Communication Scales (ESCS)*, Coral Gables, FL: University of Miami, 1996.

[19] "Aldebaran Robotics," http://www.aldebaran-robotics.com/en/.

[20] Z. Zheng, Q. Fu, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Design of a Computer-Assisted System for Teaching Attentional Skills to Toddlers with ASD," *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being*, pp. 721-730: Springer, 2015.

[21] X. Xiong, and F. De la Torre, “Supervised Descent Method for Solving Nonlinear Least Squares Problems in Computer Vision,” *arXiv preprint arXiv:1405.0601*, 2014.

[22] D. J. Feil-Seifer, and M. J. Matarić, "A Simon-Says Robot Providing Autonomous Imitation Feedback Using Graded Cueing," *International Meeting for Autism Research (IMFAR)*, 2012.

[23] B. Huskens, R. Verschuur, J. Gillesen, R. Didden, and E. Barakova, “Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention,” *Developmental neurorehabilitation,* vol. 16, no. 5, pp. 345-356, 2013.

[24] Z. Zheng, E. Young, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, “Robot-mediated Imitation Skill Training for Children with Autism,” *IEEE transactions on neural systems and rehabilitation engineering*, 2015.

[25] M.-G. Kim, I. Oosterling, T. Lourens, W. Staal, J. Buitelaar, J. Glennon, I. Smeekens, and E. Barakova, "Designing robot-assisted Pivotal Response Training in game activity for children with autism." pp. 1101-1106.

[26] D. Harel, “Statecharts: A visual formalism for complex systems,” *Science of computer programming,* vol. 8, no. 3, pp. 231-274, 1987.

[27] E. A. Lee, and S. A. Seshia, *Introduction to embedded systems: A cyber-physical systems approach*: Lee & Seshia, 2011.

[28] A. P. Association, *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*: bookpointUS, 2013.

[29] J. N. Constantino, and C. P. Gruber, "The social responsiveness scale," Los Angeles: Western Psychological Services, 2002.

[30] M. Rutter, A. Bailey, and C. Lord, "The Social Communication Questionnaire," Los Angeles, CA: Western Psychological Services, 2010.

[31] J. Wainer, K. Dautenhahn, B. Robins, and F. 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,* vol. 6, no. 1, pp. 45-65, 2014.

Chapter VII. Exploration of the Generalization from Robot-mediated Joint Attention Intervention to Human-Human Interaction

7.1 Abstract

The initial results in Chapter VI are promising. However, based on the methods employed it remains unclear whether the robot-mediated intervention has a meaningful positive impact on children's interactions with important social partners in their daily life (e.g., parents, interventionists, etc.). Ultimately, this is a fundamental question that needs to be answered in order to assess the true value of intelligent robot-mediated interventions. In this chapter, we explored whether the robot-mediated intervention provided by the Norris system could help improve the performance of children with ASD in human-human interaction. In other words, we investigated whether the joint attention skills targeted within the Norris system would generalized to other social communication skills with human partners. We conducted a pilot randomized control study with 11 children with ASD. They were randomized into an immediate participation group or a waitlist control group. Their performance in both human-robot interaction and human-human interaction was assessed. Our data suggest that participants demonstrated an increased ability to display important social communication skills in human-human interaction after participating in the brief robot-mediated joint attention intervention. In addition, participants who improved in human-robot interaction gained more improvement in human-human interaction, compared with the participants who demonstrated limited progress within the human-robot interaction.

keywords—Human-robot interaction, human-human interaction, joint attention, skill generalization

7.2 Introduction

The robot-mediated interventions discussed in Chapter VI achieved promising results. One of the most important findings was that the robotic system could hold the participants' attention across multiple sessions. This result provides the foundation for conducting a more rigorously controlled longitudinal study to further investigate the broad impact of an autonomous robot-mediated joint attention intervention system. The previous pilot studies had three important insufficiencies that needed to be addressed in this chapter:

1) The performance of the participants who experienced the robot-mediated intervention was not compared to a control group (i.e., all participants received intervention).

2) We did not examine whether the robot-mediated intervention had an impact outside the technological environment. In particular, it was not clear whether the within-system training could help the participants in real-world social communication skills with human partners.

3) The same interaction procedure was repeated in each of the 4 sessions, and thus it was not clear whether the participants would tolerate a more complex interaction protocol (i.e., a more complex interaction logic in later sessions after they were exposed to a simpler skill in in earlier sessions).

In order to address these limitations, we designed a small randomized controlled pilot study. The participants were randomized into an immediate participation (IP) group or a waitlist control (WLC) group. The participants in the IP group were assessed before and after 4 sessions of robot-mediated interventions with Norris, while the participants in the WLC group were assessed immediately before and after a waiting period. The assessments evaluated participant performance in both human-robot interaction (HRI) and in human-human interaction (HHI). After the waiting period, participants in the WLC group also had 4 sessions of robot-mediated intervention, followed by another assessment. This allowed us to conduct additional analyses in a larger sample combining the performance of all participants in the intervention procedures.

Initial results from 11 young children with ASD showed that the participants who experienced the robot-mediated intervention performed better in HHI than those who did not experience the robot-mediated intervention. Specifically, data suggest that participants demonstrated an increased ability to display important social communication skills in human-human interaction after participating in our brief robot-mediated joint attention intervention, with the robot-mediated intervention potentially improving interactive engagement across settings. In support of this idea was the finding that more than half of the participants' HRI performance improved after the robot-mediated intervention, with these children demonstrating improvement in HHI than those whose HRI performance did not improve.

The remainder of this chapter is organized as follows: Section 7.3 describes the experimental setup of the longitudinal user study; Section 7.4 presents the results of this user study; and finally, Section 7.5 concluded this chapter and highlights future research directions.

7.3 Experimental Setup

7.3.1 Participants

Initially 13 children with ASD under 36 months of age were recruited from a repository of Vanderbilt Kennedy Center. This study was approved by the Vanderbilt University Institutional Review Board. All participants had confirmed diagnoses of ASD by a clinician based on DSM-5 [1] criteria. They met the spectrum cut-off on the Autism Diagnostic Observation Schedule [2], and had existing Intelligence Quotient (IQ) data regarding cognitive abilities in the registry. Parents of these children also completed the Social Responsiveness Scale–Second Edition (SRS-2) [3] and Social Communication Questionnaire

Lifetime Total Score (SCQ) [4] to index current ASD symptoms. As opposed to previous work where children participated several months on average from their initial diagnosis, in this study we specifically attempted to target children soon after their initial medical diagnosis to mimic real-world use of this intervention as a 'priming' or 'accelerant' technology deployed while children wait for other services. This study used a randomized waitlist control design, where participants were randomized to a waitlist control (WLC) group or immediate participation (IP) group. Among the 13 participants, 7 were randomized to the IP group, and 6 were randomized to the WLC group. Two children in the WLC group dropped out. The characteristics of these children are listed in Table VII.1.

Table VII.1 Characteristics of Participants

		ADOS total raw score	***SRS-2 total raw score***	***SRS-2 Tscore***	***SCQ Current Total Score***	***IQ***	***Age***
IP	***Mean***	19.50	105.43	74.43	19.43	65.00	2.56
	SD	6.50	22.97	8.85	3.95	17.29	0.53
WLP	***Mean***	20.67	110.75	76.75	20.50	57.17	2.72
	SD	3.08	17.54	6.65	2.52	7.70	0.30
All	***Mean***	20.08	107.36	75.27	19.82	61.08	2.63
	SD	4.89	20.40	7.85	3.40	13.40	0.43

7.3.2 Interactions

Both human-robot interaction (HRI) and human-human interactions (HHI) were involved in this study, the structures these two interactions are as follows:

Human-Robot Interaction (HRI)

The HRI was provided using the Norris system introduced in Chapter VI. Here we adopted the original LTM intervention protocol discussed in Chapter VI as the single-target human-robot interaction (ST-HRI), where the robotic system prompted to one target per trial. There were 8 trials arranged in this procedure.

In addition, we introduced a double-target human-robot interaction (DT-HRI) to provide a new interaction experience for the participants with good performance in the ST-HRI procedure. The DT-HRI was added to help prevent potential ceiling effects or loss of participant interest in the longitudinal repeated robot-mediated interventions. In a DT-HRI trial, the robot spent approximately 5 minutes instructing participants to look at two targets in different directions. Trials were repeated until the total duration of the interaction reached about 5 mins. This strategy avoided a sudden termination of the intervention that might surprise the participants. Each trial had 2 levels of prompts. Starting with Prompt 1, the robot first turned its head to one target monitor and said, “First look at that!”, and then turned to the opposite monitor and said, “Then look over there!” If the participant did not succeed, Prompt 2 was given. In Prompt 2, the robot turned its head to the target, pointed at the target with its arm, and said, “First look at that!” Then the target monitor displayed a short video for 2 seconds to help attract the participant’s attention. Following that, the robot prompted to the opposite target with the same type of motion, saying, “Then look over there!” Finally, the second target monitor displayed 2 seconds of video. Prompts 1 and 2 had response windows of 8 and 10 seconds starting from the beginning of the prompt, respectively. On any prompt level, if the participant looked at both the targets within the response window following the robot’s instruction, the robot would say, “Good job!”, and the second target would display a short reward video. This DT-HRI procedure was initially tested by 6 children with ASD for its feasibility [5] and the results showed that it was well tolerated by them.

Human-Human Interaction (HHI)

The HHI represents the administration and scoring of the Screening Tool for Autism in Toddlers and Young Children (STAT) [6, 7]. STAT is a level-2 screening instrument normed on children aged 24–35 months with extended scoring systems for 3-years-olds and children as young as 14 months [8]. This interactive assessment takes 15-20 min, where an examiner interacts with the participant to test specific social communication skills. The examiner rates the participant’s performance during the interaction with formal domain scores related to key social communication skills, i.e., play, requesting, direction attention, and imitation. This assessment examines the participants’ capability of interacting with the examiner, being able to imitate and respond to what they are doing, and to direct and initiate joint attention. These domains are considered core social communication skill domains overlapping and related to the robot-mediated interaction, but not the exact same skill taught within the procedure (i.e., test of generalization of skills to person, context, and task). The sum of domain scores in STAT ranges from 0-4 with lower scores indicating more social communication skills within the HHI.

7.3.3 Experimental protocol

The experiment in this study consisted of repeated assessments across the intervention period. Each assessment session (marked by A) included the ST-HRI and the STAT to evaluate the participants’ performance in HRI and HHI. The robot-mediated intervention sessions (marked by S) included both ST-HRI and DR-HRI.

As shown in Fig. VII.1, children in the IP group received assessments before (A1) and after (A2) 4 robot-mediated intervention sessions (S1 to S4). S1 and S2 only included the ST-HRI. S3 and S4 consisted of part A and part B. Part A was the ST-HRI, and if the performance of the participant in part A satisfied one of the following criteria, this participant went on to the DT-HRI as part B:

1) The participant hit the target on prompt levels 1 to 2 in 4 or more than 4 trials in a row;

2) The participant hit the target on prompt levels 1 to 4 in all 8 trials.

Otherwise, the participant continued with another ST-HRI as part B. The duration allowed between A1 and A2 was 3-9 weeks.

The WLC group received their first assessment, marked as A0, before a waiting period of 3-9 weeks. After this waiting period, the participants were assessed for the second time, marked as A1. Following A1, the WLC group experienced 4 sessions of robot-mediated interventions as the IP group. Then, they also had the final assessment A2.

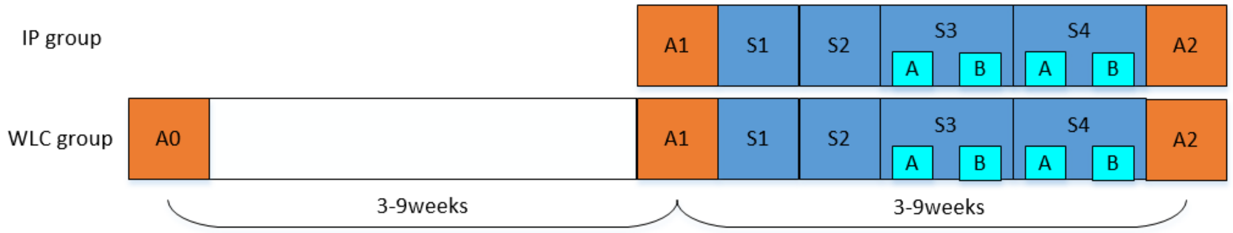


Fig. VII.1 Experiment procedure

7.3.4 *Measurements*

In this study, we evaluated the participant's performance in both HRI and HHI. The HRI performance was measured by 3 variables:

1) Average prompt level that the participants needed to hit a target: the lower the prompt level needed, the higher the performance;

2) Target hit rate: the percentage of trials where participants eventually hit a target, regardless of which prompt level was needed for the target hit. Higher values mean better performance;

3) Trial length: once a target hit happened, the interaction trial was terminated, and thus the shorter a trial, the better the performance.

In addition, we measured the attention of the participants on the robot. We defined "attention" as the percentage of the session duration that the participants spent looking at the robot. We used percentage instead of the actual duration because the length of the trials depended on the participants' performance and was not consistent.

The participants' HHI performance was evaluated by two variables:

1) In some cases, children with severe ASD symptoms were not able to sit in the assessment room and interact with the STAT examiner. Therefore, we first calculated the percentage of the participants who could participate in the STAT interaction.

2) The STAT score of the participants. The participants who could participate in STAT received rated scores from the examiner. The lower the STAT score, the lower the level of ASD symptoms shown and the better the HHI performance indicated. For the participants who could not do the STAT, we created the pseudo total score of 4 (the maximum STAT score, indicating the highest level of ASD symptoms), to facilitate numerical evaluation of the results, indicating all failed in the test protocol.

Given the aforementioned measurements, we conducted two comparisons:

First, we compared the participants who experienced the robot-mediated intervention (the IP group) and those who did not (the WLC group before they received the intervention). Here we compared the IP group's performance in A1 and A2 with the WLC group's performance in A0 and A1. The former set reflects the change in performance due to the robot-mediated intervention, and the later set reflects the changed in performance after a waiting period of similar duration as the robot-mediated intervention. This helped us differentiate the effects of the intervention (IP) from the effects of simple maturation as time passed (WLC).

Second, we compared the participants' performance before and after the robot-mediated intervention. Since the WLC group also experienced the same amount of robot-mediated intervention after the waiting period, we combined the IP and WLC groups here, and compared their performance between A1 and A2. In this way, we had a larger sample size than that of the IP group alone.

We anticipated that if the robot-mediated intervention could help the participants learn social communication skills, then their HHI performance would improve after the intervention, and this improvement should be larger than those on the waitlist who did not receive the intervention. In addition, since we introduced the DT-HRI as a more complex layer of HRI, we examined the participants' tolerance and performance with this new protocol, which may provide an option for a future update of the robot-mediated intervention. Due to the heterogeneous behavioral pattern and large individual differences among children with ASD, we also investigated the difference between the participants whose HRI performance improved and those whose did not improve after the robot-mediated intervention. This helps reveal the features of the participants who may benefit from the robot-mediated intervention more than others (i.e., we hypothesized some children would tolerate and benefit more from system than others).

7.4 Experimental Results

Two out of 6 participants in the WLC group dropped out. One child could not tolerate the HRI. Another child tolerated both the HRI and the HHI well in A0, but the family could not attend the following experiments due to time conflicts. Therefore, the overall participation rate was 84.62%.

7.4.1 Robot-mediated intervention vs No robot-mediated intervention

For this comparison we examined the performance of the IP group before (A1) and after (A2) the robot-mediated intervention with the performance of the WLC group before (A0) and after (A1) the waiting period.

As shown in Table VII.2, the HRI performance of the IP group decreased slightly (decreased hit rate, increased prompt level) from A1 and A2. Between A0 and A1, the WLC group improved in prompt level while also decreasing in target hit rate. Both groups showed increased trial lengths as well as more attention on the robot.

One of the 7 participants in the IP group could not tolerate HHI in A1, with all tolerating the HHI in A2. In addition, 85.71% of the participants in this group had better HHI performance in A2 than in A1. As shown in Table VII.2, their HHI performance improved 0.5 points. STAT is scored from 0-4 in increments of 0.25-.5, depending upon the domain. Therefore, this improvement showed that the HHI performance of the IP group improved by a clinically meaningful level on the STAT metric. In both A0 and A1, 50% of the participants in the WLC group could not tolerate HHI, and only one child's (25% of the group) HHI performance improved from A0 to A1. Overall, the WLC group's HHI performance improved by 0.25 (See Table VII.2).

In summary, the HRI performance of both groups decreased slightly. However, compared with the WLC group, the IP group achieved greater improvement in HHI. This result suggests that the robot-mediated intervention increased the ability of children to display important social communication skills in human-human interactions. We also ran 2-way ANOVA analyses with the factors of assessment time and group. Given the small sample size, we did not observe any statistically significant differences on the results listed in Table VII.2.

Table VII.2 The IP group's performance before and after the robot-mediated intervention vs the WLC groups' performance before and after the waiting period

Group			***IP***		***WLC***	
Assessments			***A1***	***A2***	***A0***	***A1***
HRI performance	***Prompt level***	***Mean***	2.38	2.83	2.41	2.14
		std	1.00	1.15	0.58	0.54
	Target hit rate	***Mean***	88%	86%	97%	88%
		std	14%	13%	6%	14%
	Trial length (s)	***Mean***	36.35	38.28	32.97	34.96
		std	7.14	8.33	4.98	6.19
Attention on robot		***Mean***	0.25	0.27	0.13	0.18
		Std	0.12	0.12	0.05	0.10
HHI performance (STAT score)		***Mean***	3.39	2.89	3.5	3.25
		Std	0.93	1.35	1	0.96

7.4.2 Pre-post comparison of robot-mediated intervention

We conducted a combined evaluation of the change of the participants' performance before and after the robot-mediated intervention utilizing both the IP and WLC groups. Since the WLC group also finished the intervention after the waiting period, they were combined with the IP group in this data analysis. From Table VII.3, we can see that the participants' HRI performance decreased slightly, while their attention on the robot increased slightly. When examining HHI performance, we found that 54.55%, 36.36%, and 9.09% of the participants' HHI scores improved, did not change, or worsened. Overall, participants' HHI performance improved by 0.52 as shown in Table VII.3. As stated before, STAT was scored in increments of 0.25-.50, and thus this improvement showed that the HHI performance was improved in a clinically meaningful increment based on the STAT scoring method. We ran 1-way ANOVA analyses on the results shown in Table VII.3. Given the small sample size, we did not observe any statistically significant differences between A1 and A2.

In S3 and S4, the participants participated in novel HRI interactions introducing a higher order of complexity (see section 7.3). Results indicate that in S3, all participants finished Part B, and 27.27% of the participants finished the DT-HRI. In S4, 81.82% of the participants finished Part B, and 36.36% of them completed the DT-HRI. This result highlights both the heterogeneity of ASD performance within system and that a subgroup of children with ASD may accept longer interactions (tolerated both part A and part B) with more complex interaction scenarios (e.g., DT-HRI) than the ST-HRI.

Table VII.3 The IP and WLC groups' performance before and after the robot-mediated intervention

Group			***IP + WLC***	
Assessments			***A1***	***A2***
HRI performance	***Prompt level***	***Mean***	2.29	2.76
		std	0.84	1.02
	Target hit rate	***Mean***	0.88	0.80
		std	0.14	0.19
	Trial length (s)	***Mean***	35.84	39.29
		std	6.53	7.66
Attention on robot		***Mean***	0.22	0.25
		Std	0.11	0.15
HHI performance (STAT score)		***Mean***	3.34	2.82
		Std	0.90	1.22

We further investigated the ADOS and the IQ scores of the participants who improved in the HRI and HHI. Higher ADOS scores reflect more severe ASD symptoms, and higher IQ reflects higher cognitive functioning. As shown in Table VII.4, we observed that children with higher ADOS and lower IQ scores showed better HRI performance. These children had more severe ASD symptoms and lower cognitive skills. Children who improved in HHI had lower ADOS score (indicates less severe ASD symptoms) and higher IQ (indicates higher cognitive functioning). Children who had fewer ASD symptoms and had higher intelligence showed more improvements in social interactions with another human. At present, our small sample size prevents us from making concrete conclusions about these findings. One hypothesis for the improved HRI performance in children with more significant ASD is that children with higher ASD symptoms may have a stronger preference for technology, such as robots. Regardless, these findings inspire us to track how ADOS and IQ scores may help us identify which subgroups of children with ASD could benefit more from robot-mediated intervention in future work.

Finally, we calculated the HHI performance of the participants whose HRI improved (at least one HRI variable improved) versus these who did not improve after the robot-mediated intervention. Results showed that 7 participants improved in HRI, and their HHI performance improved by 0.64 from A1 to A2. In contrast, the participants who did not improve in HRI only improve by 0.31 in HHI (less than a half of the HRI improved group). In other words, participants who showed larger HRI improvement also improved more in HHI. This may indicate that the robot-mediated intervention helped the participants learn social communication skills that can be applied well in HHI.

Table VII.4 Average ADOS and IQ scores of the participants whose HRI and HHI performance improved and not improved after the robot-mediated intervention

		Improved?	***Number of participants (IP + WLC)***	***ADOS***	***IQ***
HRI performance	***Prompt level***	Yes	3	21.33	60.67
		No	8	20.14	60.71
	Target hit rate	Yes	5	24	50
		No	6	17	71.40
	Trial length	Yes	4	22.75	57.75
		No	7	19	62.67
HHI performance (STAT score)		Yes	6	19	65.5
		No	5	22.75	53.5

7.5 Discussion and Conclusion

One of the most important results we observed in this study was that HHI performance improved after the robot-mediated intervention both relative to WLC and within our combined sample. The reason underlying this improvement could be that the intelligent robotic system exposed children to scenario that are meaningful within HHI, including consistently responding to the direction where the child was looking and reinforcing the procedure of responding to joint attention. The skills of engaging in communication with another agent and responding to a target following an administrator's instruction were practiced with the system, and these are skills likely deployed in HHI. This linkage between HRI and HHI has an important meaning of conducting studies beyond the technology development. It is important to track the change of the performance within system. But ultimately, evaluation of the improvement outside HRI, in other words, whether the skills learnt in HRI can be generalized in HHI is the most important aspect

Regarding the HRI performance, a subgroup of participants demonstrated more improvement post-intervention. At the group level, we observed a slight, nonsignificant decrease in the HRI performance. The specific skill focus of the HRI and quantitative approach to understanding success was entirely defined by discrete event performance. However the benefit in HHI for those participating in the HRI suggests the impact of the HRI system may not be around such skills, but a much broader ability to engage and participate in meaningful aspects of social communication and learning. We are not expecting for the HRI to replace existing clinical interventions, but we hope that it can provide children with ASD with a technological boost to accelerate their skill learning, either while waiting for human intervention or as a supplement to such. Specifically, interacting with technology may prime skills in the interaction with other people in their life. When psychological/clinical interventions are limited or not immediately accessible, for example, when the participants are on the waitlist for a clinical appointment, they can use technologies that provide consistent and engaging training to facilitate their learning, so that they will be better prepared when seeing professionals.

Another interesting result we observed in the robot-mediated intervention was that S3 and S4 showed the variability among the participants. Some of the participants could tolerate and engage in longer interactions while some of them preferred a short interaction. Importantly, we were able to observe some children who did well in the additional layer of complexity (DT-HRI). In the future, these variability could be used to prevent ceiling effects and keep participant's attention in other longitudinal HRI studies.

The user study discussed in Chapter VI showed that the participants' HRI performance improved significantly, which differs from the current study. However, the results in these different studies are not directly comparable due to several different factors. First, the participants in the previous study had a better baseline performance (target hit rate = 97% in the first session) than the current participant group (target hit rate = 88% in A1). Secondly, the length of time that the participants were allowed to finish the entire course of intervention sessions was not constrained in the previous study but was strictly scheduled in the current work. Thus some of the participants in the previous study finished 4 sessions in a few days and some others finished in a few weeks, which might impact the results. In addition, the participants in the previous study were diagnosed with ASD about 8 months (on average) before the study, while the participants in current study were diagnosed with ASD about 4 months before the experiments. Therefore, the participants in the previous study might have received more interventions between the diagnosis and attending the robot mediated sessions that could also impact the experimental results.

In summary, we found that very young children with significant ASD impairment in the wake of diagnosis, which is a very stressful period for both the children and their families, could successfully engage in a longitudinal robot-mediated intervention and displayed improvements in important social communication skills after the intervention. This study gained interest from the families of children with ASD in local community, who came to the lab and participated voluntarily in this long term study. Even though we did not observe significant change regarding the HRI performance in current data, these data are interesting enough for us to continue with the study. Ultimately, by conducting longitudinal user studies with fine-tuned machine-assisted intervention systems and reliable psychological evaluations, it will be clearer that how HRI impacts children with ASD in their daily interactions. Although we have not gotten conclusive result due to the limited data sample and preliminary experimental setup, this study shows great potential to continue in the future. In other words, this study tested a model of potential long term implication of technology on young children with ASD.

7.6 References

[1] A. P. A, *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*: bookpointUS, 2013.

[2] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," Western Psychological Services: Torrance, CA, 2012.

[3] J. N. Constantino, and C. P. Gruber, "The social responsiveness scale," Los Angeles: Western Psychological Services, 2002.

[4] M. Rutter, A. Bailey, and C. Lord, "The Social Communication Questionnaire," Los Angeles, CA: Western Psychological Services, 2010.

[5] Z. Zheng, G. Nie, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Longitudinal Impact of Autonomous Robot-mediated Joint Attention Intervention for Young Children with ASD."

[6] W. L. Stone, E. E. Coonrod, and O. Y. Ousley, “Brief report: screening tool for autism in two-year-olds (STAT): development and preliminary data,” *Journal of autism and developmental disorders,* vol. 30, no. 6, pp. 607-612, 2000.

[7] W. L. Stone, E. E. Coonrod, L. M. Turner, and S. L. Pozdol, “Psychometric properties of the STAT for early autism screening,” *Journal of autism and developmental disorders,* vol. 34, no. 6, pp. 691-701, 2004.

[8] W. L. Stone, C. R. McMahon, and L. M. Henderson, “Use of the Screening Tool for Autism in Two-Year-Olds (STAT) for children under 24 months An exploratory study,” *Autism,* vol. 12, no. 5, pp. 557-573, 2008.

Chapter VIII. Contributions and Future work

8.1 Overall Contributions

Human machine interaction (HMI) has become increasingly important in psychological research. This dissertation describes my research on the design and development of novel HMI systems for their potential use as intervention tools for children with ASD. There is a pressing need for effective treatments that will substantially impact the neurodevelopmental trajectories of young children with ASD. Although important investigations by other researchers in this area have shown immense potential for HMI in ASD intervention, there still exists several gaps that this dissertation attempts to address. First, most existing technical intervention systems for children with ASD do not target the core deficits of ASD, which may limit their impact. In addition, many of these systems need to be controlled by a human operator; this requires extra resources and training, in addition to limiting the precision and the speed of system response. Furthermore, many existing HMI systems require attaching physical sensors to subjects in order to detect interaction cues, such as gestures or gaze. Many young children (particularly those with ASD) cannot tolerate these sensors and are thus excluded from being helped by these systems. Finally, the generalization of skills learnt in human-robot interaction to human-human interaction, which reflects the ultimate value of technology-assisted intervention, has not yet been studied. The research presented in this dissertation was designed to address these problems. Specifically, the main contributions of this work are: 1) design and development intervention systems that are oriented to core deficits of ASD; 2) design and development of fully autonomous intervention systems that are able to adaptively adjust system behaviors based on the children's response in real-time; 3) autonomous detection of interaction cues in children with ASD using non-invasive methods; 4) designing and conducting pilot user studies to demonstrate the usefulness of these HMI systems on typically developing children and children with ASD in machine-assisted interventions, and 5) conducting a controlled user study to investigate whether the skills learnt in robot-mediated intervention can be generalized to human-human interaction. These contributions can be categorized into technical contributions and scientific contributions, as stated in the following two sections. Note that these contributions are closely related and thus cannot be separated completely. We summarize them in different sections for the sake of clear presentation and easy understanding. All my publications during this dissertation research is listed in the Appendix.

8.2 Technical Contributions

8.2.1 System architecture and interaction logic design

My first set of technical contributions lies in the design of automated intervention system architectures and the corresponding interaction logic for ASD intervention.

System architecture

System architecture is the global picture of a system, and the foundation of its functionality. Most of the earlier robot-mediated intervention systems were mostly open-loop systems that did not have mechanisms to provide adaptive feedback to children regarding their real-time performance. A few of them provided feedback to the participants using manual operation, which was slow and required experienced operators. Such limitations significantly impacted the efficacy and application of these systems. One of the main reasons for these limitations was the lack of autonomous interaction cue detection methods, which are the key components of designing autonomous intervention system. Consequently, the architectures of the previous systems were not embedded with autonomous interaction cue detection modules, and thus could not produce a fully autonomous intervention system. Additionally, it was also not clear how to coordinate different system components to form closed-loop interaction for intervention.

Therefore, in this dissertation research, one of the most important tasks in building the intervention systems was to design proper system architectures for closed-loop autonomous interactions. As shown in Chapter II, III, VI, and VI, we have designed several new autonomous intervention architectures, RISIA1, RISIA2, ASOTS, and Norris. These architectures include a supervisory/centralized controller as the "brain" of the system. This controller mediates the communication of different system components, and control the global interaction logic accordingly. All these architectures have autonomous interaction cue detection modules (e.g., gesture recognition methods in RISIA1 and RISIA2, as well as gaze tracking methods in ASOTS and Norris) as the "eyes" of the system that sense behavior of the participants. The sensed behavioral information was sent to executive modules (e.g., the robot module in RISIA1, RISIA2, and Norris, and the display subsystem in the ASOTS) to present adaptive prompts to the participants (e.g., robot motions/speeches, and stimuli display). Given these prompts, the participants would change their behaviors and this information was sensed by the interaction cue detection modules again. Thus interaction loops were built within the systems. In addition, we provided formal mathematical modeling that bridged the abstract architectures and the detail implementation of the systems (e.g., the Harel Statechart models in Chapter IV and Chapter VI). Although each intervention system has its own architecture, as explained in these chapters, these architectures are not limited to a certain intervention, but can be easily adjusted for training of other skills.

Interaction protocol design

In this dissertation, the interaction protocol indicates the global interaction logic applied to an intervention system. It describes the procedure of intervention and decision from the system during the interventions. Most of the previous studies simply applied free-play type of interactions, which were not particularly oriented to the core deficits of ASD, and thus did not provide targeted trainings. However, psychologists have argued that intensive interventions targeted to a particular impairment of ASD that symmetrically scaffold the learning of children can achieve the best results. While psychological and pedagogical intervention/teaching methodologies exist, few studies have shown how to implement them in a technology-assisted system. More specifically, the main difficulty is how to structure the behavior of the executive modules following appropriate educational/training principles. Therefore, in this dissertation study, we emphasize on the choice and implementation of proper intervention protocols.

Mostly, we applied the Least-to-Most (LTM) prompting hierarchy for intervention. In a LTM hierarchy, the intervention administrator allows the learner an opportunity to respond independently on each training stage and delivers the least intrusive prompt first. If necessary, more intrusive prompts, usually upgraded based on the previous prompts, are then delivered to the learner to complete each training procedure. In this study, finely tuned behaviors of the executive modules were designed to reflect LTM. In Chapter II, the robot first showed a gesture to the participant and asked him/her to copy it. If the child could not do it correctly, the robot would point out where to improve verbally, re-demonstrate the gesture accordingly, as well as mirroring the participant’s behavior in the following prompts. In Chapter IV, different levels of attention attractor with special motion and acoustics effects were added on top of simple name calling as the prompt level went up. In Chapter V and Chapter VI, the system provided more informative robot motions, speeches, audio, and video gradually if the participants could not look at the correct target monitor under simple prompts.

We can see that LTM is not limited to a specific skill, but can be used as a general guidance mechanism for machine-assisted intervention. However, to our knowledge, no mathematical model has been proposed to create a general LTM framework. Therefore, In Chapter VI, we developed a general LTM-based Robot-mediated Intervention (LTM-RI) model. Such a mathematical model can be used to teach different skills to children with ASD as well as adapt the prompts for a specific skill.

8.2.2 Interaction cue detection methods

My second set of technical contributions is the design of interaction cue detection methods. These methods are the key components in the intervention systems. As mentioned before, lack of autonomous interaction cue detection methods was the main bottleneck in previous machine-assisted intervention systems. In this dissertation, we designed autonomous gesture recognition and gaze tracking methods. Note that although each of them was applied in a particular system in this dissertation, these methods are independent components that can be easily transferred to any system that requires gesture recognition and gaze tracking.

Gesture recognition methods

Gesture recognition is the essential part of an autonomous gestural imitation intervention system, since it is the key for the system to understand the response of the participants. Most of the existing gesture recognition methods are probabilistic methods such as Hidden Markov Model, particle filtering, and SVM. However, these methods do not fit in the intervention systems for young children with ASD. First, these methods need training data, but it is difficult for young children to repeat a standard gesture multiple times accurately to create the training dataset. Second, the computational complexity of these algorithms are too high for real-time gesture recognition. Third, classification-based gesture recognition methods provide binary results, i.e., a gesture is detected or not. However, in the imitation intervention, the robot needs to recognize even partially finished gestures, and tell the exact gesture stage that the participant need to improve. In addition, some existing algorithms need invasive sensors attached to the participants for body tracking, which are unlikely to be tolerated by many young children with ASD.

To solve these problems, we developed a novel non-invasive gesture detection methods in Chapter II. We used Microsoft Kinect for skeleton tracking, and thus no body-attached physical sensor was needed. The skeleton tracking data was input to a newly designed FSM-based gesture recognition algorithm, named SGR. This algorithm was low in computation complexity that guaranteed real-time recognition. It does not require training data and can recognize even partially finished gestures with high accuracy. While SGR only recognize single gestures, in chapter III, we extended it to another algorithm, named MGRS, to recognize the combination of different single gestures. In real life learning, children may perform different gestures in parallel, and this behavior need to be properly recognized by an imitation intervention system. MGRS could recognize different mixed single gestures and spot the start and end time of each of the performed gestures, either fully or partially completed. The MGRS also showed high accuracy in validation studies.

Gaze tracking methods

Gaze tracking is an indispensable component of an autonomous intervention system for teaching attentional skills to children with ASD. It allows the system to track the gaze direction of a participant and thus give response to guide or capture the participant’s attention. The most commonly applied gaze tracking method is to use an eye tracker. However, it could not be applied in a large range interactive system as ASOTS and Norris, due to its limited detection range. In addition, the participant has to hold his/her head pose during calibration and interaction, which is impossible in a large interaction environment. Therefore, we developed two new gaze tracking methods in Chapter IV and VI to solve the problems.

In Chapter IV, we developed a new large range gaze tracking method. It utilizes a network of web-cameras and thus no invasive sensor needs to be attached to the participants. Each of the camera has a limited view and could only detect the participant's frontal head pose using a SDM algorithm when the frontal face was visible to the camera. We arranged different cameras in particular positions and angles so that the view and detection ranges of these cameras were seamlessly fused. Then, transformations were performed to normalize the detection results from different cameras into a global frame. Finally, a linear mapping function computes the frontal head orientation based on the gaze direction of the participants. This method is robust, with real-time computation speed, and accurate enough for the response to name intervention discussed in Chapter IV. Then in Chapter VI, we further improved this method by training a polynomial mapping function between the head orientation and the gaze direction of the participants. This new mapping function improves the accuracy of the algorithm, while maintaining the real-time gaze tracking speed.

8.3 Contributions to the Science of ASD Intervention

In addition to the technical contributions described above, this work contributes to the science of ASD intervention by providing controllable environments where different intervention paradigms can be assessed with precisely controlled stimuli and objective measurement. Such technologically sophisticated systems are expected to play an important role in addressing common challenges of ASD intervention, including a lack of access to trained clinicians and very high costs of treatment. Therefore, we designed and conducted experimental studies with targeted population to validate the systems that we have built. Without these user studies, it was impossible to probe the feasibility, tolerability, efficacy and other potential values of these systems for young children with ASD.

In the study presented in Chapter II, the robotic system taught children single-gesture imitation skills. Using the same interaction protocol, the effect of the robot on children's gesture use was compared with that of a human therapist. Eight children with ASD and eight typically developing (TD) children tested the RISIA1 system. We found that the participants with ASD paid more attention to the robot than to a human therapist. Compared with the human therapist, the robot also promoted significantly better imitation performance. In Chapter III, we tested the RISIA2 system for teaching mixed gestures to children with ASD. Four children with ASD and two TD children tested the system. The results showed that RISIA2 also captured the attention of these children and successfully promoted mixed gesture imitation behaviors from them.

In Chapter IV, the ASOTS system was tested by 10 toddlers with ASD and 10 TD infants. The results demonstrated that ASOTS were well tolerated by both groups. This system successfully attracted and guided the attention of both groups, and stimulated response–to-name behavior with a high success rate. This shows that ASOTS has a great potential to be used: 1) as an intervention tool for children who are diagnosed with ASD; and 2) as an early screening tool for at risk infants (e.g., siblings of children with ASD) who are too young to be diagnosed with ASD. This study also quantitatively measured the participants' performance in video- and audio-based name calling environment. Overall, the experiment revealed that participants performed better in the video-based session, where both the sound and the images of the name caller was displayed.

In Chapter V, we first evaluated the feasibility of a non-invasive robot-mediated joint attention intervention system. Due to the technical limitation, this system was not fully autonomous, but provided robot-mediated intervention with the help of human operator. A small scale longitudinal experiment with 6 children with ASD showed that children with ASD tolerated a non-invasive robotic system much better than an invasive one. In addition, their attention was constantly attracted by the robot in a few intervention sessions, and their within-system joint attention performance improved significantly after the robot-mediated intervention sessions. This work encouraged us to conduct another longitudinal user study with the fully autonomous robotics system Norris in Chapter VI. A preliminary user study with 14 young children with ASD showed that Norris was well tolerated by them. Similar results to what has been shown in Chapter V was obtained from this user study. Children's within-system performance improved, and they maintained attention on the system over multiple interaction sessions. However, it was still not clear whether the robot-mediated intervention has an impact on the children in human-human interactions. Therefore, as discussed in Chapter VII, we further conducted a more rigorous randomized controlled experiment with 11 toddlers with ASD. Results showed that compared to children who did not experience robot mediated intervention, children who completed this intervention showed improved social communication skills in human-human interaction. Furthermore, we found that the participants' social communication skills in human-human interaction improved after a few sessions of robot-mediated joint attention intervention. We demonstrated, for the first time, the long-term benefits of autonomous robot-mediated ASD intervention in joint attention.

Although not designed to replace therapists or traditional behavioral intervention, we believe that the proposed systems could be used as adjunctive intervention tools to powerfully accelerate and prime learning of key skills, either prior to accessing more traditional services or in support of such services. The current work provides initial insight into how machine-assisted intervention could move from merely an interesting idea to a technological tool that can augment intervention application and science. It is our hope that this research will pave the way for a more accurate understanding of the role of machine-assisted intervention in population with ASD.

8.4 Future Work

Advances in technology may help overcome traditional resource limitations and bridge the historical gaps between early concerns about ASD and the initiation of early intervention by building technological capacity and promoting mainstream clinical use. To date, many offered technologies have focused on isolated, discrete behaviors and skills in older children (e.g., eye-gaze, emotion recognition, skill learning), rather than attempting to utilize technologies as dynamic early social support tools that promote meaningful changes in quality of life. In the future, developing machine-assisted intervention specifically to promote early improvements in core social communication skills in young children with ASD via continuous, autonomous detection and intelligent, meaningful responses to child behavior during technologically-mediated interactions will be a very important step in both engineering and scientific research.

Although promising results were achieved in this dissertation research, the studies were preliminary, and there are important limitations to be addressed in the future. First, the sample sizes for these pilot studies were small. In the future, larger samples will be needed to conduct more powerful user studies, which would strengthen corresponding statistical analyses. In addition, the user studies presented here were not formal clinical studies. Therefore, the clinical impact of the proposed systems on everyday functioning of children with ASD is still unclear. In order to conduct such studies, the systems will need to be upgraded to fit into corresponding clinical settings, and follow-up observation and measurement will need to be conducted across different settings. Although the proposed systems are capable of detecting different types of participant behaviors, only a few interaction cues were tracked, such as gesture and gaze direction. In order to precisely evaluate the behaviors of children with ASD, more sensory channels need to be integrated. For example, some participants talked during the interaction, and their speech may reflect their needs, attitude and/or interests towards an object. Therefore, the fusion of gesture recognition and gaze tracking with speech recognition may create more precise evaluation of participants' intentions and responses in future studies.

The technologies proposed in this dissertation research could be further strengthened in several ways. First, the gesture recognition algorithms were tested only with 4 simple gestures. Including more gestures, especially complex gestures, would further validate the recognition accuracy and the scalability of the algorithm. Additionally, the gaze tracking algorithms use frontal head orientation to approximate gaze direction. This method gives satisfactory accuracy when detecting whether a participant's gaze direction falls into a region of interest. However, these algorithms may not provide precise estimation if we care about which particular point a participant looks within the region of interest. In order to address this issue, advanced high resolution cameras will be required instead of simple webcams. These cameras should be able to zoom in around the eye region of the participants, so that high resolution images of the participant's eyes can be captured. Feature extraction around the eye region may provide enough information to estimate the movement of the eye balls, and thus will offer additional features for accurate gaze estimation. When designing these new algorithms, computational complexity and hardware setup need to be carefully analyzed to guarantee real-time execution.

Another important technical component of intervention systems is the interaction logic. The current studies mostly used least-to-most prompting hierarchies. However, it is not the only option. Other successful interactive procedures, such as the most-to-least prompting hierarchy, may also fit in diverse types of interventions. The improvement of software and hardware will also enhance the communication capabilities of the robot. In the current works, the robot used simple predetermined speech and motions to express instructions and encouragement to the participants. In the future, different technologies, such as speech recognition, emotion recognition, and physiological signals, can be coordinated in the system so that the robot will be able to understand the participants' status more precisely in real time. This coordination will help further customize how the robot responds to participants. In addition, in longitudinal studies, repeating the same interaction logic will cause a ceiling effect or make children lose interest over time. To solve this problem, we will need to track long term changes in the participants' behavior. Technologies such as active learning and reinforcement learning can be applied in accordance with an individualized database of each participant's responses to recognize individual differences and behavior change trajectory. Based on this information, the system can be designed to adaptively adjust its behavior and provide a specific training plan that fits a participant's individualized demands.

Important future work also includes more investigation into whether and how targeted skills generalize from human-machine interaction to human-human interaction. The human-machine interaction results we have gotten so far are encouraging. However, our goal is to help children with ASD gain better communication skills in their daily life during human-human interactions (HHI). We have conducted a pilot study on joint attention skills with encouraging results indicating skill transfer from HMI to HHI. In the future, generalization of other skills learned within machine-assisted environments (such as imitation and social orienting) should also be assessed, since this is an important step to prove the eventual efficacy of machine-assisted intervention. These generalization studies could investigate single skill generalization, e.g., whether joint attention skills learned within human-robot interaction can be generalized to the joint attention skills in human-human interaction. It is also meaningful to study the generalization across different skills, since social communication skills are not isolated but are well related and interactive in children's daily life. In this context, it is important to explore whether a particular skill learned in human-machine interaction has an impact on other skills in human-human interaction. For example, response to name is one of the starting points of paying attention and receiving messages from caregivers. If a participant's response to name skill improves during human-machine interaction, then it may be worthwhile to study whether this skill can serve as a trigger for learning other social communication skills outside the technological environment, such as joint attention and language development.

PUBLICATIONS

Full publication list of Zhi Zheng during the dissertation research:

Journals

1. Zhi Zheng**,** Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar “Longitudinal Autonomous Robot-mediated Joint Attention Intervention for Toddlers with ASD.” Human-Machine Systems, IEEE Transactions on. (under review)

2. Jing Fan, Dayi Bian, Zhi Zheng, Linda Beuscher, Paul A. Newhouse, Lorraine C. Mion, Nilanjan Sarkar, “A Robotic Coach Architecture for Elder Care (ROCARE) Based on Multi-user Engagement Models,” Neural Systems and Rehabilitation Engineering, IEEE Transactions on. (under review)

3. Zhi Zheng, Qiang Fu, Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar “Design of an Autonomous Social Orienting Training System (ASOTS) for Young Children with Autism.” Neural Systems and Rehabilitation Engineering, IEEE Transactions on. (DOI: 10.1109/TNSRE.2016.2598727)

4. Zhi Zheng, Zachary Warren, Amy Weitlauf, Qiang Fu, Huan Zhao, Amy Swanson, Nilanjan Sarkar. "Evaluation of an Intelligent Learning Environment for Young Children with Autism Spectrum Disorder." Journal of Autism and Developmental Disorders. (in press)

5. Zhi Zheng, Eric M. Young, Amy Swanson, Zachary Warren, and Nilanjan Sarkar. “Robot-mediated Imitation Skill Training for Children with Autism.” Neural Systems and Rehabilitation Engineering, IEEE Transactions on. (DOI:10.1109/TNSRE.2015.2475724)

6. Warren, Zachary, Zhi Zheng, Shuvajit Das, Eric M. Young, Amy Swanson, Amy Weitlauf, and Nilanjan Sarkar. "Brief Report: Development of a Robotic Intervention Platform for Young Children with ASD." Journal of autism and developmental disorders (2014): 1-7.

7. Bekele, Esubalew, Julie Crittendon, Zhi Zheng, Amy Swanson, Amy Weitlauf, Zachary Warren, and Nilanjan Sarkar. "Assessing the Utility of a Virtual Environment for Enhancing Facial Affect Recognition in Adolescents with Autism." Journal of autism and developmental disorders 44, no. 7 (2014): 1641-1650.

8. Warren, Zachary E., 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 (2013): 1-9.

9. Bekele, Esubalew, Zhi Zheng, Amy Swanson, Julie Crittendon, Zachary Warren, and Nilanjan Sarkar. "Understanding how adolescents with autism respond to facial expressions in virtual reality environments." Visualization and Computer Graphics, IEEE Transactions on 19, no. 4 (2013): 711-720.

Book Chapter

1. Zhi Zheng, Esubalew Bekele, Amy Swanson, Amy Weitlauf, Zachary Warren, and Nilanjan Sarkar, "The Impact of Robots on Children with Autism Spectrum Disorder (ASD). " Autism Imaging and Devices, Taylor & Francis, Dec. 2016.

Conference Proceedings

1. Zhi Zheng, Guangtao Nie, Amy Swanson, Amy Weitlauf, Zachary Warren, and Nilanjan Sarkar. "Longitudinal Impact of Autonomous Robot-mediated Joint Attention Intervention for Young Children with ASD." In the 8th International Conference on Social Robotics, 2016.

2. Zhi Zheng, Eric M. Young, Amy Swanson, Zachary Warren, and Nilanjan Sarkar. " Robot-mediated Mixed Gesture Imitation Skill Training for Young Children with ASD." In Advanced Robotics (ICAR), IEEE International Conference on, 2015.

3. Zhi Zheng, Qiang Fu, Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar. "Design of a Computer-assisted System for Teaching Attentional Skills to Toddlers with Autism." The 17th International Conference on Human-Computer Interaction, 2015. **(Best Paper Award)**

4. Zhi Zheng, Shuvajit Das, Eric M. Young, Amy Swanson, Zachary Warren, and Nilanjan Sarkar. "Autonomous robot-mediated imitation learning for children with autism." In Robotics and Automation (ICRA), 2014 IEEE International Conference on, pp. 2707-2712. IEEE, 2014.

5. Bekele, Esubalew, Joshua W. Wade, Dayi Bian, Lian Zhang, Zhi Zheng, Amy Swanson, Medha Sarkar, Zachary Warren, and Nilanjan Sarkar. "Multimodal Interfaces and Sensory Fusion in VR for Social Interactions." In Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments, pp. 14-24. Springer International Publishing, 2014.

6. Bekele, Esubalew, Dayi Bian, Zhi Zheng, Joel Peterman, Sohee Park, and Nilanjan Sarkar. "Responses during Facial Emotional Expression Recognition Tasks Using Virtual Reality and Static IAPS Pictures for Adults with Schizophrenia." In Virtual, Augmented and Mixed Reality. Applications of Virtual and Augmented Reality, pp. 225-235. Springer International Publishing, 2014.

7. Zhi Zheng, Lian Zhang, Esubalew Bekele, Amy Swanson, Julie A. Crittendon, Zachary Warren, and Nilanjan Sarkar. "Impact of robot-mediated interaction system on joint attention skills for children with autism." In Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on, pp. 1-8. IEEE, 2013. **(Best student paper finalist**)

8. Bekele, Esubalew, Mary Young, Zhi Zheng, Lian Zhang, Amy Swanson, Rebecca Johnston, Julie Davidson, Zachary Warren, and Nilanjan Sarkar. "A step towards adaptive multimodal virtual social interaction platform for children with autism." In Universal Access in Human-Computer Interaction. User and Context Diversity, pp. 464-473. Springer Berlin Heidelberg, 2013.

9. Bekele, Esubalew, Zhi Zheng, Amy Swanson, Julie Davidson, Zachary Warren, and Nilanjan Sarkar. "Virtual reality-based facial expressions understanding for teenagers with autism." In Universal Access in Human-Computer Interaction. User and Context Diversity, pp. 454-463. Springer Berlin Heidelberg, 2013.

10. Bekele, Esubalew, Zhi Zheng, Uttama Lahiri, Amy Swanson., Julie Davidson, Zachary Warren, Nilanjan Sarkar. (2012). Design of a novel virtual reality-based autism intervention system for facial emotional expressions identification. In The 9th International conference on Disability, Virtual Reality and Associated Technologies.

Conference Abstract

1. Zhi Zheng, Qiang Fu, Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar. "3-D Social Attention Training for Young Children with ASD" International meeting for autism research, 2016. (accepted)

2. Zhi Zheng, Qiang Fu, Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar. "A 3-D learning environment for infants and toddlers at-risk for ASD: Can technology improve early social communication vulnerabilities?" International meeting for autism research, 2015.

Workshop paper

1. Zhi Zheng, Huan Zhao, Amy Swanson, Amy Weitlauf, Zachary Warren, Nilanjan Sarkar. "A Fully Autonomous Robotic System for Training Attentional Skills to Toddlers with ASD" International Workshop on Intervention of Children with Autism Spectrum Disorders using a Humanoid Robot, 2015.