Intelligent Systems for Autism Spectrum Disorders and Schizophrenia Intervention – Design, Development and User Studies

By

Esube Bekele

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, 2015
Nashville, Tennessee

Approved:

Nilanjan Sarkar, Ph.D.

Zachary Warren, Ph.D.

George Cook, Ph.D.

Mitch Wilkes, Ph.D.

Pietro Valdastri, Ph.D.

The presented work is dedicated to

my loving girlfriend Tsige Gebresslase,

my brave mother Atale Belay,

my selfless sister Wossenie Tsegaye,

and helpful brother Embiale Zemedie.

Thank you all.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Nilanjan Sarkar and Dr. Zachary Warren for their continued support and help. Their encouragement and help was invaluable and without them this dissertation would not have been possible.

I would also like to express my gratitude to TRIAD members Amy Swanson, Nicole Bardett and Amy Weitlauf for all of the hard work and for their continuous help.

I would also like to extend my gratitude to the committee members for their willingness and help during my studies and their flexibility.

I especially want to acknowledge the wonderful graduate students of the Robotics and Autonomous Systems Lab, who have been excellent colleagues as well as very good friends. Specifically, thanks go to Josh Wade, Jing Fan, Zhi Zheng, Dayi Bian, and Lian Zhang for frequent discussions, feedback, and support.

My heart felt gratitude also goes to my family and friends who persisted with me throughout my days at Vanderbilt. I would like to specially give thanks to Asfaw Beka, Hambisa Assefa, Michael Tesfahuney, and Chale Tirfe for being there when I needed them.

Finally, acknowledge the funding support by the National Science Foundation Grant 0967170, National Institute of Health Grant 1R01MH091102-01A1, and NARSAD Distinguished Investigator Grant 19825 from the Brain & Behavior Research Foundation.

TABLE OF CONTENTS

Page

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

Chapter

I. INTRODUCTION AND BACKGROUND 1

A. Overview 1
B. Significance 1
C. Related Work 3

II. SPECIFIC OBJECTIVES 8

A. Specific Aim 1: An Architecture for Robot-mediated Joint Attention task 8
B. Specific Aim 2: Design of an Intelligent Multimodal Virtual Environment for Facial Emotion Recognition Tasks for ASD and SZ Intervention 8
C. Specific Aim 3: Design and Evaluation of a Social Interaction Task in Virtual Environment 10

III. ROBOT-MEDIATED JOINT ATTENTION FOR ASD 11

A. System Development 11
B. Usability Study 17
C. Results 21
D. Discussion and Conclusion 25

IV. EMOTION RECOGNITION AND FACIAL PROCESSING IN VR FOR ASD 28

A. System Design 28
B. Methods and Procedure 35
C. Results 37
D. Discussion and Conclusion 49

V. DESIGN OF A VIRTUAL REALITY SYSTEM FOR AFFECT ANALYSIS IN FACIAL EXPRESSIONS (VR-SAAFE) AND IAPS PICTURES PRESENTATION; APPLICATION TO SCHIZOPHRENIA 52

A. The Static IAPS presentation 53
B. VR Emotion Presentation 53
C. Peripheral Monitoring Components 56
D. Offline Data Analysis 57
E. Methods and Procedure 57
F. Results 59

G. Discussion and Conclusion 68

VI. MULTIMODAL ADAPTIVE SOCIAL INTERACTION IN VIRTUAL ENVIRONMENT (MASI-VR) FOR CHILDREN WITH AUTISM SPECTRUM DISORDERS (ASD) 71

A. System Design 71
B. Data Analysis 81
C. Methods and procedure 82
D. Results and Discussion 87
E. Overall Conclusion 94

VII. POTENTIAL CONTRIBUTION 96

A. Technological Contribution 96
B. Scientific Contribution 97
C. Societal Contributions 98

APPENDIX A 99

Details of the Camera Processing Module (CPM) 99

APPENDIX B 101

Physiological Features 101

REFERENCES 103

LIST OF TABLES

Table Page

1. Levels of the Hierarchical Protocol 18
2. Profile of the TD participants in robot-mediated JA study 20
3. Profile of the participants with ASD in robot-mediated JA study 20
4. Physiological Feature Sets used in this Study 34
5. Profile of Individual Subjects in the ASD Group 36
6. Profile of Individual Subjects in the TD Group 36
7. Amount of time spent looking at Mouth and Forehead ROIs across groups. 42
8. Measures of behavioral viewing pattern 43
9. Group Differences on Performance Metrics (out of 28 presentations) 48
10. Average recognition accuracy by emotion of the college group testing. 48
11. Profile of the first 6 subjects in the patient group and the matched control group 58
12. Eye Tracking Features for the VR session 60
13. Physiological Features for the VR session 63
14. Response Bias of subjects for each Emotion 67
15. Categories of Animations 76
16. Summary of Visits 83
17. EEG features for the VR session 93
18. Physiological Features extracted for the social task 101

LIST OF FIGURES

Figure Page

1. The experiment room setup. 12
2. System schematics for closed-loop interaction. 13
3. The Humanoid Robot NAO 14
4. The hat holds the array of LEDs at side and top. 14
5. Camera Processing Module (CPM) schematics. 16
6. Flowchart of the prompt-feedback mechanism for one trial. 19
7. Across group comparison for the time spent on the human and robot therapists. 22
8. Percentage of time spent looking at robot and human therapists during respective sub-sessions for the ASD group. 22
9. Percentage of time spent looking at robot and human therapists during respective sub-sessions for the TD group. 23
10. Percentage of correct responses achieved at specific prompt levels by group and condition (each point represents percentage of correct response at that prompt level and below). 24
11. VR-based facial expressions presentation system. 28
12. The Eye tracking application and its components. 29
13. Representative characters used in the study. 31
14. Anger (top) and surprise (bottom) with two arousal levels. 31
15. Neutral (far left) and the four arousal levels for surprise. 31
16. The Eye tracking application and its components. 32
17. Gaze towards mouth and forehead regions. 38
18. Gaze towards face and non-face regions. 38
19. Gaze towards mouth and forehead regions. 39
20. Gaze towards face and non-face regions. 39
21. Gaze towards mouth and forehead regions. 40
22. Gaze towards face and non-face regions. 40
23. Gaze towards mouth and forehead regions. 41
24. Gaze towards face and non-face regions. 41
25. Intergroup comparison gaze visualizations (heat maps and masked scene maps) (a) combined gaze in the ASD group for all the trials and all participants (b) combined gaze in the TD group for all the trials and all participants. 43
26. Comparisons of behavioral eye indices. 45
27. Comparisons of physiological eye indices. 45
28. (a) Top left: original ground truth clusters. (b) Top right: the Gaussian mixtures used in the GM clustering overlaid on the ground truth clusters. (c) Bottom left: the result of the k-mean clustering. (d) Bottom right: the result of GM clustering. 47
29. Left: Results showing clustering quality of the two clustering methods using two different sets of PCA components for all the four sets of comparisons. Cor: correct and Inc: incorrect. Right: (a) Top left: result of the GM clustering, (b) Top right: result of the k-means clustering, and (c) Bot-tom: ground truth of clusters of data of from trials of ASD subjects when they were correctly identifying the emotions vs. when they were incorrect in identifying the emotions. 47
30. IAPS pictures presentation system with the arousal rating. 53
31. Overall System Diagram. 54
32. Example emotions with its different degrees of arousal. 55

33. The Eye Tracking Application Components. 56
34. Peripheral physiological electrodes placement. 56
35. Experimental setup. 58
36. Face ROIs used for gaze analysis. 60
37. Masked maps overlaid on heat map visualizations of the patient group (top) and control group (bottom). 62
38. Physiological clustering using the density peaks method. Positive vs. negative categories for SZ (top) and CTR (bottom). 64
39. Clustering accuracies using the density peaks method. 65
40. Overall performance metrics. 66
41. Per emotion raw (biased) and bias-corrected performance. 68
42. (Top) Components of emotional social interaction and (bottom) system architecture of MASI-VR. 72
43. The VR cafeteria environment for social task training. Dining area and food dispensary area. The two areas are constructed in separate rooms. 74
44. Representative characters displaying example emotions and gestural animations. 75
45. The script creator program that was developed to generate the mission files. 77
46. Finite state diagram showing a simple example level switching logic. 78
47. The peripheral physiological sensors 79
48. Eye tracking application and its components 80
49. The five facial ROIs defined on the face region. 81
50. The experimental setup. 83
51. The spoken conversation and the emotion recognition 85
52. Generalized performance metrics 88
53. (Top) bias per emotion, and (bottom) raw and bias corrected performance 89
54. (top) Gaze towards ROIs defined in Fig. 7 and gaze towards combined ROIs. 90
55. Gaze Group Pre (top) and Post (bottom). 91
56. Control (CTR) Group Pre (top) and Post (bottom). 91
57. Correlations of physiological eye indices. 92
58. Comparisons of physiological eye indices. 92
59. Comparisons of physiological eye indices. 93
60. Combined EEG feature 94
61. Top and side perspective projections of the LED arrays and the targets (LCD monitors) 99
62. Ray-line segment intersection for one target (P_0P_1) and a LEDs ray ($Q_0P(s)$) 100

LIST OF ABBREVIATIONS

ADI-R	Autism Diagnostic Interview-Revised
ADOS-G	Autism Diagnostic Observation Schedule-Generic
ASD	Autism Spectrum Disorder
BR	Blink Rate
BVP	Blood Volume Pulse
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders-4th edition
ECG	Electrocardiogram
EDA	Electrodermal Activities
EMG	Electromyogram
FNFR	Face-to-Non-Face Ratio
GSR	Galvanic Skin Response
HCI	Human-computer Interaction
HRI	Human-robot Interaction
IAPS	International Affective Pictures System
IBI	Inter Beat Interval
IRB	Institutional Review Board
OFR	Object-to-Face Ratio
PCG	Phonocardiogram
PD	Pupil Diameter
PEP	Pre-Ejection Period
PPG	Photoplethysmogram
PPVT	Peabody Picture Vocabulary Test
PTT	Pulse Transit Time
ROI	Region of Interest
SCQ	Social Communication Questionnaire
SD	Standard Deviation
SRS	Social Responsiveness Scale
SVM	Support Vector Machines
SZ	Schizophrenia
VR	Virtual Reality

CHAPTER I

INTRODUCTION AND BACKGROUND

A. Overview

This doctoral research explores the design and implementation of intelligent robotic and virtual reality systems for personalized treatment of autism spectrum disorders (ASD) and schizophrenia (SCZD or SZ) interventions. Recent advances in robotic, virtual reality, and sensor technologies are utilized by designing, implementing and testing robot-assisted intervention for children with ASD, virtual reality-based facial emotional expression recognition for children with ASD and adults with schizophrenia, and social interaction and contextual emotion understanding in a virtual environment.

B. Significance

The Role of Intelligent Technology in Autism Behavioral Therapy.
Autism spectrum disorders (ASD) are characterized by difficulties in social communication as well as repetitive and atypical patterns of behavior (*Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the diagnostic criteria from DSM-IV-TR*, 2000). According to a new report by the Centers for Disease Control and prevention (CDC), an estimated 1 in 68 children and an estimated 1 out of 42 boys (with 5 times prevalence than girls) in the United States have ASD (Developmental & Investigators, 2014). This is a 30% increase compared to the previous report of 1 in 88 (Baio, Autism, Network, Control, & Prevention, 2012) which in turn was a 78% increase since the CDC report in 2009 ("Prevalence of Autsim Spectrum Disorders-ADDM Network," 2009). The average lifetime cost of care for individual with autism is estimated to be around $3.2 million, with average medical expenditures for individuals with ASD 4.1–6.2 times greater than for those without ASD ("Autism Spectrum Disorders Prevalence Rate," 2011; Ganz, 2007; Peacock, Amendah, Ouyang, & Grosse, 2012). With these alarming prevalence figures, effective early identification and treatment of ASD is considered a pressing clinical care and public health issue ("Interagency Autism Coordinating Committee Strategic Plan for Autism Spectrum Disorder Research," 2009). At present, there is no cure or single accepted intervention for autism with lifespan outcomes of the disorder varying widely (*Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the diagnostic criteria from DSM-IV-TR*, 2000).

To address the powerful impairments and costs associated with ASD, a wide variety of potential interventions have been offered. The cumulative literature suggests substantial benefits of early, intensive, ASD specific interventions; however, outcomes vary greatly for individuals, this variation is poorly understood, and most individuals continue to display potent impairments in many areas despite significant improvements (Rogers and Vismara, 2008; Howlin, Magiata, and Charman, 2009; Warren et al. 2011). While the causes for differential response to treatment are unclear, the strongest finding for effecting change comes from interventions that incorporate intensive behavioral intervention (Warren Z. E., Veenstra-VanderWeele J, & Stone W, 2011). Intensive behavioral interventions, however, typically involves long hours from qualified therapists who are not available in many communities or are beyond the financial resources of families and service systems (Eikeseth, 2009). Therefore, such high quality behavioral intervention is often inaccessible to the wide ASD population (Chasson, Harris, & Neely, 2007; Knapp, Romeo, & Beecham, 2009) that is increasingly identified at younger ages.

Given the present limits of intervention science and the powerful nature of early impairments across the lifespan, there is urgent need for the development and application of novel treatment paradigms capable of substantially more efficacious individualized impact on the early core deficits of ASD. Given rapid progress

and developments in technology, it has been argued that innovative computer and robot oriented technologies could be effectively harnessed to provide innovative clinical treatments for individuals with ASD (Goodwin, 2008). Emerging technology (Goodwin, 2008) may ultimately play a crucial role in filling the gap for those who cannot otherwise access behavioral intervention. Moreover, these technologies such as computer technology (Bernard-Opitz, 2001; C. Liu, Conn, K., Sarkar, N., Stone, W., 2008b; Moore, 2000), virtual reality environments (Conn, 2008; U. Lahiri, Bekele, E., Dohrmann, E., Warren, Z., Sarkar, N., 2011; Mitchell, 2007; Welch, 2009), and robotic systems (K. Dautenhahn, Billard, A., 2002; C. Liu, Conn, K., Sarkar, N., Stone, W., 2007b) show the potential to make current interventions more individualized and thus more powerful (Sandall, 2005). In fact, a number of recent studies suggest that specific applications of robotic systems can be effectively harnessed to provide new important directions for intervention, given potential capacities for making treatments highly individualized, intensive, flexible, and adaptive based on highly relevant novel quantitative measurements of engagement and performance (Diehl, Schmitt, Villano, & Crowell, 2011).

A number of studies have investigated the application of advanced interactive technologies to ASD intervention, including computer technology(Goodwin, 2008), virtual reality (VR) environments (Bellani, Fornasari, Chittaro, & Brambilla, 2011), and more recently, robotic systems (Diehl et al., 2011).

Advances in robotic technology have certainly demonstrated the capacity for intelligent robots to fulfill a variety of human-like and neuro-rehabilitative functions in other populations (K. Dautenhahn, 2003; K. Dautenhahn, Werry, I., 2004), but well-controlled research focusing on the impact of specific clinical applications for individuals with ASD is very limited (Diehl et al., 2011). The most promising finding regarding robotic interaction to date has been a documented preference by some individuals with ASD, in certain circumstances, for technological interaction versus human interaction. Specifically, data from several research groups have demonstrated that many individuals with ASD show a preference for robot-like characteristics over non-robotic toys and human (K. Dautenhahn, Werry, I., 2004; B. Robins, Dautenhahn, K., Dubowski, J., 2006) and in some circumstances even respond faster when cued by robotic movement than human movement (Bird, 2007; Pierno, 2008). While this research has been accomplished with school-aged children and adults, research noting the preference for very young children with autism to orient to nonsocial contingencies rather than biological motion suggests that downward extension of this technological preference may in fact be more salient and perhaps more closely linked to the neurobiological mechanisms that are the hallmark of ASD (Annaz, Campbell, Coleman, Milne, & Swettenham, 2012; A. Klin, Lin, D.J., Gorrindo, P., Ramsay, G., Jones, W., 2009).

Virtual Reality technology possesses several strengths in terms of potential application to ASD intervention, including: malleability, controllability, replication ability, modifiable sensory stimulation, and the potential capacity to implement individualized intervention approaches and reinforcement strategies. Virtual Reality can also depict various scenarios that may not be feasible in a "real world" therapeutic settings, given naturalistic social constraints and resource challenges (Kandalaft, Didehbani, Krawczyk, Allen, & Chapman, 2012; S. Parsons, Mitchell, P., 2002). As such, VR appears well-suited for creating interactive skill training paradigms in core areas of impairment for individuals with ASD.

The Role of Intelligent Technology in Schizophrenia Therapy.

Schizophrenia (SZ) is a debilitating psychotic disorder that affects about 1% of the population, costing more than $100 billion annually in the USA. It causes emotional and cognitive impairments (Hempel et al., 2005) and is defined as a splitting of thoughts from feelings (Bleuler, 1911). Some of the psychotic symptoms such as hallucinations and delusions are partly ameliorated by antipsychotic drugs, but the route to recovery is hampered by social impairments (Couture, Granholm, & Fish, 2011). Marked social impairments are present in individuals with schizophrenia at all stages of this illness. Currently, there are no effective pharmacological treatments. Although, behavioral treatments are more effective than pharmacological treatments, their success is limited for social cognitive deficits in schizophrenia. Thus, understanding causal mechanisms of these social cognitive deficits would be the first step toward implementing effective social interventions.

Moreover, developing innovative methods of measuring social cognitive deficits in schizophrenia will contribute towards elucidating the neural and cognitive bases of abnormal social behavior.

Currently available social interventions can be helpful but low compliance rates and lack of access to such programs for most patients can be problematic. Deficits in social cognition, including emotion processing, social cue perception, empathy, mental state attributions, and theory of mind lead to poor functional outcome in SZ even after improvement in psychotic symptoms (Andreasen, 1983; Esubalew Bekele, Mary Young, et al., 2013). Thus, there is a need for efficacious cost-effective, low-burden and high-compliance interventions for social deficits in SZ, which would likely increase positive outcome. Improvement in emotion processing, a core deficit, and social understanding would be crucial for improved social outcomes. SZ patients appear to have impairments in recognizing faces and emotional expressions and disturbances in emotional functioning are major disability in SZ (Herbener, Song, Khine, & Sweeney, 2008; Kring & Moran, 2008). Much of the work on social cognitive deficits in individuals with schizophrenia has focused on emotion perception. Emotional experiences involve the interaction of multiple subcomponents (e.g., subjective experience, outward expression, and physiological reactivity, among others) which when working in concert leads to adaptive responses to the environment (Johnson & Tassinary, 2005; Keltner & Gross, 1999). Emotional experiences and expressions of individuals with schizophrenia have been of particular interest for a long time with breakdown in coherence of emotional responses Various studies investigated the self-report of emotional experiences in repose to film clips (Kring, Kerr, Smith, & Neale, 1993; Kring & Neale, 1996) and still images (Herbener et al., 2008). Individuals with schizophrenia report experiencing the same or greater amounts of emotional experiences when compared to controls (Cohen & Minor, 2010). When presented with emotionally evocative stimuli, individuals with schizophrenia do not show facial expressions of emotion even though they report experiencing an emotion (Berenbaum & Oltmanns, 1992; Earnst & Kring, 1999; Kring & Moran, 2008). Generally, the bulk of these studies have used static images of emotionally expressive faces to test emotion recognition (Healey, Pinkham, Richard, & Kohler, 2010). More specifically the International Affective Picture System (IAPS), were used to elicit emotional experience in SZ (Hempel et al., 2005). However, using stimuli that convey the temporal nature of emotional expressions will provide a clearer understanding of the nature of emotion recognition deficits in schizophrenia. Another important factor in favor of the use of dynamic stimuli and technology-assisted intervention is the absence of internal measures that could potentially alter the stimuli in an effort to understand the underlying mechanisms of the disorder itself as well as to use implicit measures as therapeutic components. In this regard, physiological signals from peripheral central nervous system could play an important role as internal measures of emotions felt by patients with SZ. The apparent disconnect between outward display of emotion by SZ patients and their actual internal feeling could be studied by understanding the involuntary peripheral physiological responses of the sympathetic central nervous system (CNS). These studies may eventually lead to using these internal measures as a way of modeling the emotional state of the patient.

C. Related Work

While the above sections discussed the importance of technology assisted intervention for both SZ and ASD intervention, the following sections present related work in the field of intelligent robotic and virtual reality systems and their applications in assistive therapy for SZ and ASD. These sections also pave the way for the next chapter and the core objective of this dissertation research by discussing the current state-of-the-art intelligent therapeutic systems in these fields and how they can be further advanced.

Robot-Mediated ASD Intervention.

Despite suggested benefits of adaptive robotic technology for individuals with ASD, both potent methodological and existing system limits present challenges to understanding feasibility and ultimate clinical utility.

Well-controlled research focusing on the impact of specific clinical applications of robot-assisted ASD systems have been very limited (Diehl et al., 2011; D. Feil-Seifer, Matarić, M., 2009; Tapus, 2007). The most promising finding documents a preference for technological interaction versus human interaction for some children in certain circumstances. Specifically, data from several research groups has demonstrated that many individuals with ASD show a preference for robot-like characteristics over non-robotic toys and humans (K. Dautenhahn, Werry, I., 2004; B. Robins, Dautenhahn, K., Dubowski, J., 2006), and in some circumstances even respond faster when cued by robotic movement than human movement (Bird, 2007; Pierno, 2008). While this research has been accomplished with school aged children and adults, research noting the preference for very young children with autism to orient to nonsocial contingencies rather than biological motion suggests a downward extension of this intervention. Broad reactions and behaviors during interactions with robots have been investigated with this age group but studies have yet to apply appropriately controlled methodologies with well-indexed groups of young children with ASD, which is needed as the bias toward nonsocial stimuli may be more salient for younger children (A. Klin, Jones, Schultz, Volkmar, & Cohen, 2002; A. Klin, Lin, D.J., Gorrindo, P., Ramsay, G., Jones, W., 2009).

Beyond the need for research with younger groups, to date there have been very few applications of robotic technology for teaching, modeling, or facilitating interactions through directed intervention and feedback approaches (Diehl et al., 2011). In the only identified study in this category to date, Duquette, et al. (Duquette, 2008) demonstrated improvements in affect and attention sharing with co-participating partners during robotic imitation interaction task. This study provides initial support for the use of robotic technology in ASD intervention. Another limitation of the current evidence base is the fact that most robotic systems studied have primarily been remotely operated and unable to perform autonomous closed-loop interactions where learning and adaptation are incorporated into the system. Hence, these approaches may be very limited in terms of application to intervention settings for extended and meaningful interactions (Billard, Robins, Nadel, & Dautenhahn, 2006; K. Dautenhahn et al., 2002; Kozima, 2005). Further, they often require significant resources for operation by necessitating simultaneous involvement of both sophisticated robotic systems and sophisticated system administrators. Open-loop systems utilize robots with pre-programmed behavior, and at the time of interaction, are either remotely operated by humans or execute the pre-programmed behaviors in simple form. Closed-loop systems, which are also referred to autonomous systems, utilize robots that alter their behavior in reaction to environmental interactions and sensor input based on control logic. Of those autonomous systems that can, not only autonomously react but also adapt their behaviors over time based on the interaction are referred to as adaptive robotic systems. Adaptive closed-loop systems have been hypothesized to offer technological mechanisms for supporting more flexible and potentially more naturalistic interaction, but have rarely been applied to specific ASD applications. Preliminary results with computer and robot-based adaptive response technology for intervention indicates the potential of adaptive technology to flexibly adapt to children with ASD (Conn, 2008; C. Liu, Conn, K., Sarkar, N., Stone, W., 2007a, 2007b; Rani, Liu, Sarkar, & Vanman, 2006). Feil-Seifer et al. (D. Feil-Seifer, Mataric, M., 2011) used distance-based features to autonomously detect and classify positive and negative robot interactions, while Liu et al. (C. Liu, Conn, Sarkar, & Stone, 2008) used peripheral physiological signals to adaptively change robot behavior based on psychological state of the participant. These are among the few adaptive robot assisted autism therapeutic systems to date.

A final limitation of the ASD robotic application literature, is the fact that studies have commonly assessed broad reactions and behaviors during interactions with robots (Kozima, 2005), rather than focusing on skills that relate to the core deficits of ASD (Diehl et al., 2011; B. Robins, Dautenhahn, Te Boekhorst, & Billard, 2004; B. Robins, Dickerson, Stribling, & Dautenhahn, 2004). One work was specifically addressing robot-mediated joint attention task but it was a case study and hence lacks broader impact and clinical relevance (B. Robins, Dickerson, et al., 2004). Despite the recent works mentioned above that have piloted specific closed-looped systems with potential applicability to ASD populations (D. Feil-Seifer, Mataric, M., 2011; C. Liu et al., 2008), they have not yet examined impact of applications to relevant core deficit areas of the

disorder. Studies have yet to investigate the utilization of intelligent and dynamic robotic interaction in attempts to directly address skills related to the core deficits of ASD.

Virtual Reality-based ASD Intervention.

Among the fundamental social impairments of ASD are challenges in appropriately recognizing and responding to nonverbal cues and communication, including challenges recognizing and appropriately responding to facial expressions (Adolphs, Sears, & Piven, 2001; Castelli, 2005). In particular, individuals with ASD may have impaired face discrimination, slow and atypical face processing strategies, reduced attention to eyes, and unusual strategies for consolidating information viewed on other's faces (G. Dawson, Webb, & McPartland, 2005). Although children with ASD have been able to perform basic facial recognition tasks as well as typically developing peers in certain circumstances (Castelli, 2005), they have shown significant impairments in efficiently processing and understanding complex, dynamically displayed facial expressions of emotion (Bölte et al., 2006; Capps, Yirmiya, & Sigman, 1992; G. Dawson et al., 2005; Weeks & Hobson, 1987).

A number of research groups have attempted to utilize computer technology to improve facial affect recognition (Golan et al., 2010; Golan & Baron-Cohen, 2006; Lacava, Rankin, Mahlios, Cook, & Simpson, 2010). A detailed and comprehensive reviews of computer-assisted technologies for improving emotion recognition (Ploog, Scharf, Nelson, & Brooks, 2012) note that specific facial recognition skills can be improved using existing systems. How well these skills transfer to real-world settings, however, remains largely unstudied. One study showed that performance in computerized training correlated with brain activation (Bölte et al., 2006). Otherwise, skill transfer, or generalization of skills beyond the computer to meaningful social situations, has been limited if examined at all.

Virtual reality (VR) (S. Parsons, Mitchell, P., Leonard, A., 2004; Welch, 2009) have been proposed for ASD intervention as VR platforms are promising for improving social skills, cognition and overall social functioning in autism (S. Parsons, Mitchell, P., 2002). Basic VR-based interaction for children with ASD began in 1990s (S. Parsons & Cobb, 2011). Various displays including immersive head mounted displays (HMD) were employed in the early phases of virtual interaction of children with ASD. However, HMD were rated as heavy and causing discomfort and as a result, desktop-based VR were preferred to HMD (Wang & Reid, 2011). Basic social skills training and navigating in the VR environment were examined with some success (Mitchell, 2007). The use of VR for understanding and interpreting basic facial emotional expressions were studied and found that children with ASD performed well in VR (Fabri, Elzouki, & Moore, 2007). Virtual reality environments offer benefits to children with ASD mainly due to their ability to simulate real world scenarios in a carefully controlled and safe environment (Josman, Ben-Chaim, Friedrich, & Weiss, 2011; S. Parsons & Cobb, 2011). Controlled stimuli presentation, objectivity and consistency, and gaming factors to motivate for task completion are among the primary advantages of using VR-based systems for ASD intervention. For multiple reasons, a VR paradigm paired with markers of performance and gaze processing may be a more effective way to teach and generalize skills of emotion recognition and social communication. Specifically, VR can provide a controlled and replicable environment where specific recognition skills can be tested and taught in a dynamic fashion. While earlier intervention approaches have tended to rely on static pictures or presentations to teach recognition skills, a VR environment approximates properties of facial expressions as they dynamically develop, appear, and change in reality. Further, the capacity for dynamic presentation means that stimuli can be controlled and altered to teach increasingly subtle displays of facial affect. Skills can be practiced in a variety of virtual scenarios (e.g., altering avatars, context, environment) to promote generalization. While it was shown that these systems can help generalization across contexts (Schmidt & Schmidt, 2008), generalization to real-world interactions remains an open question.

Explicit modalities such as audio visual for natural multimodal interaction (Lang, Bradley, & Cuthbert, 1999) were used as explicit interfaces. Whereas peripheral physiological signals (Herbener et al., 2008; Welch, 2009) and eye tracking (Lahiri, Bekele, Dohrmann, Warren, & Sarkar, 2012) were used as implicit

interfaces in such social scenarios. Identifying the psychological states of the user and hence adapt the interaction accordingly is crucial in social interactions in general and in VR in particular (Esubalew Bekele, Zhi Zheng, et al., 2013; Zeng, 2009). A growing number of studies are investigating applications of advanced VR to social and communication related intervention (Blocher, 2002; Kandalaft et al., 2012; Mitchell, 2007; S. Parsons, Mitchell, P., Leonard, A., 2004; Ploog et al., 2012). Increasingly, researchers have attempted to develop VR and other technological applications that respond not only to explicit human-computer interactions (e.g., utilization of keyboards, joysticks, etc.), but to dynamic interactions such as those incorporating eye gaze and physiological measurements (Lahiri, Bekele, et al., 2012; Wilms et al., 2010). However, most of the existing VR environments applied to assistive intervention for children with ASD are designed to build skills based on aspects of performance alone (i.e., correct or incorrect and some performance metrics), thereby limiting individualization of application. Recent research in VR systems for application of Attention Deficit Hyperactivity Disorders (ADHD), ASD, and cerebral palsy suggest making such VR systems feedback-based may result in increased engagement and individualization (Lahiri, Bekele, et al., 2012). Some of these studies used touchscreens to teach children with ASD skills such as pretend play, turn-taking, and other social skills (e.g., directed eye gaze) using a co-located cooperation enforcing interface, called StoryTable (Bauminger et al., 2007; Gal et al., 2009).

While haptic feedback can be useful in certain contexts, understanding eye gaze and physiological response during emotion recognition and social cognition tasks can be critical since both eye gaze and physiology have been shown to convey a tremendous amount of information regarding the emotion recognition process (C. Liu, Conn, K., Sarkar, N., Stone, W., 2008b; Ruble & Robson, 2007). Adaptive social interaction using implicit cues from sensors such as peripheral physiological signals (Kandalaft et al., 2012) and eye tracking (Lahiri, Warren, & Sarkar, 2012) are of particular importance. VR systems that not only gauge performance on specified tasks but also automatically detect eye-gaze or other physiological markers of engagement may hold promise for additional optimization of learning (Lahiri, Bekele, et al., 2012; Welch, 2009; K. C. Welch, Lahiri, U., Sarkar, N., Warren, Z., Stone, W., Liu, C., 2010).

There are a few recent studies that incorporate peripheral psychophysiological signals (C. Liu, Conn, K., Sarkar, N., Stone, W., 2008b; K. Welch, Sarkar, M., Sarkar, N., Liu, C., 2010) and eye gaze monitoring (Lahiri, Bekele, et al., 2012) into VR systems as applied to ASD intervention. These systems monitor several channels of physiological signals to determine the underlying affective states of the subject for individualized VR social interactions. The capacity to embed an eye-tracker and physiological monitoring within a VR environment makes it possible to individualize training sessions beyond simple measures of performance. More specifically, eye tracking may show how gaze has been used to process the emotion and as such provide essential feedback to inform intervention (Lahiri, Warren, et al., 2012). Because research has consistently documented ASD-specific processing differences in this regard (A. Klin et al., 2002; Pelphrey et al., 2002), developing platforms for gathering such information about affect recognition processing (with the potential for on-line adaptive change of the VR stimuli itself, e.g., highlighting, occluding, and shifting attention to specific features) may yield an efficient training tool for promoting both learning within the environment as well as generalization beyond the virtual environment. Moreover, explicit conversational dialog is an important part of social interaction. Recently spoken conversational modules have been incorporated into VR systems to achieve more natural interaction instead of menu driven dialog management. Instead of a large vocabulary, domain independent natural language understanding, a limited vocabulary question-response dialog management, which is focused on the specific domain, has been shown to be effective (Kenny, Parsons, Gratch, Leuski, & Rizzo, 2007; Leuski, Patel, Traum, & Kennedy, 2009).

Virtual Reality-based Schizophrenia Intervention.

Partly, SZ is also characterized by similar social and emotional impairments as ASD such as emotion identification and processing impairments. More specifically disparate emotional experiences are of particular interest as discussed above. Therefore a subset of tools used in the ASD intervention may be suitable to be

applied to SZ behavioral therapy under certain circumstances. Specifically, the use of VR technology holds promise in SZ intervention for the same reason of flexibility, controllability, and adaptability as is the case in ASD intervention. Another factor that necessitates the use of VR is the ability of incorporating implicit cues such as peripheral physiological signals with future adaptive potential as opposed to static images and isolated and non-interactive video clips. In the context of technology-based SZ intervention, Virtual Reality (VR) systems have been investigated with SZ for symptom assessment (Lan et al., 2009), training of medication management skills (Di, Huang, Sekiyama, & Fukuda, 2011), hallucinations training (Almeida, Zhang, & Liu, 2007), social perception (Bourkeφ, Scanaillφ, Culhaneφ, O'Brien, & Lyonsφ, 2006), role playing (Argyle & Dean, 1965) and improving the diagnosis of SZ (Huang, Di, Wakita, Fukuda, & Sekiyama, 2008). However there are limited applications of VR in emotion processing and identification for SZ. Moreover, these VR systems solely rely on user reporting and outward measures of performance. To mitigate these limitations, one should combine dynamic presentation of emotional expressions together with implicit physiological response and eye gaze processing. Implicit cues can be useful to understand the underlying psychological states that are not possible using performance-based systems or simple user reporting.

The inability of individuals with SZ to express their emotions outwardly despite reporting to experience the emotions means implicit cues hold important clues to understand the emotions experienced by them. Generally, research on the physiology of emotional reactions in individuals with schizophrenia has been mixed depending on the physiological response recorded and the paradigm of emotion elicitation used (Kring & Moran, 2008). Early investigations into the physiological reactivity of individuals with schizophrenia using galvanic skin response (GSR), a measure of autonomic arousal, suggested that subgroups of patients with schizophrenia show different GSR profiles given their clinical presentation (Venables & Wing, 1962). Although some other studies have not found a difference between individuals with schizophrenia and controls using emotionally evocative images (Hempel, Tulen, van Beveren, Mulder, & Hengeveld, 2007; Hempel et al., 2005). The use of electromyography (EMG) to measure activity of muscles involved in the production of emotional expressions (e.g., zygomaticus majoris and the corrugator supercilii) has provided insight into the expressive deficit in individuals with SZ (Kring, Kerr, & Earnst, 1999; Mattes, Schneider, Heimann, & Birbaumer, 1995; Wolf, Mass, Kiefer, Wiedemann, & Naber, 2006). Although visible expressions are reduced in individuals with SZ, Kring and colleagues (1999) found that patients and controls showed similar valence-dependent zygomatic activity (e.g., greater activity in response to positive images) and corrugator activity (e.g., greater activity in response to negative images). Other studies have shown reduced facial muscle activity in response to emotionally-arousing stimuli in patients but that they are activating the muscles in a valence congruent way (Mattes et al., 1995; Wolf et al., 2006). The EMG studies suggest that while patients may not be producing overt emotionally congruent expressions they are engaging the musculature associated with these expressions, albeit at an attenuated level. Other physiological modalities such as breathing rate, heart rate, and systolic blood pressure were employed for studies involving individuals with SZ (Hempel et al., 2007). Moreover, the use of the startle eye blink paradigm has shown a reversed pattern of emotional response of individuals with schizophrenia (Volz, Hamm, Kirsch, & Rey, 2003; Vrana, Spence, & Lang, 1988). Typically, the startle eye blink response is attenuated when paired with a positively valenced stimulus and enhanced when paired with a negatively valenced stimulus.

Despite the benefit of VR for controllability and its affordance of incorporating implicit physiological as well as gaze cues, most of the work involving individuals with SZ has been very limited to sill images and video clips and there is a dearth of literature on application of VR in SZ intervention in general.

The rest of the proposal is organized as follows. Chapter 2 presents the specific aims of this research project given the background presented in this chapter. Chapter 3 presents the adaptive robot-mediated joint attention task. Chapter 4 discusses the emotion recognition task in VR for ASD while Chapter 5 presents the comparative study of VR and static IAPS image for emotion recognition by individuals with SZ. Chapter 6 describes the proposed work including ongoing current work. Chapter 7 briefly highlights the potential contributions of this work followed by general concluding remarks in Chapter 8.

CHAPTER II

SPECIFIC OBJECTIVES

This dissertation research aims at developing intelligent robotic and virtual reality systems for ASD and SZ interventions that not only are sensitive of user performance but also monitor implicit affective cues inferred from peripheral body physiological signals, eye gaze, and EEG. The following sections present the specific objectives of each component of this research project and their current status.

D. Specific Aim 1: An Architecture for Robot-mediated Joint Attention task

To alleviate the lack of an adaptive intelligent robotic system that is focused on core areas of deficit in children with ASD, in this study, we developed and tested a novel closed-loop adaptive robot-mediated architecture capable of both administering joint attention prompts via both humanoid robot and human administrators. We chose early joint attention skills (C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., 2008), since these skills are thought to be fundamental, or pivotal, social communication building blocks that are central to the etiology of the disorder itself (P. Mundy, Rebecca Neal, A., 2000; Poon, 2011). At a basic level, joint attention refers to a triadic exchange in which a child coordinates attention between a social partner and an aspect of the environment. Such exchanges enable young children to socially coordinate their attention with people to more effectively learn from others and their environment. Fundamental differences in early joint attention skills likely underlie the deleterious neurodevelopmental cascade of effects associated with the disorder itself, including language and social outcomes across the ASD spectrum (C. Kasari, Gulsrud, A.C., Wong, C., Kwon, S., Locke, J., 2010; C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., 2008; Poon, 2011). The system automatically provided higher levels of prompts or contingent reinforcement via real-time non-invasive gaze detection as a marker of response. In simpler terms, the system altered its function based on the child's response to the administrator's prompt for joint attention, either by providing an additional prompt, changing the type of prompt or by providing reinforcement. We operationalized response to joint attention as the child's ability to follow an attentional directive to look toward an identified target area. We specifically examined response to joint attention prompts as deficits in social orienting and joint attention are thought to represent core social communication impairments of ASD (P. Mundy, Rebecca Neal, A., 2000; Poon, 2011) and these skills are often targeted in empirically supported intervention paradigms (C. Kasari, Gulsrud, A.C., Wong, C., Kwon, S., Locke, J., 2010; C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., 2008; Yoder, 2006).

The primary objective of this study was to empirically test the feasibility and usability of a closed-loop adaptive robotic system with regards to providing joint attention prompts and within-system adaptation of such prompts. The secondary objective was to conduct a preliminary comparison of child performance between robot and human administrators. We hypothesized (1) that our robotic system could administer joint attention tasks in a manner that would promote accurate orientation to target, and (2) children with ASD would demonstrate increased attention to the humanoid robot compared to the human administrator. We also explored whether young children with ASD would be more accurate with robot prompts than human prompts. We further hypothesized that an appropriately designed robotic system would be able to administer joint attention tasks as well as an experienced therapist, and as a consequence, can be included as a useful ASD intervention tool that, in addition to administering the therapy, can collect objective data and metrics to derive new individualized intervention paradigms.

E. Specific Aim 2: Design of an Intelligent Multimodal Virtual Environment for Facial Emotion Recognition Tasks for ASD and SZ Intervention

This study is a preliminary investigation into the development and application of a VR environment capable of utilizing gaze patterns as well as peripheral physiological signals to understand how adolescents with ASD process salient social and emotional cues in faces. Ultimately, the goal is to alter VR interactions by giving

the environment the ability to respond to patterns of performance, physiological changes and gaze. For individuals with ASD, such enhancements may improve attention to and processing of relevant social cues across dynamic interactions both within and beyond the VR environment.

The major objectives of this work are to: (1) develop an innovative VR-based facial emotional recognition system that allows monitoring of eye gaze and physiological signals, and (2) perform a usability study to demonstrate the benefit of such a system in understanding the fundamental mechanism of emotion recognition. We believe that by precisely controlling emotional expressions in VR and gathering objective individualized eye gaze, physiological responses related to emotion recognition as well as performance data, new efficient intervention paradigms can be developed in the future. The findings in this study may inform future development of affect-sensitive virtual social interaction tasks.

The gaze data as well as the physiological data collected and preprocessed in the study described above were post processed offline. The results from a preliminary comparison of facial affect recognition performance between the ASD and TD samples, and group differences regarding gaze patterns during facial affect detection as well as difference in body physiological signals when the subjects were correctly identifying the emotions versus when the subjects were making mistakes were analyzed. Although explicitly not an intervention study, the ultimate aim of this proof-of-concept and user study was to establish the utility of a dynamic VR and eye-tracking system with potential application for intervention platforms. We hypothesized that participants with ASD would show poor facial affect recognition than the comparison sample, particularly for subtler depictions of affect (e.g., disgust, contempt). We further hypothesized that participants with ASD would have longer response times and less confidence in their recognition decisions. We also hypothesized that participants would show different pattern of physiological responses to the emotional stimuli when they were correctly identifying the emotions versus when they were making mistakes. Finally, regarding eye gaze, participants in the ASD group were expected to attend less frequently to relevant facial features during the task than the control group.

The objective of this project was twofold. First, the VR-based facial emotional expression presentation system that was originally developed for children with ASD was customized with reduced set of emotions for adults with SZ. In addition to dropping the more confused emotions, i.e. disgust and contempt, the data collection was customized for the SZ group. Moreover, an IAPS picture presentation system using the same technology as the VR was developed for side-by-side comparison with emotional expression presentation in VR. Then, a study involving individuals with SZ with age and gender matched control groups was performed to understand emotion understanding both in VR as well as with a static stimuli. Specifically we seek out to answer whether individuals with SZ rate social and non-social emotional images similar in terms of arousal and valence compared to healthy controls and would the presence of dynamic interactive stimuli result in improved recognition performance.

As physiological signals were very well studied with SZ and are essential in understanding the apparent disconnect of what individuals with SZ feel and overt expressions under emotional experiences, the analysis of the physiological signals collected in the context of existent literature was completed. We seek to investigate whether individuals with SZ exhibit similar physiological responses to that of the healthy control, would there be differences in the physiological responses to social and non-social stimuli in both groups in the case of static pictures presentation. Especially, the EMG, GSR, and RSP signals have been thoroughly analyzed. We hypothesized that individuals with schizophrenia will show congruent ratings of their response to the images selected (Berenbaum & Oltmanns, 1992; Doop & Park, 2006; Kring & Neale, 1996). In terms of emotional expressivity, as measured by EMG recordings, we expect individuals with schizophrenia to show less activation of the muscle units involved in the production of emotional expressions (e.g. zygomatic and corrugator muscles). Finally, given the mixed findings in the physiology of emotional responses in individuals with schizophrenia, we do not have specific predictions for their response to the socio-emotional stimuli. We also would like to investigate whether the gaze scanning patterns of individuals with SZ differs from that of the healthy control in terms of both behavioral as well as physiological gaze responses to the static and dynamic stimuli.

F. Specific Aim 3: Design and Evaluation of a Social Interaction Task in Virtual Environment

This project is aimed at designing, developing and testing an innovative adaptive VR-based multimodal social interaction platform. The platform integrates peripheral psychophysiological signals monitoring for affect detection, eye tracking and gaze metrics for engagement based social interaction and spoken question-answer-based dialog management for a more naturalistic interaction.

The VR social task involves an embedded facial emotional expression identification task in the presence of social context, in this case conversational dialog back-and-forth. It also incorporates occlusion of key facial regions so the participant will reveal the face with proper gaze pattern within a specified time. Major emphasis was given for facial emotional expression identification in the presence of social communication and the conversational dialog is a backdrop. This extends our previous study in isolated facial emotional expression identification.

The goal of the system is to provide information on how to teach a participant to properly process an emotional face in the presence of social contexts. We hypothesize that there will be performance differences between the two groups. Evaluation of the effect of gaze pattern and forcing participants to reveal an occluded face in an effort to teach proper gaze patterns, i.e. looking at proper facial regions of interest, is given more emphasis to validate this hypothesis. Two sets of occluded facial emotional expression identification tasks were incorporated at the end of each conversational dialog mission. Pre- and post- isolated facial expression identification tasks without any social cues are also added as a means to measure baseline performance of participants in the facial expression identification task. The system that was already developed for the isolated facial emotional expression, with minor customization, will be used as the pre- and post-baseline test for this task.

The VR social task platform integrates peripheral psychophysiological signal monitoring for affective state modeling, eye tracking and gaze metrics for engagement modeling and spoken question-answer-based dialog management for a more naturalistic interaction. For such a system to simulate some semblance of naturalistic social interaction, utilization and interpretation of several components are required including conversational dialog management, body language (gesture), facial emotional expressions and eye contact in addition to the implicit user state understanding components. Such multimodal interactions help in individualization of the therapy and in cases of inaccessibility of trained therapists, it may serve as a self-contained therapeutic system.

The peripheral affective recognition and EEG data processing will be performed offline to determine underlying differences in emotional states between and within groups of the children with ASD. Visual scanning patterns as well as physiological and mental responses would be analyzed to see differences between the two ASD groups and within each group.

CHAPTER III

ROBOT-MEDIATED JOINT ATTENTION FOR ASD

This chapter describes the development as well as results from a usability study of a robot-mediated joint attention presentation frame work for children with ASD. This work is mainly focused on (1) developing an individualized and adaptive robotic therapy platform that is capable of administering joint attention therapy on its own; and (2) conducting a usability study that investigates the potential of such a robot-mediated ASD intervention as compared to a similar human therapist-based intervention. First, we designed a closed-loop adaptive robot-mediated ASD intervention architecture called ARIA wherein a humanoid robot works in coordination with a network of spatially distributed cameras and display monitors to enable dynamic closed-loop interaction with a participant, and second, we performed a usability study with this system and compared ARIA performance with that of a human therapist. We believe that such a comparative study with a robotic system in the context of response to joint attention task has not been performed before.

Though it was beyond the scope of this current study, it is our hope that eventually robot-mediated intervention can be used to teach skills and those skills can be transferred to real-world situations using co-robotic paradigms where a human caretaker will be integrated within a robotic interaction framework.

G. System Development

Architecture Overview.

A humanoid robot is the central component of the Adaptive Robot-mediated Intervention Architecture (ARIA) (Fig. 1), which is capable of performing many novel embodied actions. It presented joint attention prompts adaptively using gestures (pointing to the target by hand), gaze/head shifts, voice commands and audiovisual stimuli. The system had two computer monitors hung on specific, but modifiable locations of the experiment room capable of providing visual and auditory stimuli. The stimuli included static pictures of interest for the children (e.g., pictures of children characters), videos of similar content, or discrete audio and visual events (i.e., specific sound/video in the form of additional triggering if child did not respond to robot cues alone). These stimuli could be adaptively changed in form or content to provide additional prompt levels based on the participant's response and also served as an object of potential reinforcement and reward (i.e., contingent activation).

Participants' gaze as approximated by the head pose, in response to the joint attention prompt was inferred in real-time, by a network of spatially distributed cameras mounted on the walls of the room. Individual gaze information was fed back to a software supervisory controller that coordinated the next robotic system action.

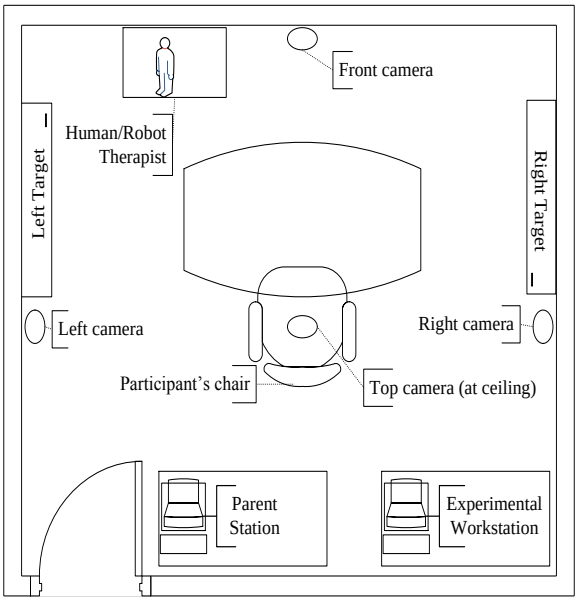


Fig. 1. The experiment room setup.

The adaptive and individualized closed-loop dynamic interaction preceded for a set of fixed duration trials. Fig. 1 shows the experimental room setup and how the cameras and the monitors that presented stimuli were placed in reference to the participant.

The robot and each camera had their own processing modules. The system was developed using a distributed architecture with a supervisory module (i.e., a software supervisory controller) making decisions about task progression. The participant wore a hat with an array of infrared (IR) LEDs sewn to its top and sides in straight lines. This hat was used in conjunction with the network of cameras to infer eye gaze of the participant. The supervisory controller sent/received commands and data to and from the robot as well as the cameras. Individual modules ran their own separate processes. The supervisory controller facilitated communication between the camera processing modules (CPM), the humanoid robot, and the stimuli controllers (SC) using a network interface as shown in Fig. 2. It also made decisions based on performance metrics computed from the sensory data collected from CPMs.

The CPMs were responsible for processing the images from the IR camera and measured approximate gaze directions in their respective projective views. The stimuli controllers controlled presentation of stimuli on the audiovisual targets. A sensory network protocol was implemented in the form of a client-server architecture in which each CPM was a client to a server that was monitoring them for raw time-stamped tracking data upon trigger from the supervisory controller.

The server accumulated the raw data for the duration of a trial, produced measured performance metrics data at the end of each trial and sent these data to a client embedded in the supervisory controller. Another server embedded in the supervisory controller generated feedback and communicated with the robot (NAO) and the stimuli controllers. Each stimuli controller had an embedded client. Communication with the robot was done using remote procedure call (RPC) of modules instantiated in the robot via a proxy using simple object access protocol (SOAP).

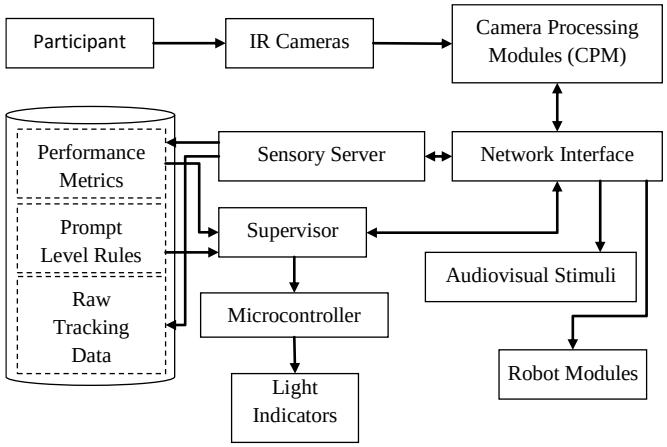


Fig. 2. System schematics for closed-loop interaction.

The Humanoid Robot, NAO.

The humanoid robot used in this project was NAO (Fig. 3), which is made by Aldebaran Robotics (www.aldebaran-robotics.com). NAO is a medium child-sized humanoid robot with a height of 58 cm and weight of approximately 4.3 kg. Its body is made of plastic and it has 25 degrees of freedom (DOF). It is equipped with software modules that enable RPC and encourage distributed processing. NAO's architecture is based on a programming architecture called NAOqi that is a broker process for object sharing and is used to attach modules developed for specific tasks. This allows to port out computationally intensive algorithms into a remote machine other than the robot.

In this work, we augmented NAO's vision using a distributed network of external IR cameras as part of the head tracker for closed-loop interaction. NAO's two CMOS vertical stereo camera sensors are low performance CMOS sensors with frame rates of 4–5 frames per second (FPS) with the native resolution of 640 x 480, which were not suitable for our task in detecting head movement in real-time. Fig. 2 shows how we

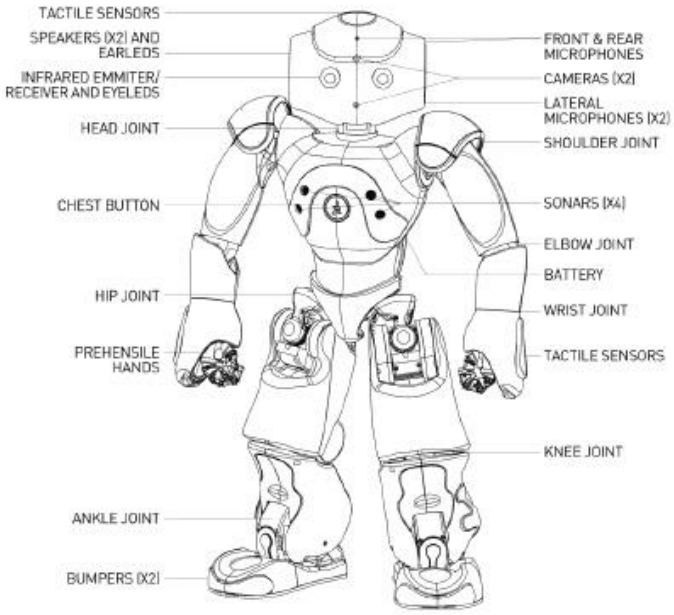


Fig. 3. The Humanoid Robot NAO

designed the communication between different parts of the system and the participant that constituted the closed-loop real-time interaction. To meaningfully compare the performance of the robot and a human therapist, we used the head tracker for both the human therapist as well as the robot. Light indicators, which were controlled by a microcontroller, were used to indicate success or failure for the human therapist. They were located behind the participant and in front of the human therapist.

Rationale for the Development of a Head Tracker.

In responding to joint attention (RJA) tasks, the administrator of the task is interested in knowing whether the child looks at the desired target following a prompt (P. Mundy, Block, J., Delgado, C., Pomares, Y., Van

Fig. 4. The hat holds the array of LEDs at side and top.

Hecke, A.V., Parlade, M.V., 2007). Gaze inference is usually performed using eye trackers. However, such systems often place restrictions on head movement and detection range. They are also expensive and are generally sensitive to large head movements, especially the type of head movements needed for RJA tasks. Moreover, the eye trackers need calibration (often multiple times) with each participant and the range of the eye trackers typically requires that the participants be seated within 1-2 feet from the tracker/stimuli. Joint attention tasks, on the other hand, require large head movements and the objects of interests are typically placed between 5-10 feet from the participants. While it is necessary to know whether the child is looking at a target object in joint attention tasks, unlike typical eye tracking tasks the need to know the precise gaze coordinates is less important in JA tasks. Commercial marker based head trackers (such as TrackIR4, and inertial cubes) were initially considered. However, they are specifically designed for certain applications (e.g., video games) and their tracking algorithm is not flexible enough for JA task, which requires very large head movements around the room. As a result, we decided to develop a low cost head tracker that would not be limited to large head movements.

In order to determine whether the child is responding to RJA commands, a near-IR light marker-based head tracker (Fig. 4) was designed using low cost off-the-shelf components for gaze inference with an assumption that gaze direction can be inferred by knowing the head direction using projective transformations.

The Head Tracker.

The head tracker was composed of near-infrared cameras, arrays of IR light emitting diodes (LEDs) sewn on the top and the sides of a hat as markers, and the camera processing modules (CPMs) (Fig. 4). The tracker had a set of CPMs that processed input images captured by the top and side cameras that were tracking the arrays of top and side LEDs on the hat, respectively.

The Head Tracker has the following components:

1) IR Cameras: The IR Cameras were custom modified to monitor only in near-infrared from inexpensive Logitech Pro 9000 webcams. The IR filter for the original camera was removed carefully and was replaced with a glass of similar refractive index and thickness.

2) Camera Processing Module (CPM): These modules were equipped with contour-based image processor in the XYZ color space and simple decision tree like rule-based object detector given geometric and color characteristic features. The XYZ space is characterized by identifying brightness cues better than other color spaces. This was a suitable property for detection of LEDs in the IR spectrum.

The camera matrix (hence, the focal length) and distortion coefficients were estimated using the Levenberg–Marquardt optimization algorithm (More, 1978). A camera calibration routine was performed with the common square grid fiducial.

There were two stages of the camera processing modules: offline preprocessing stage and an online detection stage (Fig. 5). The offline preprocessing stage included the camera calibration and the offline manual feature selection. Regions of interests (ROI) containing only LEDs were manually selected. The input images

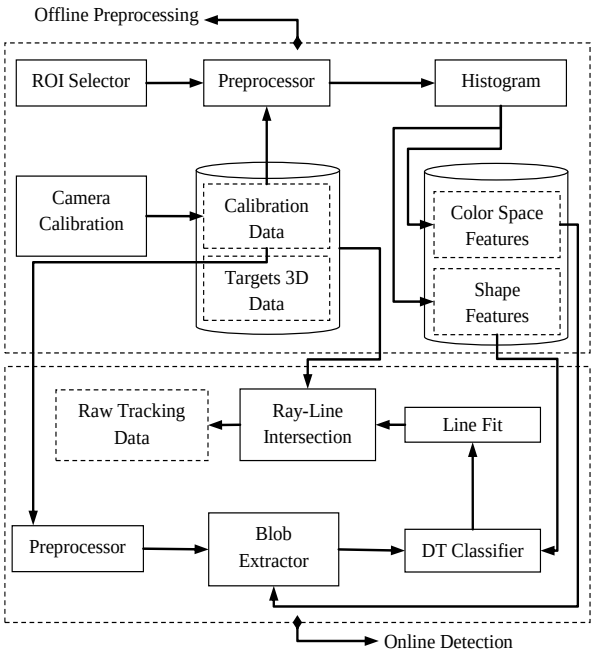


Fig. 5. Camera Processing Module (CPM) schematics.

were then masked by the selected ROI. Preprocessing (such as noise filtering, un-distorting using the camera distortion coefficient) was performed before extracting features. Then, color space histogram statistics (such as minimum, mean, maximum in each of the 3 color channels) and geometric shape features (such as area, roundness, and perimeter) were computed inside the regions of interest. These features were then stored for online real-time detection stage.

In the online detection stage, the input images were un-distorted using the distortion coefficients and then a 9 x 9 Gaussian smoothing filter was applied to remove artifacts. The blob extractor applied the color space features in each channel of the XYZ color space to obtain candidate blobs. The resultant single channel images were logically combined so that all qualifying blobs satisfied all the color space constraints. To further remove artifact blobs, extreme geometric feature pruning and morphological opening were performed. The candidate blobs were then passed through a simple classifier, which classified the blobs as those belonging to LED blobs and those belonging to the background using the geometrical features. A linear function was fit to those blobs that qualified the classification, which was then extended indefinitely as a ray. Finally, a line segment (projection of the targets) and ray (extension of the fitted line) intersection test was performed to approximate the gaze of the participant in the respective projective plane. When the child moved his/her head so did the LEDs arrays and so did their 2D projections. Hence the ray moved around the projection plane. The targets were at rest relative to the camera's frames of references. The intersections of rays with the projections of the targets in the top and side projection planes gave the *x* and *y*-coordinates of the approximate gaze point, respectively, on the corresponding image planes. Therefore, a ray-line segment intersection could be performed to compute the intersection coordinate in the respective dimensions. A brief description of the ray-line segment intersection algorithm used in this work is given in (Bekele, Lahiri, Davidson, Warren, & Sarkar, 2011) and Appendix A. Note that this system considered yaw and pitch angles of the head

movement but not the roll, which was considered less important for this study.

H. Usability Study

To test and verify the feasibility of the ARIA system, a usability study was designed based on common paradigms for indexing early joint attention and social orienting capacity, such as the Early Social Communication Scales (ESCS) and the Autism Diagnostic Observation Schedule (ADOS) (C. Kasari, Freeman, & Paparella, 2006; P. Mundy et al., 2003). We developed the behavior protocol used in this study to mirror standard practice in assessment and treatment of children with autism regarding early joint attention skills. Consistency in prompt procedures across robot and human administrators was of primary importance given our aim to determine whether a robot administrator could carry out a joint attention task similar to a human administrator. We therefore used only one prompting method. We chose a least-to-most prompt (LTM) hierarchy (Demchak, 1990), which essentially provides support to the learner only when needed. The method allows for independence at the outset of the task with increasing help only after the child has been given an opportunity to display independent skills. Prompt levels arranged from least-to-most supportive are also commonly used in 'gold-standard' diagnostic or screening tools (e.g., ADOS, STAT, and ESCS). The LTM approach allowed us to identify the lowest level of support needed by each child to emit the correct response (i.e. looking toward the joint-attention target). Our paradigm, adopting the least-to-most prompting systems of these common assessments, involves an administrator cuing a child to look toward specific targets and adding additional cues or reinforcement based on performance.

The child sits facing the human therapist or the robot administrator in their respective sessions and the human therapist and the robot presented the joint attention task using the hierarchical prompt protocol described below. It should be noted that this is a preliminary usability study using a well-accepted prompting strategy commonly used in ASD assessment and intervention; but the study is not an intervention. Future study applying the system and indexing learning over time would be a natural extension to this comparative behavioral study.

Hierarchical Prompt Protocol.

A hierarchical prompt protocol was designed in a least-to-most fashion with higher level of prompt provided on a need basis. In this robot-mediated joint attention task, the robot or the human therapist initiated the task (IJA) using speech as a first level of interaction. They cued the participant to look at a specific picture that was being displayed on one of the two computer monitors hung on the wall. The next higher prompt level adds pointing gesture to the target by the robot or the human therapist. The next higher prompt level adds an audio prompt at the target on top of the pointing gesture cue and the verbal cue. The final level cue adds a brief video in addition to pointing gesture and the verbal cues. The audio and video prompts at the target were not directly addressing the child rather they are novelty added to attract the child attention. There were 6 levels of prompts in each trial. These hierarchical prompts were administered stepwise if no or inappropriate response was detected and progressed in least-to-most fashion (Table 1). Exact similar prompts protocol was followed by both the human therapist as well as the robot. The way they were presented could, however, be slightly different as the robot might not be capable of producing all subtle non-verbal cueing exactly as the human therapist. For instance, the non-biological (robotic) movement of the head and the arms were starkly different from the smooth human limb motions. Moreover, the robot has limitations on its eyes and eye gaze was approximated by a head turn on the robot. The LEDs on the robot's eyes could have been toggled to give a sense of direction but it would be unnatural and it would make comparison to a human therapist compounded.

The head tracker was triggered by the supervisory controller as the robot issued a prompt to activate the camera processing modules for the specified trial duration to accumulate time-stamped data of the head

Table 1. Levels of the Hierarchical Protocol

Prompt Levels (PL)	Robot Prompt for a child named Max
PL 1	"Max, look" + shift (robot's) head to target
PL 2*	If NR$ after 5 s : "Max, look" + shift head to target
PL 3	If NR after 5s: "Max, look at that." + head shift + point to target
PL 4*	If NR after 5s: "Max, look at that." + head shift + point to target
PL 5	If NR after 5s: "Max, look at that." + head shift + point to target + audio clip sounds at target
PL 6	If NR after 5s: "Max, look at that." + head shift + point to target + audio clip sounds at target and then video onset for 30s

*PL2 and PL4 are intentionally repeated versions of PL1 and PL3, respectively. $NR means no or inappropriate response.

movement so as to infer where the participant was looking. At the end of the trial duration, several performance metrics were computed and sent to the supervisory controller. These performance metrics were used to generate rewards to be executed by the robot. For example, if the response was correct the robot would say "Good Job!" and encouragement (e.g., static picture shown on the monitor turned into a movie clip). On the other hand, if there was no response or incorrect response, the robot issued the next level of prompt.

The pictures, audio, and video clips were carefully selected from children's TV programs such as Bob the Builder, Dora the Explorer, etc. Segmented clips of characters performing specific actions from six age appropriate preschool shows characters were selected for inclusion. Clips selected were components of these shows wherein a dance, performance, or action 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 patient population. Within each show, videos were selected to be part of the prompt and feedback. Each video was selected so that it has short duration action filled segment so as to be used for the six level of prompting. The videos for the reinforcement were selected to make the reinforcement enjoyable to the participant. The reinforcement videos contained segments that are filled with play, joy and multiple characters. The prompting videos were short segment (5 seconds) compared to a longer reinforcement segment of 30 seconds. In almost all of the videos, the characters are not directly addressing the participant. The audio/video played on the targets was used only as a simple attention grabbing mechanism rather than an additional source of prompts on their own. Each audio and video clip played was directly matched with the initial picture displayed on the presentation targets. This means, if initial target was a picture of Scooby-doo, the audio and video prompt as well as the reinforcement videos would be that of Scooby-doo with the only difference of multiple characters (people) in a single segment in the case of the reinforcement video. The target monitors were 24 inch (58 cm x 36 cm). Each monitor was placed at a horizontal and vertical distance of 148 cm and 55 cm, respectively, as seen from the top camera reference frame and they were hung at a height of 150 cm from ground and 174

cm from the top camera (ceiling) (Fig. 1).

The Joint Attention Task and Procedure.
Each participant took part in one session of experiment lasting approximately 30 minutes. A typical session consisted of four sub-sessions of each 2–4 minutes long: two human therapist and two robot therapist sub-

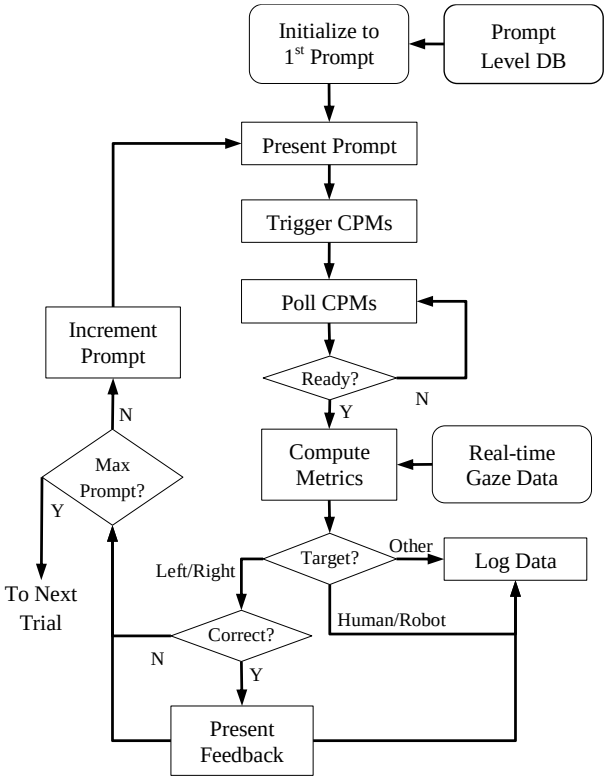


Fig. 6. Flowchart of the prompt-feedback mechanism for one trial.

sessions interleaved (H-R-H-R or R-H-R-H). Each sub-session contained four trials, with two trials randomized per target. Each trial was 8 s long with 3-5 s of prompt and an additional 3-5 s of monitoring interval. Fig. 6 presents the flow of the process.

In the beginning, a researcher (a trained therapist) described the procedures and the tasks involved in the study to the participant and his/her parent(s) and obtained written parent consent. An assent document was then read for children above 3 years of age. The child was instructed to tell the researcher or parents if he or she became uncomfortable with the study and could choose to withdraw at any time. The participant was then seated and buckled into a Rifton chair to start the first sub-session. The child remained buckled in the chair during sub-sessions (approximately 2–4 minutes long) but encouraged to take a break from the chair between sub-sessions. For the first 3 participants of both groups (children with ASD and typically developing children), the robot was presented first while for the rest of participants in each group, the human therapist started out the session. The prompt presenter (either the robot or the human therapist) first cued the participant just verbally. For example, if the participant's name was Max, "Max, Look!" (Table 1). If the participant did not respond, the robot hierarchically increased the prompting level to include pointing to the target as well as audio and video trigger for 5 seconds (Table 1).

Table 2. Profile of the TD participants in robot-mediated JA study

Participant (Gender)	Age	SRS (cutoff=60)	SCQ (cutoff=15)
TD1 (f)	4.72	47	2
TD2 (m)	5.27	39	0
TD3 (m)	4.74	45	2
TD4 (f)	3.20	46	5
TD5 (m)	5.18	50	11
TD6 (m)	2.46	46	3
Average (SD)	4.26 (1.05)	45.5 (3.3)	3.83 (3.53)

Subjects.

A total of 12 participants (two groups: children with ASD and a control group of typically developing (TD) children) completed the tasks with their parents' consent. Initially, a total of 18 participants (10 with ASD and 8 TD) were recruited. Six TD (m: n=4 and f: n=2) of age 2–5 y (M=4.26 y, SD=1.05 y) and six participants with ASD (m: n=5 and f: n=1) of age 2–5 y (M=4.7 y, SD=0.7 y) successfully completed the study. The details of the TD and ASD group who completed the study are given in Table 2 and Table 3.

Out of the total 10 ASD and 8 TD children, 4 ASD and 2 TD participants were not able to complete the study. Three out of the four subjects with ASD withdrew because they could not tolerate wearing the hat. One was distressed during the interaction and as such the session was discontinued. Two typically developing children were not willing to participate after parent consent. Tables 3 shows the profile of the participants with ASD who completed the study. All participants with ASD were recruited through existing clinical re-

Table 3. Profile of the participants with ASD in robot-mediated JA study

Subject (Gender)	ASD1 (m)	ASD2 (m)	ASD3 (f)	ASD4 (m)	ASD5 (m)	ASD6 (m)	Average (SD)
Age	5.14	3.24	4.92	5.27	4.49	5.17	4.70 (0.7)
ADOS-G (cutoff=7)	10	13	14	25	25	9	16.0 (6.58)
ADOS CSS (cutoff=8)	6	6	7	10	10	5	7.33 (1.97)
ADOS-G RJA	0	1	1	2	3	0	1.17 (1.07)
SRS (cutoff=60)	51	58	70	85	81	77	70.33 (12.24)
SCQ (cutoff=15)	5	11	8	21	20	15	13.33 (5.91)
IQ	101	54	102	50	49	73	71.5 (22.65)

search 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 2–5 years of age, had to have an established diagnosis of ASD based on the gold standard in autism assessment, the Autism Diagnostic Observation Schedule Generic (ADOS-G) (Lord et al., 2000), as well as participants' willingness (and parents' consent) and the physical and mental ability to adequately

perform the tasks.

In order to ensure that TD participants did not evidence ASD-specific impairments and to have a quantification of ASD symptoms across groups, TD parents completed ASD screening/symptom measurements (Table 3): the Social Responsiveness Scale (SRS) (Constantino & Gruber, 2002) and the Social Communication Questionnaire (SCQ) (Rutter, Bailey, Lord, & Berument, 2003). Efforts were made to match the groups with respect to age and gender whenever possible.

I. Results

Validation Results.

First, the head tracker was validated using an infrared (IR) laser pointer. Validation of the head tracker was performed by attaching an infrared laser pointer to the hat and aligning it in such a way that the direction of the laser pointer approximated the gaze direction. Projection of the laser pointer onto one of the targets (computer monitors) was recorded using a 2 cm x 2 cm grid displayed on the monitor. Twenty such validation points were uniformly distributed across the screen and 10 seconds of data was recorded and averaged at each point. It was found that the head tracker had average validation errors of 2.6 cm (1.2^0) and 1.5 cm (0.7^0) in x and y coordinates at approximately 1.2 m distance, respectively, with good repeatability. Since joint attention requires large head movements, an average 2.6 cm error was acceptable (for details, refer (Bekele et al., 2011)).

The overall system was also tested with two typically developing children, ages two and four years. The system worked as designed and all the data were logged properly and the robot could administer the JA task accurately. This preliminary testing helped to fix issues and made the system ready for the usability study.

Preferential Looking Towards the Robot.

The results of this study indicated that children in both ASD and TD groups spent more time looking at the robot than the human therapist. A paired dependent t-test statistic was performed for statistical significance analysis of the results. The children in the ASD group looked at the robot therapist for 52.76% (17.01%)[1] of the robot sub-session time (Fig. 7). By comparison, they looked at the human therapist for 25.11% (11.82%)[1] of the human sub-session time. As such, the ASD grouped looked at the robot therapist 27.65% longer than the human therapist, $p < 0.005$. Fig. 8 shows individual comparison of the time children in the ASD group spent looking at the therapist in the robot and human sub-sessions, respectively. The children in the TD control group looked at the robot therapist for 54.27% (17.65%)[1] of the robot sub-session time. By comparison, they looked at the human therapist for 33.64% (16.04%)[1] of the human therapist sub-session time. As such the children in the TD group looked at the robot therapist for 20.63% longer than the human therapist, $p < 0.005$. Fig. 9 shows individual comparison of the time spent looking at the therapist in the robot and human sub-sessions, respectively, for the TD group.

The results indicated a statistically significant preferential orientation towards the robot as compared to the human therapist for both groups. Individual and group comparisons indicate that the children spent more time looking at the robot than human therapist. This difference was most pronounced for children in the ASD group as this group looked at the human therapist for a smaller percentage of human sub-sessions than did those children in the TD group (The TD group looked 8.53% more time on the human therapist than the children in the ASD group).

[1] The format is Mean (Standard Deviation) throughout the paper.

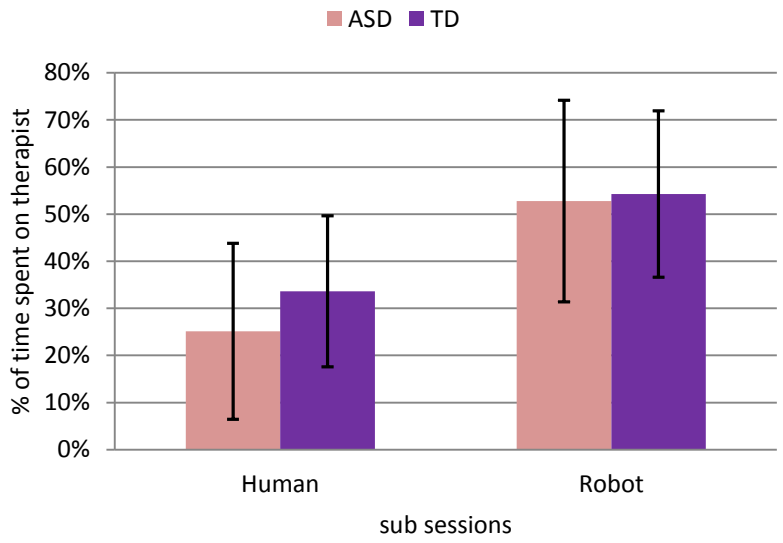


Fig. 7. Across group comparison for the time spent on the human and robot therapists.

The robotic system together with the dynamic target stimuli was able to perform the joint attention task administration with 95.83% success to the ASD group and with 97.92% success to the TD group. Success was measured as a percentage of successful trials (trials that resulted in eventual response to the targets at any of the six levels of prompt) out of the total available trials. The human therapist was able to achieve 100% success for both groups. Taken together with preferential looking data, the robot was able to perform the task with success rates similar to that of the human therapist with the potential advantage of increased engagement. Success to attend to the target before the dynamic audio/video target stimuli was also measured to rate the performance of the robot in administering the trials. The robot was able to result in success 77.08% for the ASD group and 93.75% for the TD group while the human therapist was able to achieve successes of

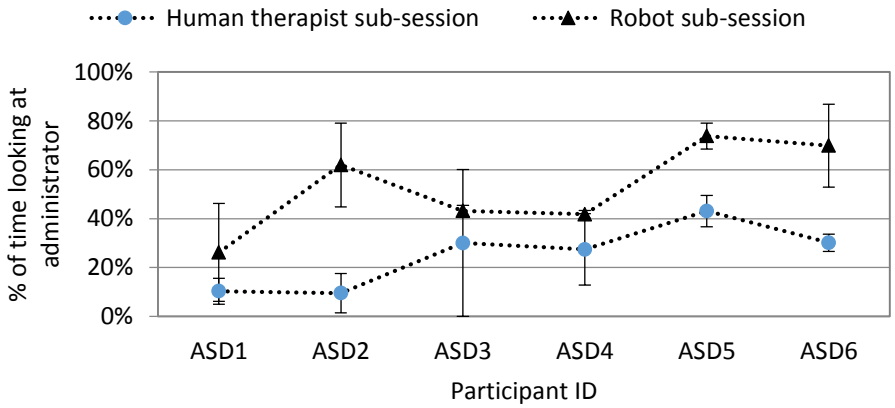


Fig. 8. Percentage of time spent looking at robot and human therapists during respective sub-sessions for the ASD group.

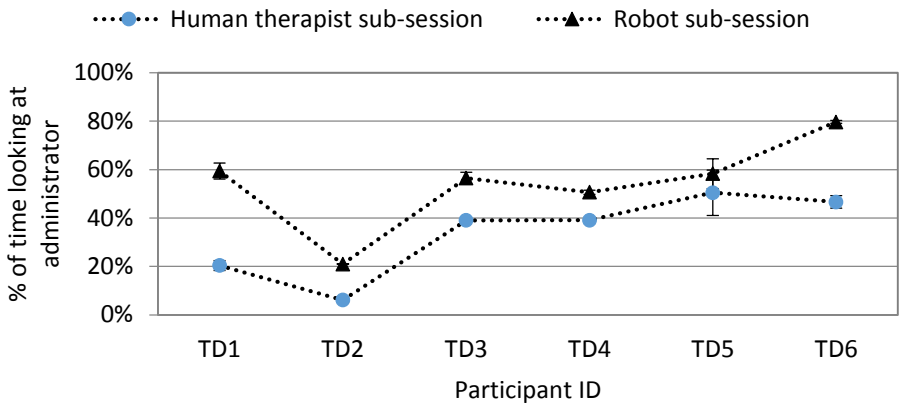


Fig. 9. Percentage of time spent looking at robot and human therapists during respective sub-sessions for the TD group.

93.75% and 100% for the ASD and TD groups, respectively. This shows that the robot alone resulted in lower success rate, which highlights the importance of the dynamic stimuli presented on the targets themselves. Relatively, this is not the case for the control group.

Frequency of Looking to Target and Number of Required Levels for Success.

Objective performance metrics were computed to quantitatively measure: "hit frequency" (i.e., frequency of looks to the target after the prompt was issued within a trial) and "levels required for success" (prompt level required for success hitting target), expressed as prompt level/total levels in a given sub-session.

Too large a hit frequency is undesirable as it indicates erratic gaze movement rather than engagement. Too small a hit frequency might be considered a "sticky" attention scenario (Landry & Bryson, 2004). The children in the ASD group showed a hit frequency of 2.06 (0.71)[1] in the human therapist sub-session and 2.02 (0.28)[1] in the robot sub-session, which were comparable.

Children in the TD group showed hit frequency of 2.17 (0.65)[1] in the human therapist sub-session and 1.67 (.35)[1] in the robot sub-session. There was no significant difference in hit frequency for ASD and TD groups, $p > 0.1$.

Levels required for success measures the realization of the prompt and the ability to respond to initiation of joint attention. Children in the ASD group required 14.58% more of the total number of levels available in the robot sub-sessions than in the human therapist sub-sessions. Children in the TD group required 9.37% more of the total number of levels available in the robot sub-sessions than in the human therapist sub-sessions. Both differences for the two groups were statistically significant, $p < 0.05$ (Fig. 10).

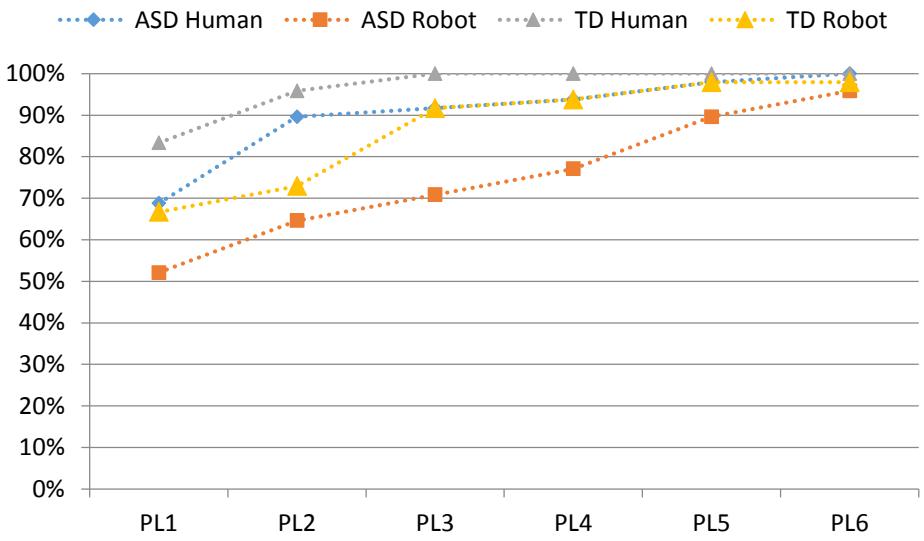


Fig. 10. Percentage of correct responses achieved at specific prompt levels by group and condition (each point represents percentage of correct response at that prompt level and below).

As seen above, preferential looking toward robot was statistically significant for both groups. Further, a trained observer and a parent completed ratings of behavioral engagement. Both observers reported that most children appeared excited to see the robot and its actions. The higher number of levels required for success in robot sub-sessions may also be due to this attentional bias towards the robot.

General Comparisons of ASD group with TD group.

Generally, both groups spent more time on the robot as opposed to the human therapist sub-sessions ($p < 0.005$, both groups), exhibit a higher latency before first hit on the robot sub-session than the human therapist ($p < 0.05$, ASD, $p < 0.005$, TD), lower percentage of time spent on the targets in the robot sub-session than the human therapist sub-session ($p < 0.05$, ASD, $p < 0.005$, TD), higher average number of levels were required in the robot sub-session than the human therapist sub-session ($p < 0.05$, both groups)

In the human therapist sub-sessions, children in the TD group looked towards the human therapist more frequently than did children in the ASD group. In the robot sub-sessions, children in the TD group looked towards the robot at roughly equivalent rates as the ASD group; however, they looked towards the human therapist 8.53% more than children in the ASD group. This suggests excellent engagement with the robot regardless of group status. As expected, human therapists garnered the attention of children in the TD group more than children in the ASD group. However, the difference was not statistically significant with $p > 0.1$ (independent two-sample unequal variance test).

Regarding levels required for success, the children in the TD group required fewer prompt levels for success in general than children in the ASD group. In the robot sub-session the TD group required 11.46% fewer prompt levels than ASD group while they required 6.25% fewer prompt levels than the ASD group on average in the human therapist sub-session.

Generally, children in the ASD group attended more to the robot and were less attracted by the targets than their TD counterparts. The relatively higher number of prompt levels required for children with ASD might be best attributed to their attention bias for the robot than the nature of the disorder itself. The relatively small and insignificant differences in these performance metrics for children in the ASD group may be explained

by participant characteristics (i.e., mostly high functioning with closer to normal RJA scores. See Table 3 for the participants' profiles).

J. Discussion and Conclusion

In the current pilot study, we studied the development and application of an innovative closed-loop 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 feasibility and usability of a robotic system capable of intelligently administering joint attention prompts and adaptively responding based on within system measurements of performance. We also conducted a preliminary comparison of child performance across robot and human administrators to evaluate the main hypothesis of this study. This comparison was performed for children in both groups. Since this study is mainly focused on demonstrating the long theorized intrinsic interest, that is children in general and children with ASD in particular show preferential interests toward robots, comparing the two groups in any of the performance metrics is discussed tangentially. Hence, the main focus is on comparing how the robot performs as compared to the human administrator.

Both TD and ASD children spent more time looking at the humanoid robot and were able to achieve a high level of accuracy across trials. However, across groups children required higher levels of prompting to successfully orient within robot administered trials. This may be due to novelty of the robot, which might decline over time and should further be examined in a larger study. No specific data suggested that ASD children exhibited preferences or performance advantages within system as compared to their TD counterparts within system. These results were not statistically significant as such except the case of target success before target activation. Preschool children with ASD looked away from target stimuli at rates comparable to typically developing peers.

Children with ASD and typically developing children were able to ultimately respond accurately to prompts delivered by a humanoid robot and a human administrator within the standardized protocol. Children with ASD also spent significantly more time looking at the humanoid robot than the human administrator, a finding replicating previous work suggesting attentional preferences for robotic interactions over brief intervals of time (K. Dautenhahn et al., 2002; Duquette, 2008; Kozima, 2005; Michaud, 2002; B. Robins, Dautenhahn, & Dickerson, 2009). Further, children with ASD displayed comparable levels of gaze shifts during correct looks, suggesting that differences in attention were likely not simply a reflection of atypically focused gaze toward the technological stimuli or random looking. In terms of tolerability, we anticipated a certain, if not large fail rate, across the ASD sample in terms of willingness to wear the LED cap even for a brief interval of time (i.e., less than 15 minutes). The completion rate of 60% for the ASD group was promising, but ultimately highlights the need for the development of non-invasive systems and methodologies for realistic extension and use of such technologies with a young ASD population with common sensory sensitivities (Rogers & Ozonoff, 2005). Likewise, such a non-invasive system may help overcome the challenges that of head-tracking methodologies for marking and approximating gaze. In the current protocol targets were placed out of peripheral range of vision to necessitate head movement for tracking; however, precise gaze detection would afford for more robust systems and methodologies in future investigations.

Collectively these findings are promising in both supporting system capabilities and potential relevance of application. Specifically, preschool children with ASD directed their gaze more frequently toward the humanoid-robot administrator, they were very frequently ultimately capable of accurately responding to robot administered joint attention prompts, and they were also looking away from target stimuli at rates comparable to typically developing peers. This suggests that 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 capable of taking advantage of baseline enhancements in non-social attention preference (Annaz et al., 2012; A. Klin, Lin, D.J., Gorrindo, P., Ramsay, G., Jones, W., 2009) in order to meaningfully enhance skills related to coordinated attention. The current system only provides a preliminary structure for examining ideal instruction and prompting patterns.

Future work examining prompt levels, the number of prompts, cumulative prompting, or a refined and condensed prompt structure would likely enhance future applications of any such robotic system.

While our data provides preliminary evidence that robotic stimuli and systems may have some utility in preferentially capturing and shifting attention, at the same time both children with ASD and TD children required higher levels of prompting with the robot administrator when compared to a human administrator in the current study. It is entirely plausible that such differences were related to unclear or suboptimal instructions within system or initial naïve response patterns of children, given that children had no previous exposure to an unfamiliar robot and copious exposure to human directives, prompts, and bids. In this context children had to figure out what the robot was doing and, in turn, expecting them to do. The finding that this was both the experience of the TD and ASD children lends potential support to this explanation. If this was the case, improvements in performance over time might be seen with a refined robot system, including optimized prompts and instructions, and could yield greater success over time based on preferential attention. However, it is also entirely plausible that such a difference highlights the fact that humanoid robotic technologies, in many of their current forms, are not as capable of performing sophisticating actions, eliciting responses from individuals, and adapting their behavior within social environments as their human counterparts (K. Dautenhahn, 2003; Diehl et al., 2011). Though NAO is a state-of-the-art commercial humanoid robot, its interaction capacities have numerous limits. Its limb motions (driven by servo motors) are not as fluid as human limb motions, it creates noise while moving its hand that is not present in the human limb motion, and flexibility and degrees of freedom (DOF) limitations produce less precise gestural motions, and its embedded vocalizations have inflection and production limits related to its basic text-to-speech capabilities. In fact, our data ultimately suggest that children fundamentally performed best with human prompting across all trials as compared to this type of humanoid robotic interaction. As such, these data suggest that it is unlikely that the mere introduction of a humanoid robot that performs a simple comparable action of a human in isolation will drive behavioral change of meaning and relevance to ASD populations. Robotic systems will likely necessitate much more sophisticated paradigms and approaches that specifically target, enhance, and accelerate skills for meaningful impact on this population. Closed-loop technologies (D. Feil-Seifer, Mataric, M., 2011; C. Liu et al., 2008) that harness powerful differences in attention to technological stimuli, such as humanoid robots or other technologies may hold great promise in this regard.

There are also several methodological limitations of the current study that are important to highlight. The small sample size examined and the limited time frame of interaction are the most powerful limits of the current study. As such, while we are left with data suggesting the potential of closed-loop 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. Eventual success and clinical utility of robot-mediated systems hinges upon their ability to accelerate and promote meaningful change in core skills that are tied to dynamic neurodevelopmentally appropriate learning across environments. We did not systematically intend to assess learning within this system; rather we indexed simple initial behavioral responses within system application. As such, questions regarding whether such a system could constitute an intervention paradigm remain open. Further, the brief exposure of the current paradigm, in combination with unclear baseline skills of participating children, ultimately cannot answer questions as to whether the heightened attention paid to the robotic system during the study was simply the artifact of novelty or of a more characteristic pattern of preference that could be harnessed over time. Further study with large sample size is needed to determine (1) whether the preferential gaze towards the robot is from the novelty of the robot; and (2) ways of employing this attention and diverting these towards the target over time for meaningful skill learning. In addition, the robots non-biological limb movements and inability to move the eyes separately from the head were some of the limitations that might inhibit exact comparison to a human therapist.

Another important technical limitation was the approximation of gaze with 3D head orientation. It must be emphasized that head orientation approximating gaze does not necessarily equate to actual eye gaze especially when the children were peering. However, as discussed above, targets were placed at the right and left

extremes necessitating the participants to move their heads significantly to attend the targets. The requirement to wear a hat was also a major limitation with 33% dropout rate overall. Though, this dropout rate is similar or less than minimally-invasive clinical devices such as physiological monitoring devices, it highlights the need to develop a non-contact remote eye gaze tracker. This will also solve the issues related to peering. Finally, although we made attempts to ensure that children with ASD had received evaluations with gold-standard assessment tools (e.g., ADOS, clinician diagnosis), we did not have rigorous assessment data on the comparison sample on these same instruments. As such, our ability to comment on the specific clinical characteristics matched with performance differences regarding this technology is limited. Moreover, we did not perform a baseline assessment before the robot and human therapist sessions. Instead, we used the RJA section score of the ADOS as indicative of the joint attention skill for the children in the ASD group.

Despite limitations, this work is the first to our knowledge to design and empirically evaluate the usability, feasibility, and preliminary efficacy of a closed-loop interactive robotic technology capable of modifying response based on within system measurements of performance on joint attention tasks. Few other existing robotic systems (D. Feil-Seifer, Mataric, M., 2011; C. Liu et al., 2008) 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 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.

CHAPTER IV

EMOTION RECOGNITION AND FACIAL PROCESSING IN VR FOR ASD

In this work, we present a detailed description of the VR system, results from a preliminary comparison of facial affect recognition performance between our ASD and TD samples, and group differences regarding gaze patterns during facial affect detection. Although explicitly not an intervention study, the ultimate aim of this proof-of-concept and user study was to establish the utility of a dynamic VR and peripheral monitoring system with potential application for intervention platforms. We hypothesized that participants with ASD would show poor facial affect recognition than the comparison sample, particularly for subtler depictions of affect (e.g., disgust, contempt). We further hypothesized that participants with ASD would have longer response times and less confidence in their recognition decisions. Finally, regarding eye gaze, participants in the ASD group were expected to attend less frequently to relevant facial features during the task than the control group.

K. System Design

A VR-based facial emotional expressions presentation system, which incorporated eye tracking and peripheral physiological monitoring, was developed to study the fundamental differences in eye gaze and physiological patterns of adolescents with autism while presented with emotional expression stimuli. The system was composed of three major applications running separately while communicating via a network in a distributed fashion. There were two phases of this study: the online phase represents stimuli presentation, and eye tracking and physiological monitoring, while the offline phase consists of offline data processing and analysis.

Fig. 11 shows the online monitoring components of the overall system. The VR task presentation engine used the popular game engine Unity (www.unity3d.com) by Unity Technologies. The peripheral psychophysiological monitoring used wireless BioNomadix physiological signals acquisition device by Biopac Inc. (www.biopac.com). The eye tracker employed in the study was the Tobii X120 remote desktop eye tracker by Tobii Technologies (www.tobii.com). In the online interaction phase of the system, eye tracking, physiological, performance and regions of interest (ROI) data were logged for offline processing and analysis.

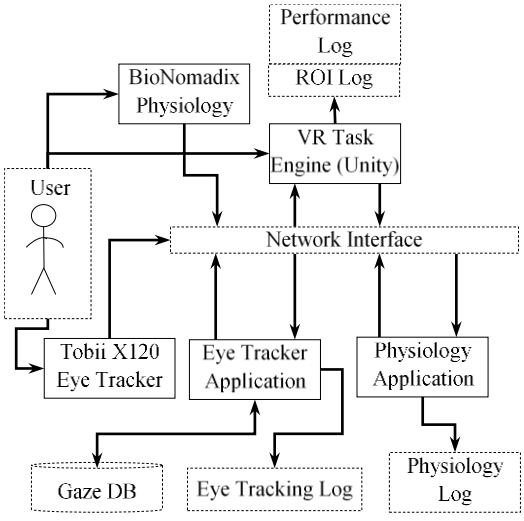


Fig. 11. VR-based facial expressions presentation system.

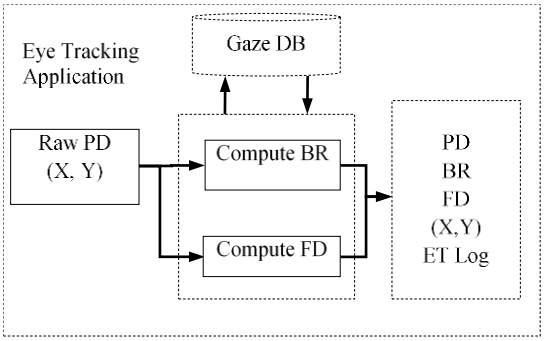


Fig. 12. The Eye tracking application and its components.

The Eye Tracking Application.

The eye tracker application was developed using Tobii software development kit (SDK). The remote desktop eye tracker, Tobii X120, was used with 120 Hz frame rate that allows a free head movement of 30 x 22 x 30 cm (width x height x depth) at 70 cm distance. Its firmware runs on a server that serves eye tracking data via UDP (user datagram protocol) network sockets. This makes it easier to stream the eye tracking data to multiple applications.

The main eye tracker application computed eye physiological indices (PI) such as pupil diameter (PD) and blink rate (BR) and behavioral indices (BI) (U. Lahiri, Warren, Z., Sarkar, N., 2011) such as fixation duration (FD) from raw gaze data. The FD is correlated with attention on a specific region of visual stimuli whereas the eye physiological indices PD and BR are indicative of sensitivity to emotion recognition (C.J. Anderson, Colombo, & Shaddy, 2006; U. Lahiri, Warren, Z., Sarkar, N., 2011).

For each data point, gaze coordinates (X, Y), PD, BR, and FD were computed and logged together with the whole raw data, trial markers and timestamps. The eye tracker application ran two separate network clients: one to monitor the data visually as the experiment progresses and one to record, pre-process and log the eye tracking data.

Fig. 12 shows details of the eye tracking application and its components in the online recording, pre-processing and logging stage. The fixation duration computation was based on the velocity threshold identification (I-VT) algorithm (Salvucci & Goldberg, 2000). We chose the I-VT algorithm for its robustness and simplicity. The algorithm sets a velocity threshold to classify gaze points into saccade and fixation points. Generally, fixation points are characterized by low velocities (e.g.: < 100 deg/sec) (Salvucci & Goldberg, 2000). We used 35 pixels per sample (~60 deg/sec) as our velocity threshold. Spurious fixations processing was not considered in this online interaction phase. Offline post-processing rejected inadmissible fixations. The blink rate was computed using condition code returned from the eye tracker whereas the pupil diameter was obtained by averaging data from both eyes when both eyes' data were available, and from only one eye data when the other eye was not in the tracking range.

The Physiological Monitoring Application.

The physiological monitoring application collected 4 channels of physiological data and was developed using the Biopac SDK and BioNomadix wireless physiological acquisition modules with a sampling rate of 1000 Hz. Like the eye tracker application, this application also received trial and session markers from the task presentation application via its embedded client. The physiological signals monitored were: electrocardiogram (ECG), pulse plethesymograph (PPG), skin temperature (SKT), and galvanic skin response (GSR).

Due to the social communication impairments in adolescents with ASD, there are often inherent challenges in having individuals identify, describe, and often display (e.g., nonverbally communicate) specific internal affective states (C. Liu et al., 2008). Physiological signals are, however, not affected by these impairments and can be useful in understanding the internal psychological states of children with ASD (Groden et al., 2005). Among the signals we monitored, GSR, PPG, and ECG are directly related to the sympathetic response of the autonomic nervous system (ANS) (Cacioppo, Tassinary, & Berntson, 2007). When there is increased sympathetic activity due to external factors such as stress, heart rate, blood pressure, and sweating are all elevated (Cacioppo et al., 2007). We chose these three signals together with skin temperature to analyze physiological patterns during offline analysis to see pattern differences in the presence of stress. We hypothesized that the children would be subject to less stress when they could identify the emotional expressions correctly as compared to when they misidentify them and consequently would have different physiological responses. We used clustering techniques to investigate if the data from these two sets of trials could be clustered as two separate groups in a feature space and whether they could be labelled with good accuracy when compared to the ground truth. For this particular study, we used clustering as there was no training set available for classification, and also as the primary interest in this study was identifying differences as opposed to classifying each group with actual labels.

The VR Task Presentation Engine.

The development of the virtual reality environment involved a pipeline of design and animation software packages. Characters were customized and rigged in online animation and rigging service, Mixamo (www.mixamo.com), and Autodesk Maya. They were animated in Maya and imported into the Unity game engine for final task presentation.

Character Modeling and Rigging.
The characters used in this project were customized in Mixamo to suit the teenage age group targeted for the usability study, i.e., 13-17 years. A total of seven characters including four boys and three girls were selected and customized. Fig. 13 presents three representative characters.

Fig. 13. Representative characters used in the study.

Fig. 14. Anger (top) and surprise (bottom) with two arousal levels.

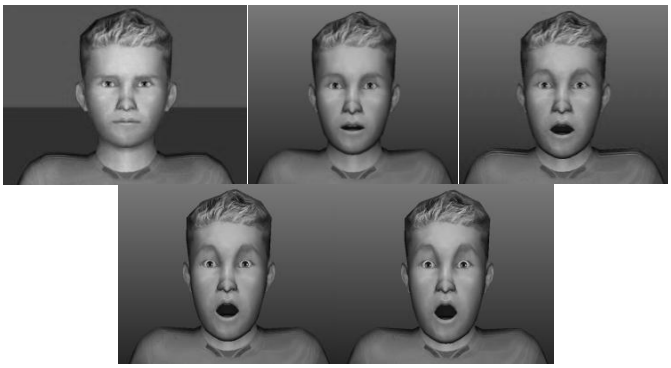


Fig. 15. Neutral (far left) and the four arousal levels for surprise.

Each character was rigged with a skeletal structure consisting of 94 bones. Twenty of these bones were involved with the face structure that was used for facial emotional expressions. Since the main focus of this project was displaying facial emotional expressions, greater emphasis was given to the face structure. The

face rig was attached to a facial emotional controller using set-driven keys. The rest of the body was controlled by inverse kinematics (IK) controllers, except the fingers in which case direct forward kinematics was employed. Besides the facial rig, due attention was given to the quality of the characters.

Facial Expression Animations and Lip-syncing.
Facial emotional expressions and lip-syncing were the major animations of this project. Range of weights were assigned from 0 (no deformation) to 20 (maximum deformation) for each emotional expression. The universally accepted seven emotional expressions proposed by Ekman were used in this project (Ekman, 1993). The expressions are: enjoyment, surprise, contempt, sadness, fear, disgust, and anger. The range of each facial expression had four arousal levels: low, medium, high, and extreme. All the animations were created in Maya using set keys from the set-driven keys because Unity does not currently support any other form of animation import other than set key animations. Each emotional expression was animated from neutral facial expression to the four levels of each emotion. The four levels were chosen by careful evaluation by clinical psychologists involved in this project. Fig. 14 shows two examples of emotional expressions with two arousal levels (medium and high) and Fig. 15 shows the neutral and all the four arousal levels of the surprise emotion. In addition to these facial expressions, seven phonetic viseme poses were created using the same set-driven key controller technique. The phonemes are L, E, M, A, U, O, and I. These phonemes were used to create lip-synced speech animations for storytelling. The story was used to give context to the emotional expressions as described above. A total of 16 stories were lip-synced for each character.
A total of 28 (7 emotional expression x 4 levels of each emotion) animations and 16 story lines lip-synced animations were created. The rigs of each avatar were standardized to make transferring animations easier. All the animations done on one character were then copied to all the remaining characters using attribute copying utility script.

Once all the characters had all the lip-synced animations and all the emotional expression animations resulting in a total of 315 animations, they were batch exported using a utility script into Unity. All the game logic was scripted in Unity as described above.

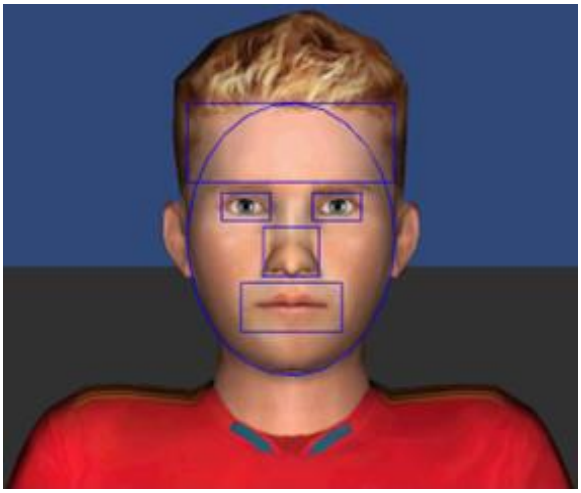


Fig. 16. The Eye tracking application and its components.

Offline Analysis.

The task performance data, the gaze data for each ROI, and the eye tracker and physiology data logs were analyzed offline.

Eye Tracking Data Analysis.

The gaze data analysis was performed to determine behavioral viewing patterns of adolescents with ASD as compared to that of their typically developing (TD) peers. The behavioral indices such as where they were looking in terms of screen coordinates were clustered into ROIs defined around the key facial bones. The clustering results were then averaged over trials for each subject and the aggregate results were used to compare where the adolescents with ASD looked on average compared to the TD adolescents. The defined ROIs represented the following regions: forehead, eyes (left and right), nose, and mouth. The face region was modelled by a combination of an ellipsoid and a rectangular forehead region (Fig. 16). Facial regions outside of the 5 defined regions of interest were categorized as "other face regions" while all the background environment regions outside of the face regions were defined as "non-face regions". This gave a total of seven regions into which all the gaze data points were clustered.

The other behavioral index considered for analysis was the fixation duration. The raw fixation duration was computed for each gaze point during the online interaction. The raw data was first filtered to remove excessively small and large fixation durations. Typical fixation duration and saccades last between 200 and 600 ms and less than 100 ms, respectively (Salvucci & Goldberg, 2000). The filtered fixation duration data was used to compute the average fixation duration (FDave). Another important behavioral eye index associated with fixation duration, called the total sum of fixation counts (SFC), was also computed from the filtered fixation duration data.

The eye physiological indices, i.e., the blink rate and the pupil diameter were also post processed. Missing pupil data due to blinks and presence of noise were filtered from the PD data. The BR data was also filtered based on typical blink ranges. Typical human blinks range between 100 and 200 ms (Shiffman, 2001). The PD data were used to reject incorrectly registered blinks at missing data points.

Physiological Pattern Analysis.

The collected physiological data was analyzed to decipher any pattern differences between two situations, 1) when the subject correctly identified the emotion, and 2) when he/she did not correctly identify the emotion.

First, the signals were filtered to reject outliers and artifacts and were smoothened. Then, the individual baseline mean was subtracted from the data to remove effects of individual variations. The signals were then standardized to be zero mean and unity standard deviation for further feature extraction. For ECG and PPG, the peaks were detected after baseline wanders removal following the artifact removal.

Feature Extraction.

From the four channels of physiological signals collected, i.e., ECG, PPG, SKT and GSR, 16 features were extracted. Table 4 presents all the feature sets used in the physiological analysis. These features were chosen because of their correlation with engagement and emotion recognition process as noted in psychophysiology literature (Cacioppo et al., 2007; C. Liu et al., 2008; K. Welch, Sarkar, M., Sarkar, N., Liu, C., 2010). For example, cardiovascular activities such as inter-bit interval (IBI) represents the rate at which the cardiovascular activity changes and can be used to distinguish arousal levels of an emotion. Electrodermal activity as measured via GSR is indicative of response to external stimuli that might make the subject tense or anxious. The pulse transit time (PTT) is a measure of the time the blood takes to travel from the heart to the finger tips. This specific feature was computed using the peaks of both ECG and PPG signals.

Clustering Analysis.

The extracted features were mapped to a lower dimensional space using principal component analysis (PCA). The PCA analysis revealed that only the first 7 components were sufficient to contain 99.9% of the information contained in the original signal and the first 3 projected components constituted more than 90% of the original information contained in the original feature space.

Table 4. Physiological Feature Sets used in this Study

Channels	Features	Units
ECG	Mean IBI	ms
	SD IBI	N/A
PPG	Mean PTT	ms
	SD PTT	N/A
	Mean IBI	ms
	SD IBI	N/A
	Mean PPG peak	Micro Volts
	Max PPG peak	Micro Volts
GSR	Tonic Mean	Micro Siemens
	Tonic Slope	Micro Siemens/s
	Phasic Response Rate	peaks/s
	Phasic Mean Amplitude	Micro Siemens
	Phasic Max Amplitude	Micro Siemens
SKT	Mean Temp	Degree Fahrenheit
	SD Temp	N/A
	Temp Slope	Degree Fahrenheit/s

IBI: Inter-Beat-Interval, and PTT: Pulse Transit Time

Using the first 2 PCA components, clustering analysis was performed using k-means clustering and Gaussian mixture clustering (GMM). The categories considered were trials when individuals with ASD were correct and when they were incorrect, as well as when TD adolescents were correct and when they were incorrect in identifying the emotions displayed by the avatars. Comparative analysis was performed to see if there was any significant pattern difference between the two sets of trials and two groups, i.e., to see if the physiological pattern of adolescents with ASD was different when they were correctly identifying the emotions compared to when they could not identify the emotions. The same analysis was done for the TD group. The results were compared to the ground truth to determine the accuracy of clustering. Note that these analyses are to show that with the proper choice of learning algorithm, the data set in the feature space could be separable for the two sets of trials within each group as well as across the two groups. This is useful because if the physiological data in the projected space in the presence of external factors (e.g., stress inducing tasks such as the task of identifying emotional faces), they can be used to reduce stress inducing interactions in a future VR task adaptively. For instance, if the subject is feeling stressed interacting with a virtual social peer and the stress

can be automatically detected using appropriate learning algorithms, the virtual interaction can be altered in real-time so as to allow the subject to experience less stress.

Objective Performance Metrics.
In addition to the eye gaze tracking and physiological data, we also measured objective performance metrics to measure the overall effectiveness of the adolescents in the ASD group in identifying the emotional faces when compared to their typical controls. We measured correctness as percentage of total number of trials, asked how confident they were with their choices and their latency to respond. All these measures were averaged for all trials across subjects.

L. Methods and Procedure

A usability study was conducted to validate the system and to study the behavioral and physiological pattern difference of adolescents with ASD and those of typically developing adolescents.

Experimental Setup.
The VR environment ran on Unity. Eye tracking and peripheral physiological monitoring were performed in parallel on separate applications that communicated with the unity-based VR engine via a network interface as described above. The VR task was presented using a 24" flat LCD panel monitor. The experiment was performed in a laboratory with two rooms separated by one-way glass windows for caregiver observation. The caregivers sat in the outside room. In the inner room, the subject sat in front of the task computer. A therapist was present in the inner room to monitor the process. The task computer monitor was also routed to the outer room for caregiver observation. The session was video recorded for the whole duration of participation.

Subjects.
A total of 10 high functioning subjects with ASD with average to above average intelligence (Male: n=8, Female: n=2) of ages 13 – 17 (Mean age (M) =14.7, standard deviation (SD) =1.1) and 10 age matched TD controls (Male: n=8, Female: n=2) of ages 13 – 17 y (M=14.6, SD=1.2) were recruited and participated in the usability study. All ASD subjects were recruited through existing clinical research programs and had established clinical diagnosis of ASD. All subjects in the ASD group fell well above the clinical threshold (Table 5). The gold standard in clinical ASD diagnosis, the Autism Diagnostic Observation Schedule-Generic (ADOS-G) current revised algorithm score (Gotham, Risi, Pickles, & Lord, 2007) and the severity score (ADOS-SS), were used to recruit the ASD subjects. IQ of the ASD subjects was obtained from existing clinical research database.

Table 5. Profile of Individual Subjects in the ASD Group

Subject (Gender)	Age	ADOS-G (cutoff=7)	ADOS-CSS (cutoff=8)	SRS (cutoff=60)	SCQ (cutoff=15)	IQ
ASD1 (f)	17	9	6	88	23	101
ASD2 (m)	15	13	8	80	28	115
ASD3 (m)	14	8	5	80	8	117
ASD4 (m)	16	15	9	69	25	133
ASD5 (m)	14	13	8	74	11	121
ASD6 (m)	13	11	7	82	14	133
ASD7 (m)	15	13	8	81	16	119
ASD8 (f)	14	11	7	87	12	125
ASD9 (m)	14	13	8	77	9	108
ASD10 (m)	15	14	8	81	26	115
Average	14.7	12.0	7.4	79.9	17.78	118.7
(SD)	(1.1)	(2.1)	(1.11)	(5.34)	(7.36)	(9.55)

The control group was recruited from the local community. To ensure control group subjects did not exhibit ASD related symptoms, we asked the parent to complete social responsiveness scale (SRS) (Constantino & Gruber, 2002) and social communication questionnaire (SCQ) (Rutter et al., 2003). Parents of both groups completed these forms. In addition, the Wechsler Abbreviated Scale of Intelligence (WASI) (Wechsler, 2008) was used to measure IQ of the TD subjects. The IQ measures were used to potentially screen for intellectual competency to complete the tasks. The profile of the TD subjects is given on Table 6.

Table 6. Profile of Individual Subjects in the TD Group

Subject (Gender)	Age	SRS (cutoff=60)	SCQ (cutoff=15)	IQ
TD1 (m)	14	41	0	113
TD2 (m)	13	36	0	130
TD3 (f)	17	35	0	110
TD4 (m)	15	39	3	114
TD5 (m)	14	41	6	102
TD6 (m)	14	43	4	90
TD7 (f)	13	36	0	105
TD8 (m)	16	35	3	105
TD9 (m)	15	42	2	115
TD10 (m)	15	45	1	127
Average	14.50	39.30	1.90	111.10
(SD)	(1.38)	(3.44)	(1.97)	(11.14)

All the TD subjects were well below the clinical cut-offs for the SRS and the SCQ.

Tasks.

The VR-based facial emotional understanding system presented a total of 28 trials corresponding to the 7 emotional expressions with each expression having 4 levels. Each trial was 30-45 seconds long. For the first

25-40 seconds, the character narrated a lip-synced context story that was linked to the emotional expression that followed for the next 5 seconds. The avatar exhibited a neutral emotional face during storytelling. Subjects were instructed to rate the emotions based on the last 5 seconds of interaction. The story was used to give context to the displayed emotions. The context of the stories ranged from incidents at school to interactions with families and friends that were suitable for the targeted age group.

A typical laboratory visit was approximately one hour long. During the first 15 minutes, a trained therapist read approved assent and consent documents to the subject and the parent, and explained the procedures. Once the subject finished signing the assent document, he/she began the task. While the parent completed the SRS and SCQ forms, the subject wore the wearable physiological sensors with the help of a researcher. Before the task began the eye tracker was calibrated. The calibration was a fast 9 point calibration that took about 10-15 s.

At the start of the task, a welcome screen greeted the subject and described what was about to happen and how the subject was to interact with the system. Immediately after the welcome screen, the trials started. At the end of each trial, questionnaires appeared on screen prompting the subject to label the emotion he/she thought the avatar displayed and how confident he/she was in his/her choice. The total participation time was about 20-25 minutes. The emotional expression presentations were randomized for each subject across trials to avoid ordering effects. To avoid other compounding factors arising from the context of the story, the story was recorded with monotonous tone and there was no facial expression displayed by the avatar during that story telling period.

M. Results

Gaze Pattern Comparisons.

We quantified gaze pattern in percentage as number of gaze points to a specific ROI over total number of gaze points. Comparative analysis was performed to distinguish pattern differences within the ASD and TD groups and across the groups. The pattern difference analysis also compared behavioral gaze pattern differences during the context story telling part of the trial and the last 5 seconds of the trial in which the avatar displayed the facial emotional expressions. We have used statistical significance paired t-test to compare intra group variations and independent two sample unequal variance t-test to compare inter-group variations.

Gaze Comparisons During Emotion Recognition.

These results compared the gaze patterns of the ASD group to that of the control group for the last 5 seconds of the trials when the avatar displayed facial emotional expression. Data were averaged across trials for each subject. Statistically significant gaze difference between the two groups was found for the mouth and the forehead ROIs.

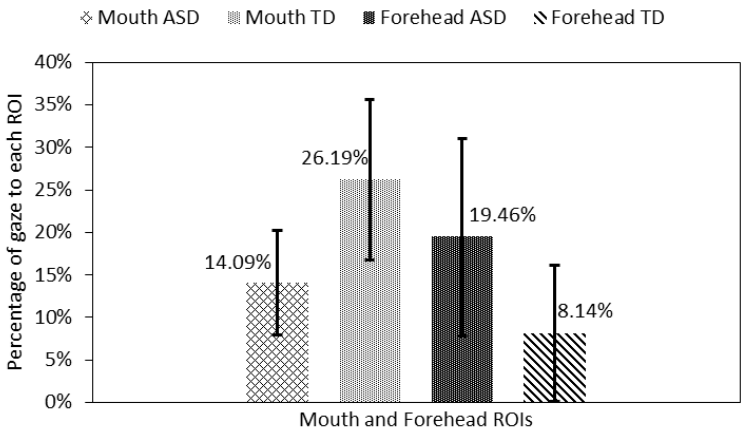


Fig. 17. Gaze towards mouth and forehead regions.

The adolescents with ASD looked 11.32% ($p<0.05$) more towards the forehead area and 12.1% ($p<0.05$) less to the mouth area than the TD subjects (Fig. 17). Note that both the forehead and the mouth areas were the primarily morphed ROIs in most of the emotional expressions. Adolescents with ASD also looked 5.58% less to the eye area and 1.89% less to the nose area. But these differences were not statistically significant.

The total face area ROIs (these do not include the face regions that are not marked) were combined and compared with non-face ROIs. Adolescents in the ASD group looked 6.41% ($p<0.05$) more to the non-face region and they looked 8.25% ($p<0.05$) less to the combined facial ROIs as compared to the TD control group (Fig. 18).

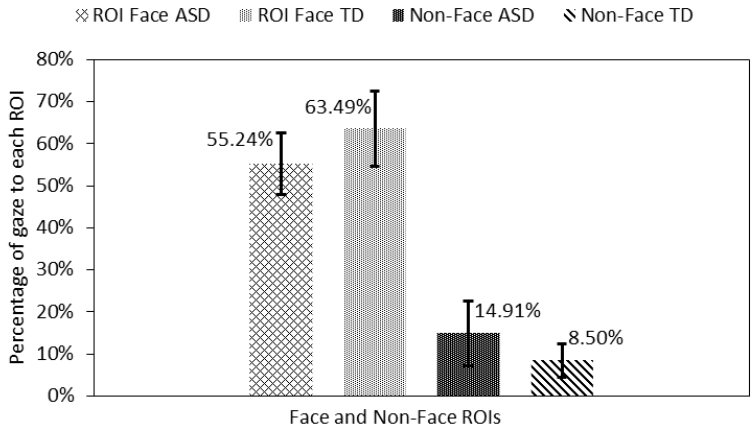


Fig. 18. Gaze towards face and non-face regions.

Gaze Comparisons During Story-telling.

We also compared the gaze patterns of the ASD group to that of the TD group for the context story telling portion of the trials (Fig. 19). Note that the facial expression of the avatar was neutral during this time. The

statistically significant gaze differences between the two groups were found for the mouth and the forehead ROIs.

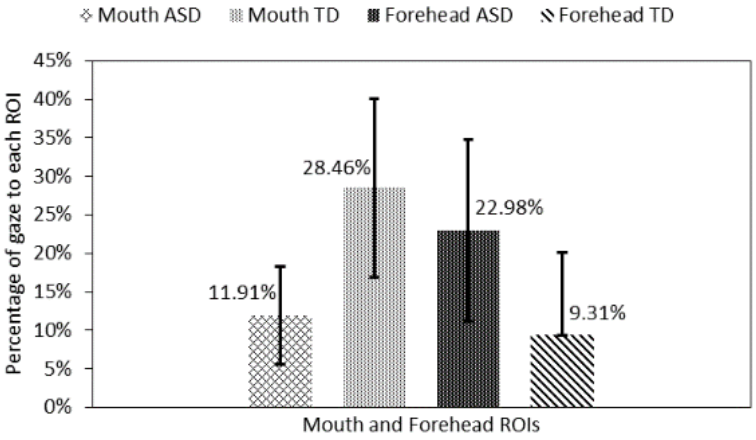


Fig. 19. Gaze towards mouth and forehead regions.

The adolescents with ASD looked 13.67% ($p<0.05$) more towards the forehead area and 16.55% ($p<0.05$) less to the mouth area than the adolescents in the TD group (Fig. 19). It is interesting to note that similar looking patterns towards the mouth and the forehead area were observed in both emotion recognition and storytelling cases. Adolescents with ASD also looked 1.59% less towards the eyes ROI and 2.39% less to the nose ROI. However, these differences were not statistically significant.

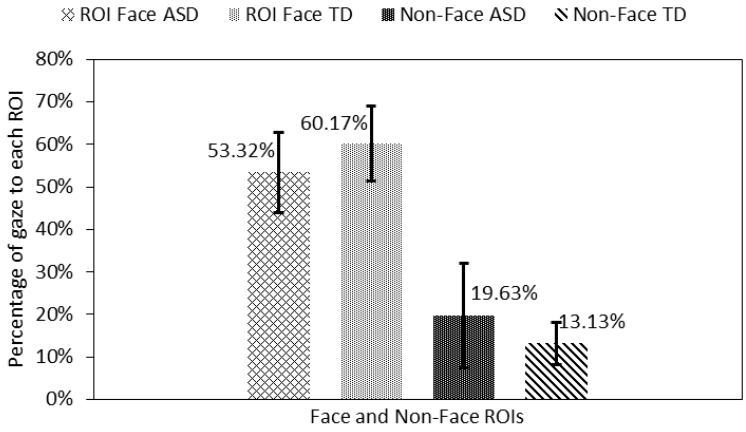


Fig. 20. Gaze towards face and non-face regions.

Adolescents in the ASD group looked 6.5% more towards the non-face region and 6.85% less towards the combined facial ROIs as compared to the TD control group (Fig. 20). But, these differences were also not statistically significant.

Gaze Comparisons When Subjects Correctly Identified the Emotion.
Here we compared the gaze patterns of the ASD group to that of the TD group when both groups correctly identified the emotions displayed by the avatars. Statistically significant gaze difference between the two groups was found only for the mouth ROI. This analysis is performed for the last 5 seconds of the trials when the avatar changed its emotional expression from neutral to a target state.

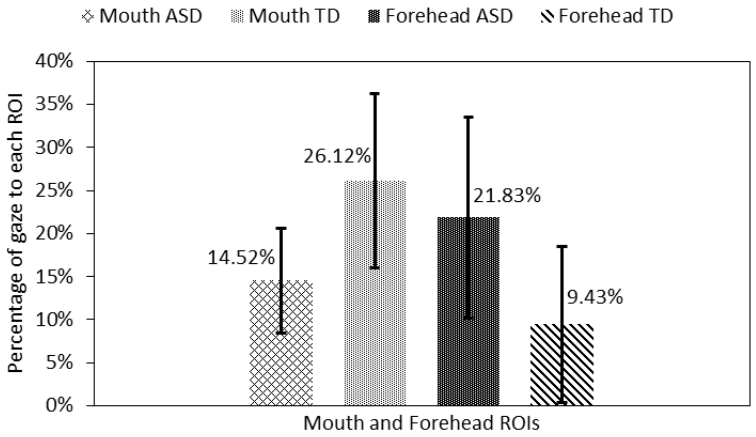


Fig. 21. Gaze towards mouth and forehead regions.

The adolescents with ASD looked 12.4% more towards the forehead area and 11.6% ($p<0.05$) less to the mouth area than the adolescents in the TD group (Fig. 21). Further, adolescents with ASD also looked 7.86% less towards the eyes area and 2.51% less to the nose area. But these differences were not statistically significant.

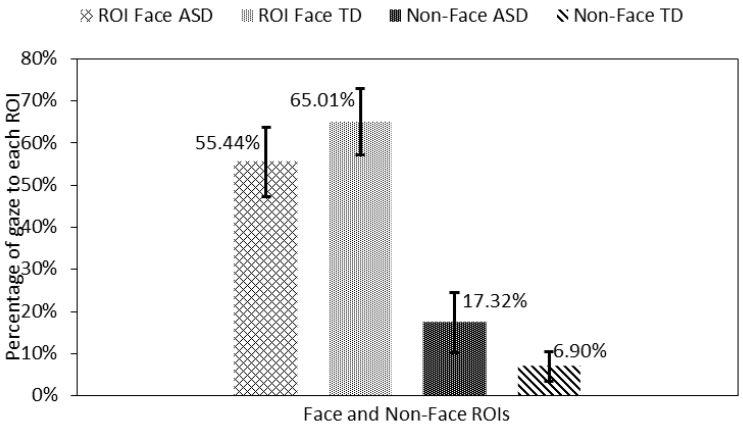


Fig. 22. Gaze towards face and non-face regions.

Finally, adolescents in the ASD group looked 10.42% ($p<0.05$) more towards the non-face region and 9.57% ($p<0.05$) less to the combined facial ROIs as compared to the TD control group (Fig. 22).

Gaze Comparisons When Subjects Incorrectly Identified the Emotion.
Here we compared the gaze patterns of the ASD to that of the TD group when both groups were incorrect in identifying the emotions displayed by the avatars. Statistically significant gaze differences between the two groups were found only for the mouth and forehead ROIs. This analysis is performed for the last 5 seconds of the trials when the avatar changed its emotional expression from neutral to a target state.

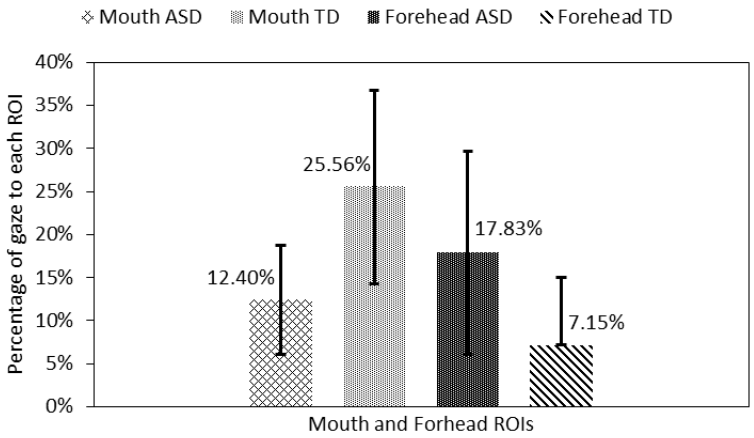


Fig. 23. Gaze towards mouth and forehead regions.

The adolescents with ASD looked 10.68% ($p<0.05$) more towards the forehead area and 13.16% ($p<0.05$) less towards the mouth area than the adolescents in the TD group (Fig. 23). Further adolescents with ASD also looked 3.2% less towards the eyes area and 1.09% less towards the nose area.

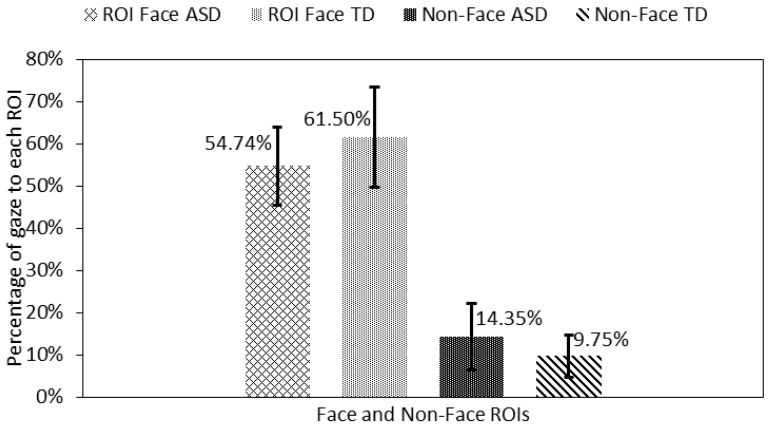


Fig. 24. Gaze towards face and non-face regions.

Adolescents in the ASD group looked 4.6% more towards the non-face region and 6.76% less towards the combined facial ROIs as compared to the TD control group (Fig. 24). But, these differences were not statistically significant.

Table 7. Amount of time spent looking at Mouth and Forehead ROIs across groups.

	Neutral			Emotion Expression		
	Mouth*	**Forehead***	**Face**	**Mouth***	**Forehead***	**Face***
ASD	11.91%	22.98%	86.87%	14.09%	19.46%	85.09%
TD	28.46%	9.31%	80.37%	26.19%	8.14%	91.50%
t-value (df=18)	-3.76	2.56	-1.47	-3.24	2.41	-2.23
p-value	0.0014	0.0196	0.1582	0.0045	0.0269	0.0388

We then further examined the percentage of time participants spent looking at different parts of the avatar's face during the neutral and emotional expression conditions across the two groups (see Table 7). Participants in both groups spent similar amounts of time looking at the avatar's face as well as eyes, nose, and "other face areas." Significant differences emerged in the amount of time spent looking at the avatar's mouth and forehead, however, with ASD participants looking more at the forehead (neutral: M = 22.98%, $t = 2.56$, $p < .05$; emotion: M = 19.46%, $t = 2.41$, $p < .05$) and TD participants looking more at the mouth (neutral: M = 28.46%, $t = -3.76$, $p < .01$; emotion: M = 26.19%, $t = -3.24$, $p < .01$) across both conditions. A visual map of these different looking times is presented in Fig. 16.

Intra Group Gaze Comparisons.

We also performed within group gaze pattern difference analysis for both the ASD and the TD groups between those trials when the adolescents identified the displayed emotion correctly versus those trials when they did not.

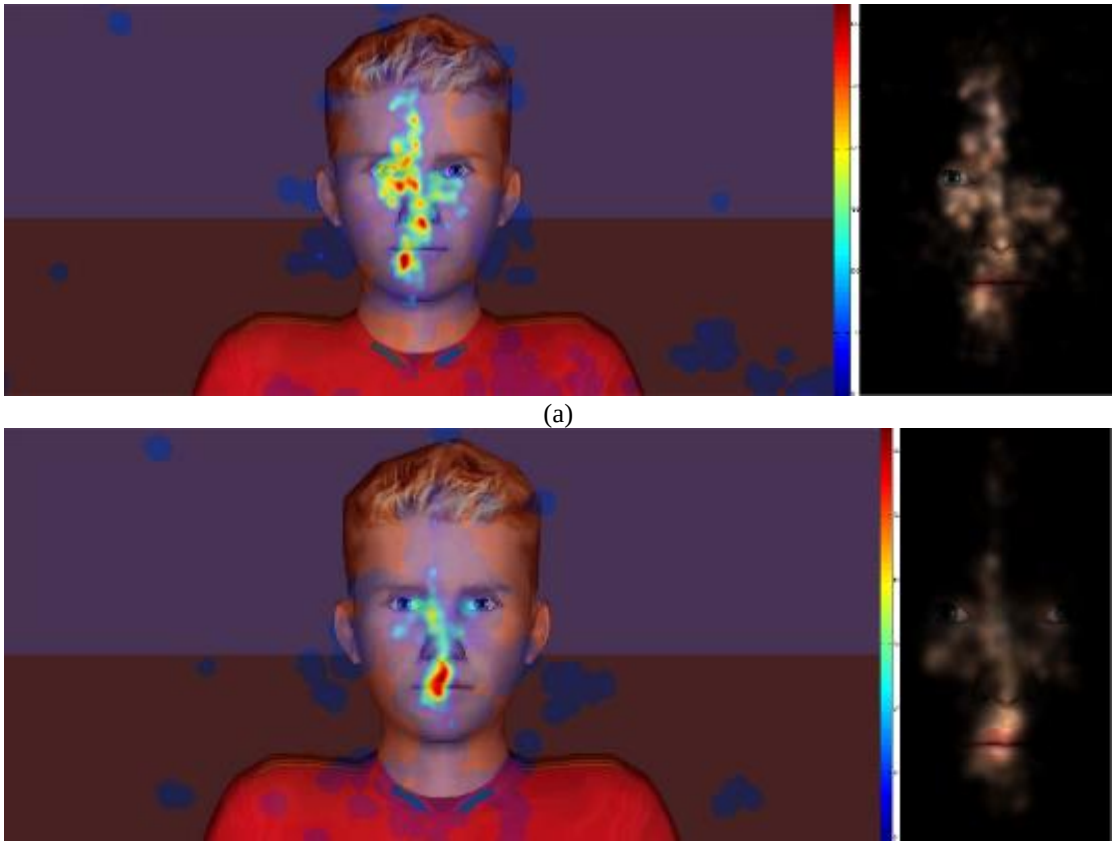
The adolescents in the ASD group looked 2.12% ($p<0.05$) more towards the mouth ROI when they were correct than when they were incorrect while their TD counterparts looked 0.56% more towards the mouth area for same situations (Fig. 25). The result of the TD group was not statistically significant. There were variations in other ROIs as well, but none of them were statistically significant.

Eye Behavioral Indices (BI).

Average fixation duration (FDave), the average saccade path length (SPLave) and sum of fixation counts (SFC) were used as behavioral viewing pattern measures in this study. These indices were computed as discussed above. These behavioral indices are indicative of engagement to particular stimuli and are correlated with social functioning for individuals with autism (A. Klin et al., 2002). Generally, adolescents in the ASD group had lower FD and SFC than the control group in all the four comparison sections described above (Fig. 26 and Table 8). However, none of them were statistically significant.

Table 8. Measures of behavioral viewing pattern

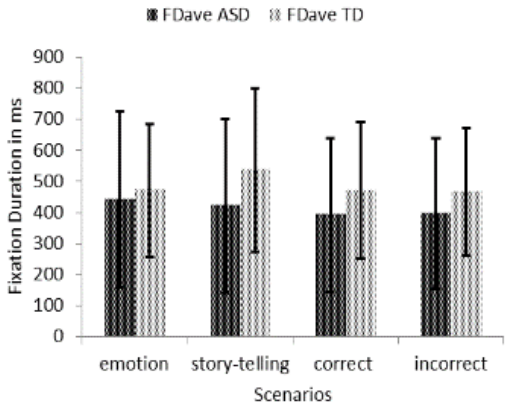
	PDave (mm)	BRave (bpm)	FDave (ms)	SPLave (pix)	SFC (no unit)
ASD	3.21	5.34	414.84	116.08	26.90
TD	3.61	12.26	471.59	128.27	32.26

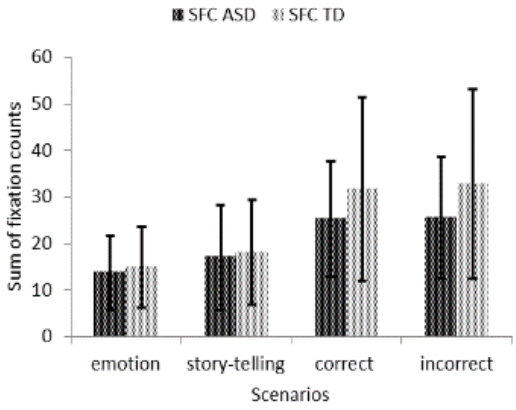


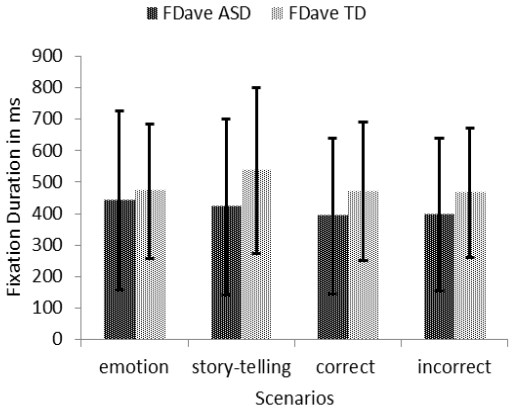
(a)

(b)

Fig. 25. Intergroup comparison gaze visualizations (heat maps and masked scene maps) (a) combined gaze in the ASD group for all the trials and all participants (b) combined gaze in the TD group for all the trials and all participants.


FDave ASD
FDave TD
Fixation Duration in ms
900
800
700
600
500
400
300
200
100
0
emotion
story-telling
correct
incorrect
Scenarios


SFC ASD
SFC TD
Sum of fixation counts
60
50
40
30
20
10
0
emotion
story-telling
correct
incorrect
Scenarios


FDave ASD
FDave TD
Fixation Duration in ms
900
800
700
600
500
400
300
200
100
0
emotion
story-telling
correct
incorrect
Scenarios

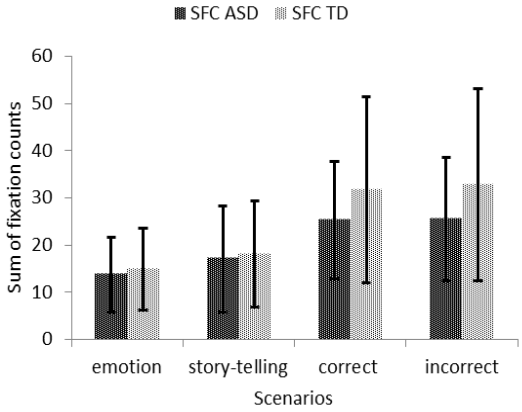

SFC ASD
SFC TD
Sum of fixation counts
60
50
40
30
20
10
0
emotion
story-telling
correct
incorrect
Scenarios

Fig. 26. Comparisons of behavioral eye indices.

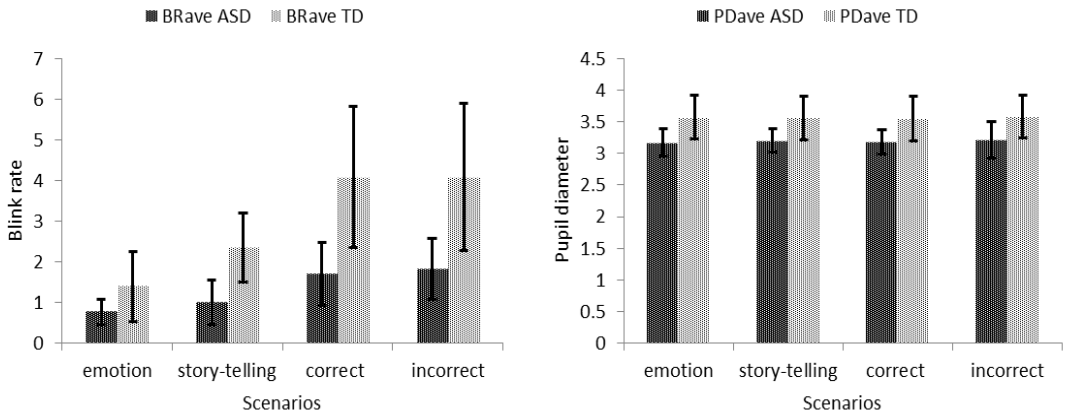


Fig. 27. Comparisons of physiological eye indices.

Eye Physiological Indices (PI).

The physiological patterns of the eyes of the subjects were represented by the average pupil diameter (PDave) and the average blink rates (BRave). PD is indicative of how engaged a subject is and literature suggests that there are variations of these indices between individuals with ASD and TD individuals (U. Lahiri, Warren, Z., Sarkar, N., 2011) given the same stimuli. Individuals with autism were shown to have abnormal eye blink conditioning compared to TD subjects (Sears, Finn, & Steinmetz, 1994). Generally, adolescents with autism exhibited lower pupil diameter and blink rates in all the four comparison scenarios. Specifically, they had 11.2% ($p<0.05$) less PD on average than adolescents in the TD group during the emotion display cases. They also showed 57.44% ($p<0.05$) fewer blinks on average during the story-telling cases and 55.39% ($p<0.05$) fewer blinks on average when they were incorrectly identifying the emotions compared to their typical counterparts (Fig. 27). Unlike reported high blink conditioning for individual with ASD in general (Sears et al., 1994), these low blinks could be attributed to 'sticky attention' that this population sometimes exhibits (Landry & Bryson, 2004).

Physiological Pattern Analysis.

As described above, identifying physiological pattern differences when there were external factors such as stress within a social task in which emotion identification is a part, is of particular importance for the development of an affective state detection system for a future adaptive VR social interactive task. To investigate if the differences in physiological patterns are strong enough to be classified by supervised training methods for online classification, we used unsupervised clustering. If a feature space is separable using unsupervised methods, this may imply that there is sufficient pattern to be learned by supervised algorithms to classify the data accordingly. For this purpose, the physiological data of the adolescents in both groups was separated into data from trials when the adolescents were correct and trials when they were incorrect in identifying the displayed emotion. The combinations of these four dataset were clustered using k-means and Gaussian mixture (GM) clustering methods using two datasets at a time. This resulted in four comparisons. The ground truths were two pair of classes of ASD group correct, TD group correct, ASD group incorrect and TD group incorrect. In both algorithms the data was first clustered into four clusters and the four clusters were re-clustered back to two clusters. Then the accuracy was computed between the clustering result and the ground

truth to measure how well a machine learning algorithm could differentiate between these physiological data of ASD and TD groups. Accuracy here represents cluster quality.

The k-means and the GM achieved accuracies of 55.36% and 57.5%, respectively, separating the data of adolescents with ASD when they were correct from when they were incorrect. On the other hand, the two algorithms were able to separate those of the TD group with accuracies of 55% and 71.43%, respectively. For the across group comparison, when both groups were incorrect the k-means and the GM clustered the data, with accuracies 53.17% and 73.81%, respectively, while they separated the data when both groups were correct, with accuracies of 61.36% and 73.38%, respectively. Fig. 28 shows an example for the case when both groups were correctly identifying the displayed emotions. It shows the original ground truth clusters (a), the Gaussian mixtures overlaid on top of the original ground truth clusters (b), the result of k-means clustering (c), and the result of GM clustering (d), when both groups were correct. Note that the four Gaussian mixtures (b) were used to first cluster the data into four preliminary clusters and finally they were re-clustered to give just two clusters (d) as the original ground truth clusters. Although, the within pattern differences of the ASD group was less than the TD group, the results indicated that there were distinguishable physiological pattern differences between the physiological responses in these two task situations.

Based on the analysis, it is observed that k-means clustering qualities were not as good as those of the Gaussian mixture clustering (Fig. 29). In Fig. 28, the GM clusters (d) more closely resemble the original clusters (a) as compared to the k-means clusters (c). For this analysis, the first two PCA components, which represent more than 80% of the original 16 features data set, were used. Results from ten iterations were averaged.

A more robust clustering algorithm such as hierarchical clustering might result in better clusters than even the GM clusters. Projecting the features using kernels could also be another alternative to increase cluster quality.

In general, these results indicate that there are clear pattern differences in physiological responses of adolescents with ASD while performing social tasks such as identifying emotional faces. Given enough training data, these pattern differences can be learned using supervised non-linear classifiers to enable adaptive VR-based social interaction in the future. Liu et al. (C. Liu, Conn, K., Sarkar, N., Stone, W., 2008b), for instance, showed that it is possible to use such physiological measures to create an adaptive closed-loop robotic interaction in real-time using support vector machines (SVM) with Gaussian kernels.

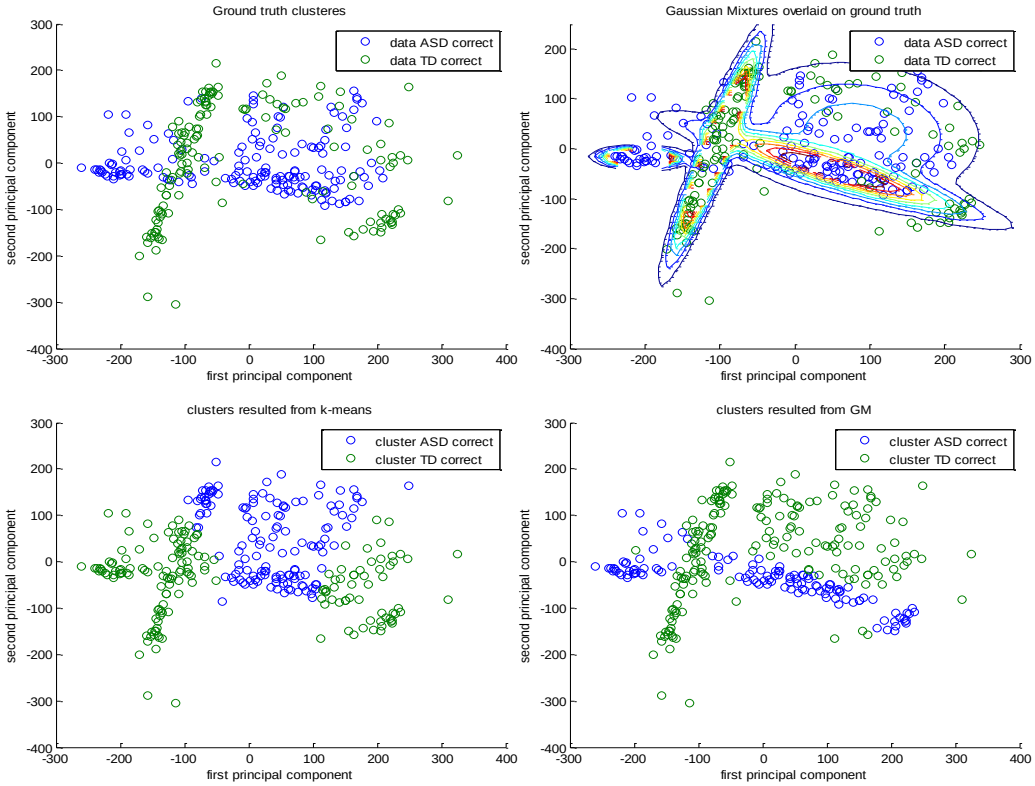


Fig. 28. (a) Top left: original ground truth clusters. (b) Top right: the Gaussian mixtures used in the GM clustering overlaid on the ground truth clusters. (c) Bottom left: the result of the k-mean clustering. (d) Bottom right: the result of GM clustering.

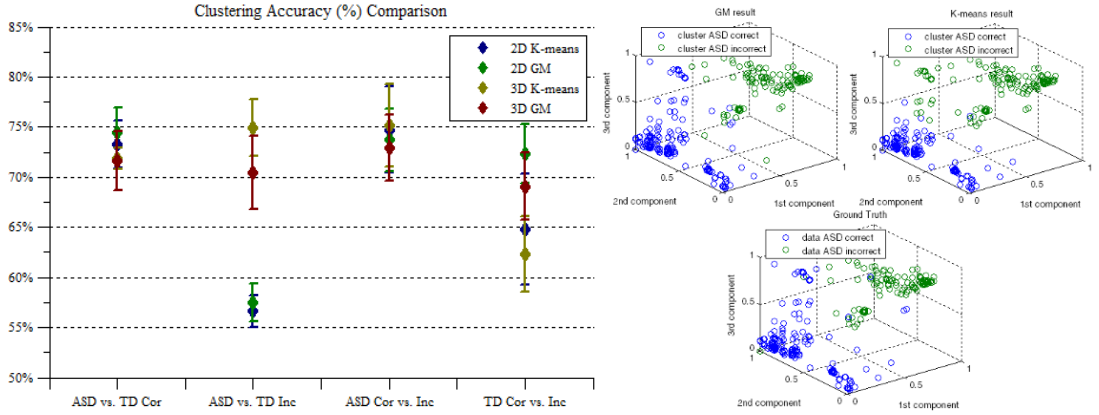


Fig. 29. Left: Results showing clustering quality of the two clustering methods using two different sets of PCA components for all the four sets of comparisons. Cor: correct and Inc: incorrect. Right: (a) Top left: result of the GM clustering, (b) Top right: result of the k-means clustering, and (c) Bot-tom: ground truth of clusters of data of from trials of ASD subjects when they were correctly identifying the emotions vs. when they were incorrect in identifying the emotions.

Performance.

We examined participant performance using three metrics: accuracy, response latency, and ratings of response confidence. We conducted independent samples t-tests to analyze performance in these categories. To test our hypothesis that participants with ASD would be less accurate at identifying emotions, we first compared performance accuracy between ASD and TD groups using number correct (out of 28 total presentations) as the dependent variable (see Table 9). There was no significant difference in overall performance accuracy between groups ($t = 1.23$, $p = 0.23$; ASD: $M = 16.30$, $SD = 2.67$; TD: $M = 14.50$, $SD = 3.78$). Groups also performed similarly accurately when identifying types of emotion displayed.

Table 9. Group Differences on Performance Metrics (out of 28 presentations)

Metric	ASD		TD		Statistics	
	Mean	SD	Mean	SD	t- value (df=18)	p-value
Overall Accuracy (%)	16.30 (58.22%)	2.67 (9.53%)	14.50 (51.79%)	3.78 (13.49%)	1.23	0.23
Response Time* (seconds)	11.30	4.47	7.50	0.97	2.63	0.02
Confidence Ratings*	75.52%	10.69%	90.09%	5.75%	-3.80	0.001

The average score of the adolescents with ASD was 6.42% higher than the adolescents in the TD group while they were 14.57% ($p<0.05$) less confident in their choices and took 4 seconds ($p<0.05$) more than the control group (Table 9). However, the performance score differences were not statistically significant. The low score in both groups was due to the apparent high misclassification of the emotion contempt as disgust and surprise as fear and vice versa. These patterns were observed for both groups. We also performed a rating by a group of typical college students and they too highly misclassified the above-mentioned emotions. Therefore, a part of these low scores can be attributed to the design limitations of these emotion expressions as opposed to the inability of the adolescents to identify the emotions.

Table 10. Average recognition accuracy by emotion of the college group testing.

Emotion	ASD		TD		Statistics	
	Mean	SD	Mean	SD	t- value (df=18)	p-value
Anger	75.00%	20.41	70.00%	28.38	0.45	0.66
Contempt	17.50%	20.58	15.00%	17.48	0.29	0.77
Disgust	70.00%	30.73	62.50%	21.25	0.64	0.53
Fear	42.50%	20.58	22.50%	27.51	1.84	0.08
Joy	57.50%	23.72	55.00%	19.72	0.26	0.80
Sadness	90.00%	17.48	90.00%	17.48	0.00	1.00
Surprise	52.50%	21.89	47.50%	27.51	0.45	0.66

Although our facial expression animations were developed from images widely used in research, we ran follow-up analyses to ensure responses were not related to factors unique to the newly developed stimuli. We examined the data for any consistent response patterns among answers that were incorrect (i.e., misclassified facial expression). Specifically, we analyzed response patterns for all incorrect answers across all participants and for all trials (Table 10). Results revealed that across both groups, participants' most common patterns of error related to misclassification of contempt as disgust (52.5% of the time), disgust as anger (25%), and joy as surprise (21.3%). When examining overall correct identification, participants accurately identified joy, anger, fear, and sadness more than 50% of the time across all conditions, showing less accuracy with surprise (50%), contempt (16.25%), and fear (32.5%).

N. Discussion and Conclusion

We have developed a VR-based controllable facial expression presentation system that was able to collect eye tracking and peripheral psychophysiological data while the subjects were involved in emotion recognition tasks. Specifically, we developed controllable levels of facial expressions of emotion based on long-standing research documenting certain universal expression patterns (Ekman, 1993) as well as a desktop virtual reality presentation to avoid issues and sensitivities individuals with ASD exhibit with immersive virtual reality displays such as head mounted displays (HMD) (Harris & Reid, 2005). Subsequently, a usability study involving 10 adolescents with ASD and 10 typical controls was performed to evaluate the efficacy of the system as well as to study behavioral and physiological pattern differences in how individuals processed different valences of these expressions. In this pilot study, we used a novel VR system and a range of newly developed avatar facial expressions to examine performance and process differences between adolescents with and without ASD on tasks of facial emotion recognition. Although, this particular study did not employ direct interaction between the user and the system, the study is a precursor to a more interactive adaptive multimodal VR social platform that is under development. Such capabilities are expected to be useful in understanding the underlying heterogeneous deficits individuals with ASD often display in processing and responding to nonverbal communication of others. In turn, such a system will hopefully contribute to the development of novel intervention paradigms capable of harnessing these technological advancements to improve such impairments in a powerful, individually specific manner. The system successfully presented the facial emotional expressions and collected the synchronized eye gaze and physiological data.

Our hypotheses were partially supported, such that group differences emerged when examining response latency, eye gaze, and confidence in responses, but not participants' ability to correctly identify emotions. Results did not suggest powerful differences between our ASD and TD groups in terms of emotion recognition. Contrary to our expectations, no significant difference was found between groups in terms of overall performance accuracy. In fact, in some instances individuals with ASD recognized expressions with greater accuracy than the TD group. While this was somewhat surprising given that individuals with ASD often have potent impairments regarding nonverbal communication, it is not entirely inconsistent with existing literature. Other researchers have found that individuals with ASD, particularly those with average to above average intelligence as in our sample, can often be taught to accurately identify basic static emotions. Previous work has noted that children, particularly children with very high cognitive abilities as was the case in the current study, are capable of identifying basic static emotions at a high level of accuracy. However, we had originally hypothesized that our dynamic displays of emotion within the VR environment might prove challenging for children with ASD. Ultimately, this result supports the possibility that if children with ASDs have difficulty with affect recognition, the problem may not be at the level of basic naming or recognition which mirrors decades worth of extant literature on affect recognition using static images (Adolphs et al., 2001; Castelli, 2005; G. Dawson et al., 2005) rather in complex recognition tasks and in the presence of social context stimuli. However, often individuals with ASD still struggle with integration of this recognition skill in fluid interaction across environments.

Hypothesized significant group differences did emerge when additional performance metrics and gaze patterns were examined. We did in fact find interesting differences in how facial expressions were processed and decoded across between these groups. Participants in the TD group showed generally symmetrical eye gaze to relevant areas on the face during the animations and demonstrated a significant bias toward relevant components of the facial expression (i.e., mouth ROI). However, those in the ASD group showed a pattern of more distributed attention with less focus to relevant stimulus features. Adolescents in the ASD group also paid less attention to the eye area than the TD group on average, although these differences were not statistically significant. Importantly, in this particular study, all the animations did not manipulate the eye areas and hence the eyes were static in many instances and thus attracted less attention which may explain why a bias may not have been evident in this region. Another interesting finding regarding face processing was that adolescents with ASD spent much more time examining faces prior to response and they were often less confident in their ratings. Finally, although our system was not designed to map specific physiological responses, our offline analysis demonstrated that meaningful physiological pattern differences could be detected during system performance. Such differences may be potentially modelled onto constructs of stress and engagement over time in order to further enhance and endow our system with the ability to understand processing and performance.

Results indicated interesting differences in performance regarding identification of certain emotions as well as differences in how individuals with ASD often processed emotions. With regard to performance two pairs of emotional expressions, i.e., contempt and disgust as well as fear and surprise were often confusing for subjects to successfully discriminate. In preliminary ratings of these emotions during our development phase, an independent group of college students found similar problems regarding these pairs of emotion. One potential explanation of this confusion and challenge in discriminating these universal faces within system construction may be related to flexible limits of expression utilized in the current study. All the emotional expressions were developed using taped actor performances. The key facial regions such as eye brows, the mouth, and the cheek area were manipulated based on the benchmark variations in those regions of the taped performances. A total of 20 facial bones were employed to make the emotional expressions realistic. However, there were some limitations on the extent to which the facial bone rigs could manipulate the key skin areas and in some cases extreme levels of certain emotions it resulted in unnatural and ambiguous mesh deformations. Subsequent system refinement may enhance such discrimination. Results also demonstrated that extreme and low levels of emotions were hard to be detected by both groups, whereas the medium and high arousal levels were reliably identified in most cases. This also suggests more flexible methods for driving key naturalistic variation in expressions may be necessary for creating a system that presses appropriately for a fuller range of expression recognition in this population. Notably the riggings utilized to display animations for the involved avatars had a majority of connection features (i.e., relevant features that dynamically changes to display the emotion) within the very region of interest where TD participants focused a majority of their gaze (i.e., mouth ROI). Each facial rig for each character contains 6 out of the 17 facial bones (35%) compared to just one bone on the forehead (5.88%) and 4 bones around each eye (23.5%). This result mirrors previous findings regarding facial processing and eye gaze for people with ASD (C.J. Anderson et al., 2006; Castelli, 2005; Celani, Battacchi, & Arcidiacono, 1999; G. Dawson, Webb, Carver, Panagiotides, & McPartland, 2004; Hobson, 1986; Hobson, Ouston, & Lee, 1988), suggesting atypical attention to irrelevant features. Such a difference in gaze attention may help explain why participants in the ASD group took longer to achieve similar results as the TD group regarding what facial expressions were being displayed. The faster response times of the TD group may be due to focusing quickly on relevant features when making judgments about social stimuli. Indeed, previous research has shown that when children with ASD are primed to look at salient facial features during a facial emotion recognition task, they focused on non-core facial feature areas than the control group (A. Klin et al., 2002; Pelphrey et al., 2002). As such, developing tools such as VR displays of emotion that can dynamically display emotive expressions and potentially guide and alter gaze processing and attention to enhance facial recognition, may prove a valuable intervention approach over time. Specifically, such a system might be capable of enacting changes not simply recognizing emotions,

but changes related to how such emotions are recognized. Further, addressing social vulnerabilities on a processing level may result in changes that more powerfully generalize than current approaches for enhancing skills in social interaction, as real-world social interactions often require fast and accurate interpretation of, and response to, others' verbal and nonverbal communication.

In combination, these findings suggest our system was able to elicit and drive meaningful differences related to how individuals processed information from faces. Such capabilities are expected to be useful in understanding the underlying heterogeneous deficits individuals with ASD often display in processing and responding to nonverbal communication of others. In turn, such a system will hopefully contribute to the development of novel intervention paradigms capable of harnessing these technological advancements to improve such impairments in a powerful and individually specific manner.

Although these findings are promising there are several important limitations to note. First, this was a static performance driven system, and physiological indices were not incorporated into online performance or modification. Further, our design of the emotional expressions, while based on decades of research and theory, was not adequately able to push for accurate identification of certain emotions across groups. Finally, while the system created some sense of a social scenario, such interactions were very limited in the application. The present study is preliminary in nature. Our goal was to test participant response to our novel animated facial expression stimuli. Cognitive scores in the current sample are higher than general population norms, which may limit generalization of findings. This sample characteristic may be irrelevant, however, as all of our teenaged participants (both with and without ASD) were less accurate in identifying more subtle expressions than the college students included in a previous verification sample. Observed response variability, then, may be better explained by maturation of social cognition rather than general cognitive ability, per se. Our study's small sample size, although characteristic of exploratory studies in general, weakens the statistical power of the results.

Despite limitations this initial study demonstrates the value of future work subtly adjusting emotional expressions, integrating this platform into more relevant social paradigms, and embedding online physiological and gaze data to guide interactions with potential relevance toward fundamentally altering and improving how individuals with ASD process nonverbal communication within and hopefully far beyond VR environments. We anticipate that providing a safe environment in which to practice such social skills, with ongoing monitoring of performance, engagement, and processing will lead to individuals with ASD having more confidence and successful navigation of parallel tasks in the real world.

CHAPTER V

DESIGN OF A VIRTUAL REALITY SYSTEM FOR AFFECT ANALYSIS IN FACIAL EXPRESSIONS (VR-SAAFE) AND IAPS PICTURES PRESENTATION; APPLICATION TO SCHIZOPHRENIA

There exist a large number of studies in physiology and eye gaze based affect analysis of SZ patients. Most of these results, however, are based on static stimuli or video clips. In recent years, VR has been used for SZ intervention but mostly as a performance-based system without having any built-in internal measures. The scope of this present work is to bridge this gap and develop a VR-based SZ intervention system that is integrated with both peripheral physiology and eye gaze monitoring systems with the aim of precisely understanding the implicit response to VR-based dynamic representation of facial expressions. *We believe that by precisely controlling emotional expressions in VR and gathering objective individualized eye gaze and physiological responses related to emotion recognition as well as performance data, new efficient intervention paradigms can be developed in the future. The novel VR-based system and the findings in this study may inform future development of affect-sensitive virtual social interaction tasks.*

In this chapter, we present a novel VR-based system that incorporates implicit cues from peripheral physiological signals (C. Liu, Conn, K., Sarkar, N., Stone, W., 2008a, 2008b) and eye tracking (Andreasen, 1984) for the understanding of facial emotional expression. We compare how a SZ group and a matched group of healthy non-psychiatric adults performed emotion recognition tasks when presented in the form of static IAPS slides and when presented in a VR environment with the avatars expressing emotions dynamically. Both the IAPS presentation system and the VR system were composed of three major components: the presentation environment, the eye tracking component, and the peripheral physiology monitoring component. The presentation environments were based on Unity3D game engine by Unity Technologies (http://unity3d.com). A remote desktop eye tracker by Tobii Technologies (www.tobii.com) called Tobii X120 was employed for gaze tracking. A wireless physiological signals acquisition device called BioNomadix by Biopac Inc. (www.biopac.com) with 8 channels of peripheral physiology electrodes was used to record the physiological signals. Each component ran separately while communicating via a network interface. The system development is completed and a pilot study with 12 patients with schizophrenia and 12 health controls was conducted. Preliminary results from the pilot study involving brief comparison of the VR system and the IAPS presentation system in terms of physiological responses, performance metrics and eye gaze parameters with the 12 patients with SZ and matched controls is presented.

O. The Static IAPS presentation

We developed a picture presentation system using Unity3D that displayed the full screen images on a 24" flat screen monitor at 1024x768 resolution (the original resolution of the IAPS images). The pictures were preselected from the pool of about 600 IAPS pictures (Lang et al., 1999). They were categorized into 6 major groups, namely, social positive (pictures of erotica), social negative (violence pictures), social neutral (people in normal scenery), non-social positive (pictures of food), non-social negative (pictures of dirty and unpleasant scenery), and non-social neutral (normal pictures of objects). The emotional pictures were broadly divided into social and non-social and within each broad category, they were further categorized into positive, nega-

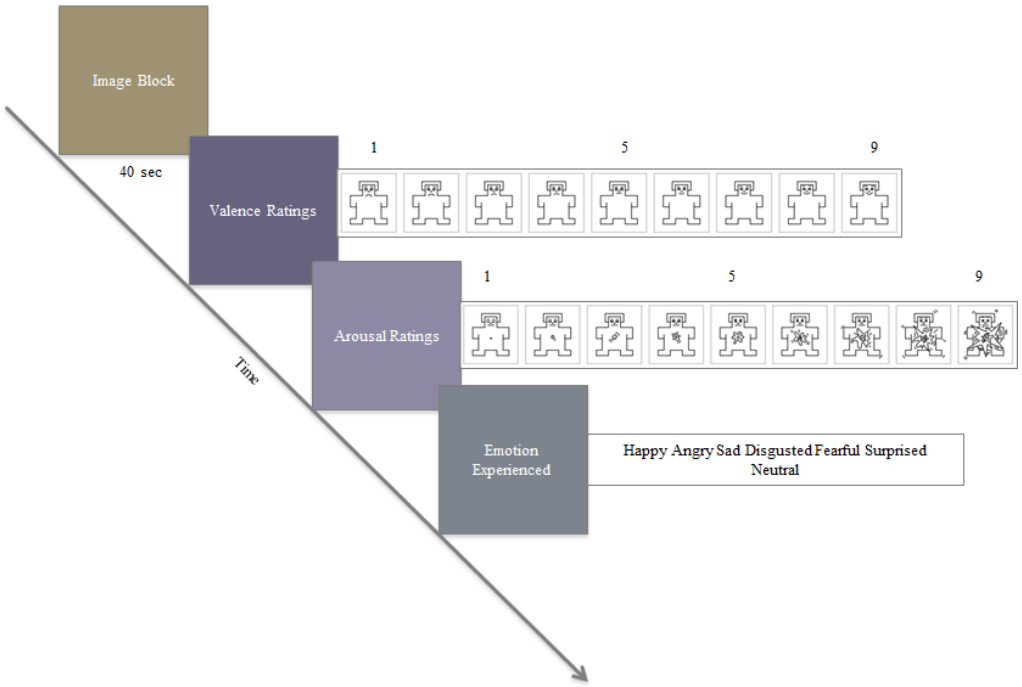


Fig. 30. IAPS pictures presentation system with the arousal rating.

tive and neutral. All the 6 categories consisted of 4 pictures each. The erotica pictures were selected appropriately for men and women subjects. After a 10 second presentation of the picture, the subjects were presented with choices to rate their emotional experience on how aroused the pictures in the preceding category made them feel (in a pictorial scale of 1-9, see Fig. 30), the valence of the emotion they felt (in a pictorial scale of 1-9) and the actual emotion they felt (out of 5 emotions and neutral). The subjects were seated around 70-80 cm from the computer screen during the whole IAPS pictures presentation session.

P. VR Emotion Presentation

We describe how the overall integrated system was designed. The system was originally designed for autism spectrum disorder (ASD) intervention (Esubalew Bekele, Zhi Zheng, et al., 2013). Given that there are several commonalities in social impairments between SZ and ASD (Cheung et al., 2010) such as atypical emotion processing and lack of affective display, the original system was deemed appropriate for use with SZ patients with minor modifications. Since the readership for ASD and SZ is likely to be different and since the experiment cannot be described without system details, we present the complete VR-based system in this

paper. The VR system was composed of three major components: the task presentation environment, the eye tracking component, and the peripheral physiology monitoring component. The task presentation environ-

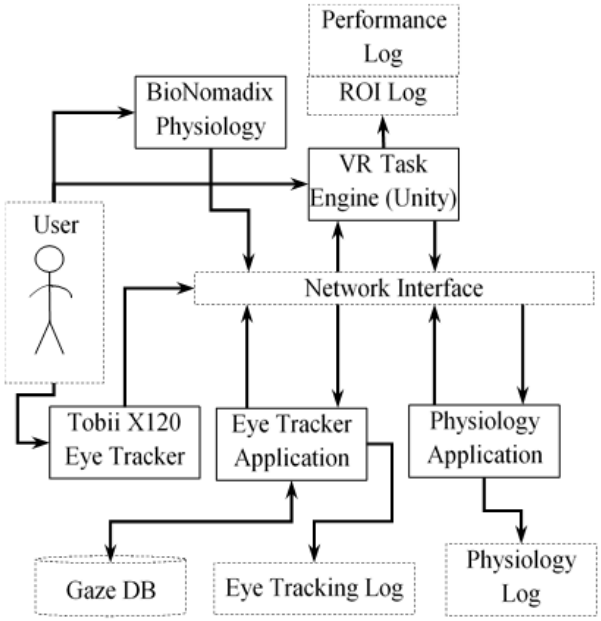


Fig. 31. Overall System Diagram.

ment was based on Unity3D game engine by Unity Technologies (www.unity3d.com). A remote desktop eye tracker by Tobii Technologies (www.tobii.com) called Tobii X120 was employed for gaze tracking. A wireless physiological signals acquisition device called BioNomadix by Biopac Inc. (www.biopac.com) with 8 channels of peripheral physiology electrodes was used to record the physiological signals. Each component ran separately while communicating via a network interface (Fig. 31).

Facial emotional expressions and lip-syncing were the major animations of this project. Range of weights were assigned from 0 (no deformation) to 20 (maximum deformation) for each emotional expression. The universally accepted seven emotional expressions proposed by Ekman were used in this ASD project (Ekman,

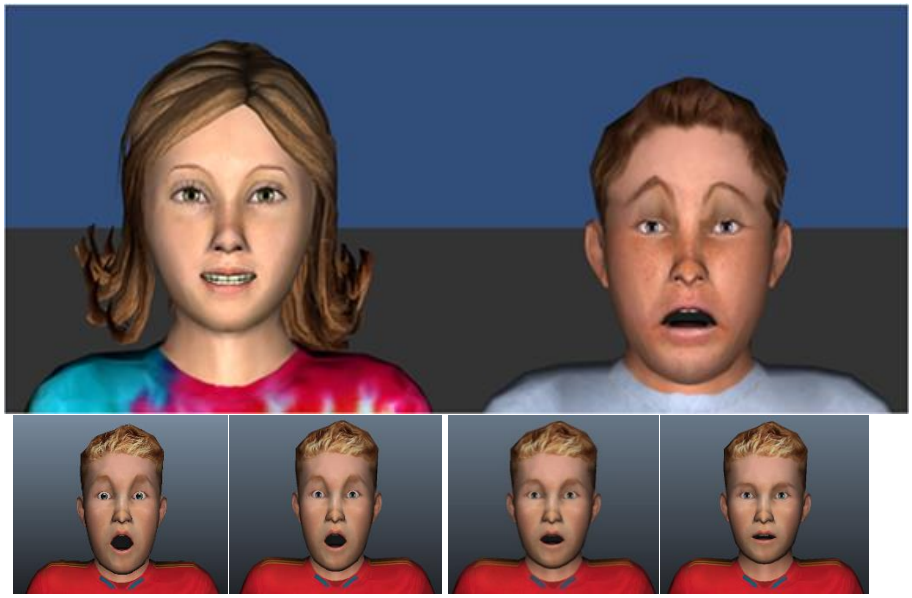


Fig. 32. Example emotions with its different degrees of arousal.

1993). The expressions were: enjoyment, surprise, contempt, sadness, fear, disgust, and anger. The range of each facial expression had four arousal levels: low, medium, high, and extreme. The four levels were chosen by careful evaluation by clinical psychologists involved in this project. In addition to these facial expressions, seven phonetic viseme poses were created using the same set-driven key controller technique. The phonemes are L, E, M, A, U, O, and I. These phonemes were used to create lip-synced speech animations for storytelling. The story was used to give context to the emotional expressions. A total of 16 stories were lip-synced for each character. Due to the similarity of emotion recognition impairment in ASD and SZ, we customized the system to suit the new target group with 5 emotions (joy, surprise, fear, anger, and sadness). The avatars were customized and rigged using an online animation service, mixamo (www.mixamo.com) together with Autodesk Maya. All the facial expressions and lip-syncing for contextual stories narrated by the avatars were animated in Maya. A total of seven avatars including 4 boys and 3 girls were selected. Close to 20 facial bone rigs were controlled by set driven key controllers for realistic facial expressions and phonetic visemes for lip-sync. Each facial expression had four arousal levels (i.e., low, medium, high, and extreme, see Fig. 32). A total of 315 (16 lip-synced stories + 28 emotion expression plus neutral for each character) animations were developed and imported to Unity3D game engine for task presentation.

Before we proceeded any further, we conducted a separate pilot study with a control group of college students to determine which of these emotions were ambiguous in our design. We found that most subjects confused contempt with disgust and fear with anger. Thus we dropped contempt and fear and used the rest 5 emotions in our tasks. The created animations and characters were imported into unity and unity was used as the main VR engine to present the emotion facial expressions.

The logged data was analyzed offline to illustrate differences in physiological and gaze responses between the patient and the control groups.

Q. Peripheral Monitoring Components

The eye tracker recorded at 120 Hz frame rate allowing a free head movement of 30 x 22 x 30 cm (width x height x depth) at an approximately 70 cm distance. We used two applications connected to the eye tracker: one for diagnostic visualization as the experiment progressed and another one to record, pre-process and log the eye tracking data. The main eye tracker application computed eye physiological indices (PI) such as pupil diameter (PD) and blink rate (BR) and behavioral indices (BI) (Lahiri, Warren, et al., 2012) such as fixation duration (FD) from the raw gaze data (Fig 33).

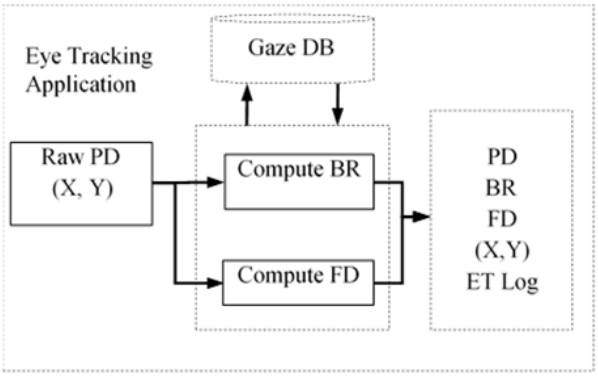


Fig. 33. The Eye Tracking Application Components.

The wireless Bionomadix physiological monitoring system with a total of 8 channels of physiological signals was running at 1000 Hz. The physiological signals monitored were: pulse plethesymogram (PPG), skin temperature (SKT), galvanic skin response (GSR), 3 electromyograms (EMG), and respiration (RSP) (Fig. 34). Due to the apparent disconnect between what patients with SZ feel and their outward expressions, they are not usually expressive of their internal affective states and these states often are not visible externally (Bleuler, 1911). Physiological signals are, however, relatively less affected by these impairments and can be useful in understanding the internal psychological states and pattern (Hempel et al., 2005). Among the signals we monitored, GSR and PPG are directly related to the sympathetic response of the autonomic nervous system (ANS) (Cacioppo et al., 2007).

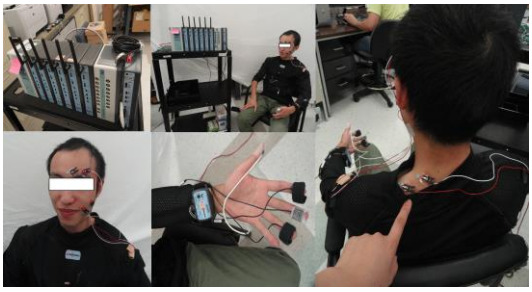


Fig. 34. Peripheral physiological electrodes placement.

R. Offline Data Analysis

The collected physiological data were processed to extract useful features and decipher any differences between the two subject groups for conditions of selected emotional expressions presentation and neutral baseline condition. We specifically chose features from PPG, GSR, RSP and SKT for this analysis. These features were chosen because of their correlation with engagement and emotion recognition process as noted in psychophysiology literature (Cacioppo et al., 2007; Hempel et al., 2005; C. Liu, Conn, K., Sarkar, N., Stone, W., 2008a; Welch, 2009). The PPG were used to extract heart rate (HR), which is a cardiac index used to measure stress and certain emotions (E. Bekele et al., 2013). The GSR is decomposed into two major components, i.e., phasic and tonic components, and from them features such as skin conductance response rate (SCRrate) and mean skin conductance level (SCL) were extracted. The RSP signal was used to extract the breathing rate (BR). The mean skin temperature (SKT) was obtained from the SKT signal. For the eye tracking data, we extracted the following features: pupil diameter (PD), fixation duration (FD), sum of fixation counts (SFC), saccade path length (SPL), and blink rate (BR). Statistical two sample unequal variance t-test was used to quantify the significance of the differences.

S. Methods and Procedure

Experimental Setup.
The presentation engine ran on Unity while eye tracking and peripheral physiological monitoring were performed in parallel using separate applications on separate machines that communicated with the Unity-based presentation engine via a network interface. The VR task was presented using a 24'' flat LCD panel monitor (at resolution 1980 x 1080) while the IAPS picture was presented on the same monitor with a resolution of 1024 x 768 in order to preserve the original resolution of the images. The experiment was performed in a laboratory with two rooms separated by one-way glass windows for observation. The researchers sat in the outside room. In the inner room, the subject sat in front of the task computer. The task computer display was also routed to the outer room for observation by the researchers. The session was video recorded for the whole duration of the participation (Fig. 35). The study was approved by the Institutional Review Board (IRB) of Vanderbilt University.

Subjects.
A total of 12 patients with SZ (Male: n=8, Female: n=4) of ages (M=45. 7, SD=9.4) and an age and IQ matched 12 healthy non-psychiatric subjects (Male: n=6, Female: n=6) controls of ages (M=44.9, SD=9.9) were recruited and participated in the usability study. All patient subjects were recruited through existing clinical research programs and had established clinical diagnosis (Table 11).

Table 11. Profile of the first 6 subjects in the patient group and the matched control group

Demographic Information	Healthy Controls	Schizophrenia	t	p
	Mean (SD)	Mean (SD)		
Age	44.9 (9.9)	45.7 (9.4)	-0.21	0.83
Gender	6 F / 6 M	4 F / 8 M	$chi^2 = 0.68$	0.4
Education, years	13.4 (2.1)	15.4 (2.2)	-2.2	0.03
IQ	107.1 (6.8)	106.4 (7.4)	0.22	0.82
SAPS	12.9 (7.7)	N/A	N/A	N/A
SANS	24.5 (11.4)	N/A	N/A	N/A
SPQ				
Positive	N/A	2.75 (4.1)	N/A	N/A
Negative		4.00 (3.8)		
Disorganized		2.58 (3.8)		
CPZ (mg/kg/day)	359.62 (247.6)	N/A	N/A	N/A

Premorbid intelligence was assessed using the North American Adult Reading Test (Blair & Spreen, 1989). Chlorpromazine equivalent (mg/day; (Andreasen, Pressler, Nopoulos, Miller, & Ho, 2010)). Semi-structured clinical interviews assessing symptoms over the past month. Brief Psychiatric Rating Scale (BPRS; (Overall & Gorham, 1962)); Scale for the Assessment of Positive Symptoms (SAPS; (Andreasen, 1984; Ikezawa, Corbera, Liu, & Wexler, 2012)); and the Scale for the Assessment of Negative Symptoms (SANS; (Andreasen, 1983)). SPQ = Schizotypal Personality Questionnaire (Van Orden, Limbert, Makeig, & Jung, 2001); CPZ = Chlorpromazine Equivalent Dose; N/A = not Applicable

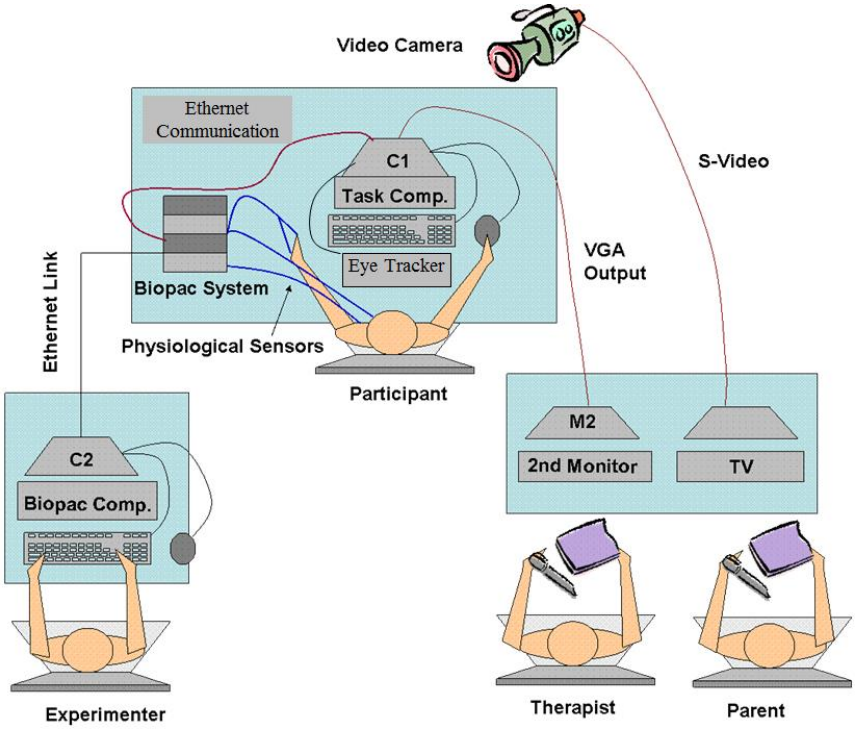


Fig. 35. Experimental setup.

The control group was recruited from the local community. The IQ measures were used to potentially screen for intellectual competency to complete the tasks.

Tasks.

The VR-based system presented a total of 20 trials corresponding to the 5 emotional expressions with each expression having 4 levels. Each trial was 12-15 s long. In each trial, first, the character narrated a context story that was linked to the emotional expression that followed for the next 5 seconds. The avatar exhibited a neutral emotional face during story telling. The IAPS picture was presented in such a way that each category was presented as a block and rating was performed after each category resulting in a total of 6 trials of 10 s for each picture in the category whereas ratings in the VR systems was after each trial of emotion expressions. Therefore, each IAPS trial consisted of four pictures from the same category. It has to be noted that all the four pictures in a category were selected carefully for equivalence as far as eliciting equivalent emotional responses were concerned. A typical laboratory visit was approximately one hour and 30 minutes long. During the first 15 minutes, a trained therapist prepared the subject for the experiment by placing the physiological sensors on the participant. Before the task began, the eye tracker was calibrated. The calibration was a fast 9 points calibration that took about 10-15 s. At the start of each task, a welcome screen greeted the subject and described what was about to happen and how the subject was to interact with the system. Immediately after the welcome screen, the trials started. At the end of each trial, questionnaires popped up asking the subject what emotion he/she thought the avatar displayed and how confident he/she was in his/her choice in the VR system. The questionnaires for the IAPS pictures asked the level of arousal and valence of the emotion they felt together with the emotion they felt by watching the pictures. The emotional expression presentations were randomized for each subject across trials to avoid ordering effects. To avoid other confounding factors arising from the context stories, the stories in the VR session were recorded with a monotonous tone and there was no specific facial expression displayed by the avatars during that context period. Totally, each subject performed three consecutive tasks in a visit. The IAPS pictures presentation task, the VR facial expression task with the context stories and the VR facial expression task without the contextual stories.

T. Results

The Collected eye gaze and physiological data for both groups were analyzed as described in the offline data analysis section above. In this section, we first present the comparative feature level analysis for both eye gaze and physiological data. Subsequently we present clustering analysis based on physiological data to show discriminability of elevated responses towards the VR emotional responses. For the feature level comparison, five features each from gaze as well as physiological signals were extracted to compare elicited responses during the facial emotional recognition tasks in the virtual environment. The data from prominent positive (joy and surprise) and prominent negative emotions (anger and disgust) were combined with high and extreme levels of arousal to generate the positive and negative category dataset, respectively. We also extracted baseline features for the physiological data to compare the responses in these categories in the clustering analysis. For gaze analysis various regions were defined as shown in Fig 36.

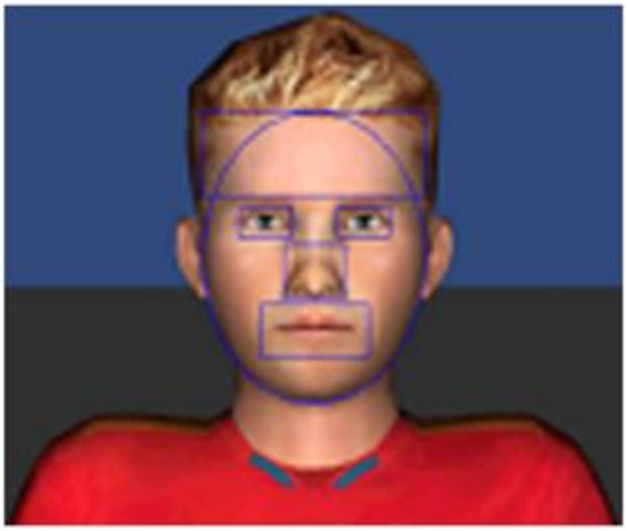


Fig. 36. Face ROIs used for gaze analysis.

Eye Gaze Features Analysis

Most of the eye gaze indices showed differences between the two groups of participants (Table 12). In all the cases considered, the patient group showed more saccadic eye movement as indicated by higher SFC, lower FD and shorter SPL. These indices are known to correlate with one's engagement (U. Lahiri, Warren, Z., Sarkar, N., 2011) and pupil constriction was associated with engagement (Christa J Anderson, Colombo, & Jill Shaddy, 2006). Therefore, based on these data, the patients were less engaged than the control group in the VR session. Moreover, notably lower blink rates for the patient group indicated abnormal gaze pattern with lower than average blinks per minute while processing such stimuli (U. Lahiri, Warren, Z., Sarkar, N., 2011). Pairwise statistical t-test analysis indicated that during the positive emotions presentations the patient

Table 12. Eye Tracking Features for the VR session

			PD (mm)	FD (ms)	SFC (N/A)	SPL (pix)	BR (bpm)
Positive Category	SZ	Mean	2.63	155.98	42.58	70.32	1.83
		SD	1.15	174.67	31.00	80.13	1.16
	CTR	Mean	2.87	308.35	51.98	83.81	3.31
		SD	0.39	273.85	42.12	45.32	1.66
Negative Category	SZ	Mean	2.51	147.53	39.94	58.00	2.06
		SD	1.22	129.06	27.18	40.30	1.49
	CTR	Mean	2.90	329.82	53.94	77.55	3.08
		SD	0.38	304.77	43.55	42.22	1.95

group had statistically significantly lower fixation duration (SZ M=155.98, CTR M=308.35, $p<0.05$) and blink rates (SZ M=1.83, CTR M=3.31, $p<0.05$) as compared to the healthy control group. During the negative emotions presentations, 4 out of the 5 eye features including PD (SZ M=2.51, CTR M=2.9, $p<0.05$), FD (SZ M=329.82, CTR M=147.53, $p<0.05$), SPL (SZ M=58, CTR M=77.55, $p<0.05$), and BR (SZ M=2.06, CTR M=3.08, $p<0.05$) were statistically significantly lower for the patient group when compared with the control group. No statistically significant difference arose between the positive and negative emotion presentations within each group. Similar pattern of less engagement was observed for the patient group as compared to the control group during the VR session even when emotion without the contextual stories were presented.

Intergroup Eye Gaze Visualization

To determine if there were any differences in gaze scanning pattern between the patient group and the control group and between positive emotions and negative emotions intragroup, we performed clustering of the gaze points on to distinct facial regions of the face of the avatars, which are shown in Fig. 5. For qualitative analysis of these gaze clusters, we generated heat map and masked map visualizations of the scanning gaze patterns to compare group variations. We generated visualization for all the five emotional expressions that were presented in the VR experiments for all the avatars averaged across all the participants in each group. In almost all the cases, the patients exhibited asymmetrically atypical scanning patterns when compared to the control group. Fig. 37 shows an example comparison for one avatar between the patient (top) and the control (bottom) groups for the emotion anger averaged across all trials and all participants within each group. It can be observed that the patients were looking in an asymmetric pattern and were focused more on context irrelevant areas such as the forehead and even outside of the facial region. On the contrary, the healthy control group participants focused more on the context relevant areas such as the mouth and eyes with a more symmetric scanning gaze pattern. In summary, the visualizations indicate atypical facial scanning pattern by the patients group with more asymmetric and dispersed gaze pattern as compared to a more symmetric concentrated gaze to the important regions of interest such as the eyes and the mouth by the healthy controls.

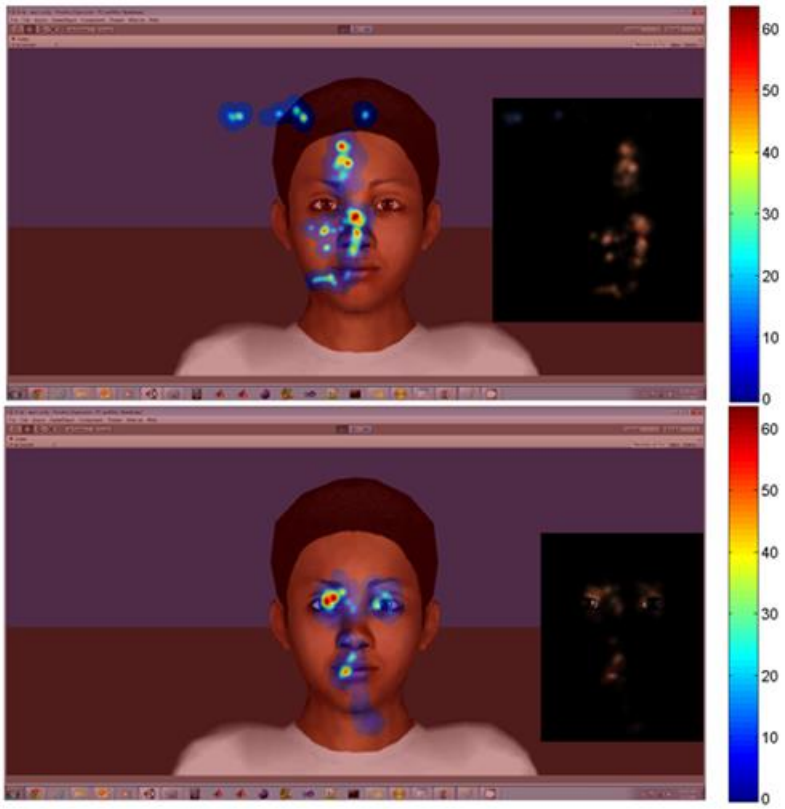


Fig. 37. Masked maps overlaid on heat map visualizations of the patient group (top) and control group (bottom).

Physiological Features Analysis

As shown in Table III, the patient group had higher emotional response indicators including higher heart and skin conductance response rates and comparable breathing rate compared to the control group on both the positive and negative emotion presentations. Both skin temperature (SKT) and skin conductance levels (SCL) were statistically significantly different with patients exhibiting significantly lower SKT and SCL in both the positive (SKT: SZ/CTR M=90.27/94.81, SCL SZ/CTR M=6.79/9.24, $p<0.05$) and negative (SKT: SZ/CTR M=90.16/94.96, SCL SZ/CTR M=6.49/8.79, $p<0.05$) emotions categories. Also, the breathing rate (BR) was higher in the negative emotions category for the patient group (BR SZ P/N M=17.96/22.39, $p<0.05$).

Table 13. Physiological Features for the VR session

			HR (bpm)	SKT (^{0}F)	BR (bpm)	SCL (μS)	SCR (rpm)
Positive Category	SZ	Mean	87.83	90.27	17.96	6.79	3.12
		SD	13.83	9.17	9.84	3.65	2.79
	CTR	Mean	85.43	94.81	18.49	9.24	2.42
		SD	15.41	2.48	6.16	6.87	2.77
Negative Category	SZ	Mean	87.88	90.16	22.39	6.49	2.91
		SD	14.37	9.10	12.55	3.41	2.70
	CTR	Mean	86.53	94.96	20.67	8.79	2.13
		SD	14.08	2.57	11.23	6.50	2.65
Baseline	SZ	Mean	83.02	90.41	19.71	6.77	1.83
		SD	9.28	8.90	3.57	4.17	0.74
	CTR	Mean	83.39	95.01	18.56	9.75	2.19
		SD	10.87	2.01	3.99	7.68	0.73

Features for the baseline where participants were monitored for 3 minutes without any stimuli at the beginning of each session are shown in Table 13. Generally, heart rate increased from the baseline in both positive and negative conditions for both groups. Breathing rate (BR) was greater in the negative category and slightly lower in the positive category for both groups. SKT and SCL remain closely similar or slightly lower to that of the baseline in both conditions for both groups. SCR increased significantly from baseline in both positive and negative conditions for the SZ group. For the control group, SCR was slightly higher in the positive condition but slightly lower for the negative condition. Greater SCR of SZ group in both the positive and negative categories suggests higher phasic arousal, irrespective of the stimuli; this result indicates an undifferentiated response to emotional conditions in schizophrenia.

Physiological Clustering Analysis

To get a clearer understanding of which features are responsible for the differences between the two groups with the two conditions and the baseline, we further explored physiological responses by performing clustering of data points from trials of the positive and negative conditions with the full 8 channels of physiology and 81 features extracted out of them. Although, the limited feature level comparison with the selected 5 features are easy to apply to demonstrate response differences between the two groups and within each group, feature level analysis is not suitable for the whole 81 feature set.

The clustering analysis helps probe whether there are differences in physiological response pattern in the presence of positive and negative emotional stimuli as compared to the baseline and also with each other in both groups. If clustering analysis reveals clear pattern differences, supervised models could be built for a closed loop physiology based online interaction in the future. Two clustering methods were used in this work, which were the widely used k-means and a recent technique based on density peaks (Green, Williams, &

Davidson, 2003). However, we observed that k-means had difficulty in extracting non-spherical clusters and hence was not able to capture highly non-linear pattern differences. Moreover, results from the density clustering method outperformed those from k-means for this particular dataset. Therefore, in this paper, we present our analysis from the density based clustering method. Fig. 38 shows example clustering analysis for the control (top) and the SZ (bottom) group using the density peak clustering method (Green et al., 2003). The decision graph indicates minimum distance to density peak points from each data point (δ) versus the cluster density around each data point (ρ). Cluster centroids are selected using product of these two parameters, meaning points with higher density that are surrounded with other higher density points are likely cluster centroids. Using both these cues, cluster centroids were chosen and the resulting clustering is indicated in the bottom multidimensional scaled dissimilarity of the data points. For details of the clustering methods, please, refer to (Green et al., 2003). We compared the negative and positive emotions of the VR stimuli for both groups. We additionally compared these two emotional stimuli with the baseline physiological responses of each group. We further performed sequential forward feature selection to see the most prominent features in the physiological responses.

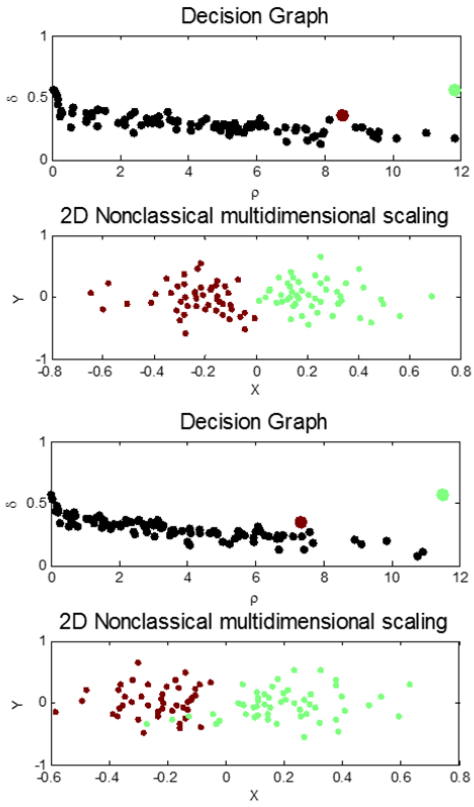


Fig. 38. Physiological clustering using the density peaks method. Positive vs. negative categories for SZ (top) and CTR (bottom).

Fig. 39 presents the clustering results from the density clustering method. Accuracy was computed as a percentage of data points that are correctly clustered using ground truth clusters. Each pair of comparisons out of the two conditions, i.e., positive (pos) and negative (neg) emotion categories, the two groups, i.e., schizophrenia (SZ) and healthy control (CTR) and the baseline (BL) are taken as the two clusters in each

comparison. The horizontal axes indicate the features included in the sequential forward selection (SFS) feature selection method. Feature [1, N] means features from 1 to N including features 1 and N. The top and the middle plots show that responses to positive and negative emotions were significantly greater than the baseline in both groups. This demonstrates that the VR stimuli was successful in eliciting emotional responses compared to baseline. The results also indicate that there is no clear difference between responses to the positive and negative emotion categories, demonstrated by the low accuracy, close to chance or 50%, for the two clusters. Lower clustering accuracy indicates absence of differences between the two clusters considered. Of particular importance is the variation in features as shown by the results of SFS. In both groups and conditions compared with the baseline, it appears that facial EMG features (features 10 to 39) and respiratory related features (features 76 to 81) contributed more pronounced emotional responses. GSR features (features 40 to 45) were the next contributing features. The cardiovascular features (features 1 to 10) did not produce sufficient response in both groups.

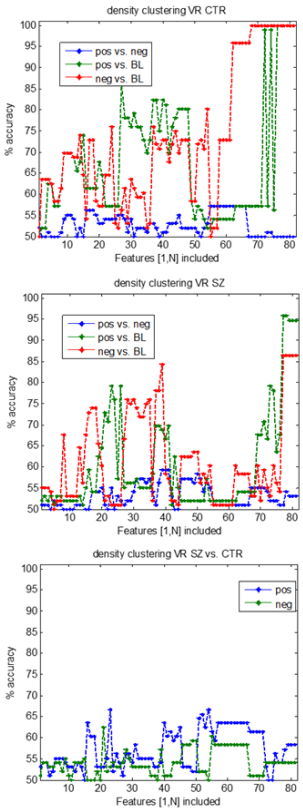


Fig. 39. Clustering accuracies using the density peaks method.

Performance Data Analysis

Three performance metrics were considered for analysis, i.e., percentage performance of correctly identifying the emotional expressions, how confident the subjects were with completely confident being 100% and completely unconfident being 0%, and the time the subjects took to make their selection once the emotion presentations were finished. The performance data was first collapsed into the two virtual reality sessions,

i.e., VR facial expression with vignette stories (VR1) and VR facial expression without stories (VR2). Fig 40 indicates that there were performance differences between the two groups in both sessions. The patients exhibiting lower performance compared to the control group in both VR1 and VR2. However, these differences were not statistically significant. Significant performance differences arise between the two VR sessions within the patient (VR2/VR1 M=78.75%/62.5%, $p<0.05$) and the control groups (VR2/VR1 M=86.67%/70.42%, $p<0.05$). These performance differences between the two groups could be explained in relation to the vignette stories in VR1. The addition of contextual stories that were recorded with monotonous tone (no vocal emotional prosody) may have introduced greater ambiguity that interacted with the identification of emotion expressed by the avatar. This finding suggests that isolated emotion recognition is easier than emotion recognition in the presence of contextual social interaction. The capacity and flexibility of VR for presenting and manipulating multiple attributes, features and components of stimuli dynamically are big advantages that the traditional static pictures and non-modifiable video clips cannot emulate. VR environment can be programmed to present a very wide range of vocal intonations and other social interaction parameters to maximize the potential training of interpretation of emotional faces in a social context.

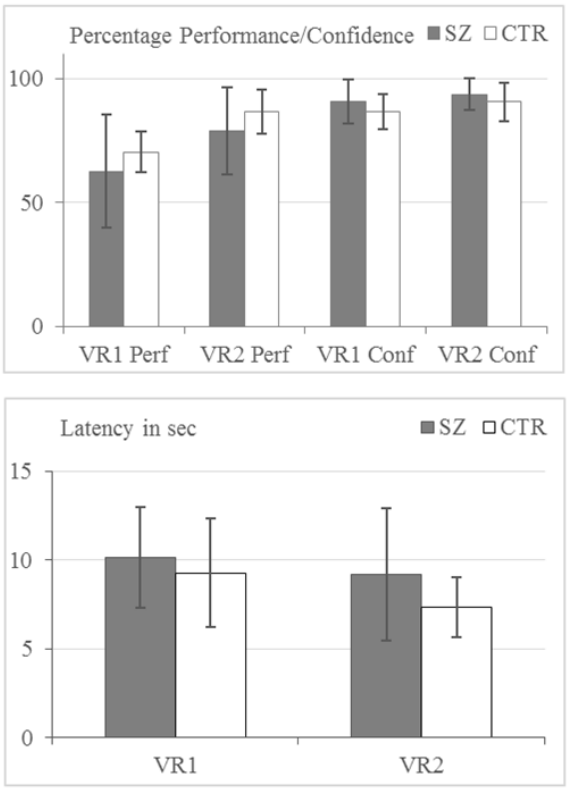


Fig. 40. Overall performance metrics.

Patients and controls did not differ in their confidence ratings for recognition responses in both the VR sessions. The control group were significantly more confident in VR2 than VR1. This suggests that the monotonous (no prosody) stories could have contributed to the misrecognition; introduction of the recorded

vignettes may have reduced confidence in their ability to correctly identify the emotion expressed the avatars. This suggests that monotonous presentation of social context may not be the best approach moving forward in the development of social training VR programs. On the last performance metric, the response latency, the patient group took relatively more time to make their choices than the control group in both VR sessions. However, these differences were not statistically significant.

Table 14. Response Bias of subjects for each Emotion

Subject Response	VR1			
	SZ		CTR	
	n	%	n	%
Anger	56	23.33%	49	20.42%
Disgust	48	20.00%	59	24.58%
Joy	35	14.58%	32	13.33%
Sadness	56	23.33%	59	24.58%
Surprise	45	18.75%	41	17.08%
	VR2			
Anger	55	22.92%	41	17.08%
Disgust	44	18.33%	54	22.50%
Joy	46	19.17%	35	14.58%
Sadness	43	17.92%	50	20.83%
Surprise	52	21.67%	60	25.00%

Emotion recognition performance is known to be biased by subjects' preferred choice regardless of the stimuli presented, i.e., personal bias (Karson, Bigelow, Kleinman, Weinberger, & Wyatt, 1982). To account for personal bias, we first broke down the performance into each emotion. The top plot of Fig. 41 shows that of the 5 emotions presented, the control group had a higher performance than the schizophrenia group except on anger in both VR sessions and joy (enjoyment) in VR2 session. Moreover, both groups had a higher performance in VR2 session than in VR1 in all emotions except disgust for the patient group.

We then computed the response bias, in which subjects reported a certain emotion regardless of the stimuli. Table 14 shows the number and percentage response biases. The expected response bias for 5 emotions is 20%. Most of the biases are not far from 25%, which indicates balanced responses. Emotions with response bias that are too far from the expected 20% value are generally biased. We used these emotion response biases to adjust the raw performance by dividing the raw performance by the percentage bias and rescaling to 100%. The bottom plot of Fig 10 shows the response bias corrected performance. After the bias adjustment, in VR1, the control group performed higher than the patients in all emotions except disgust whereas they performed higher than the patient group in all emotions except surprise in VR2. The notable result after the bias adjustment is, both groups performed better in VR2 than VR1 in all the emotions. Taken together, these results suggest that patients do not show an erratic or idiosyncratic response bias to the emotional expressions

of the avatars. Such similar performance profiles might indicate that patients are able to pick up on the emotion specific signals in the expression, albeit, at a slightly more impaired level.

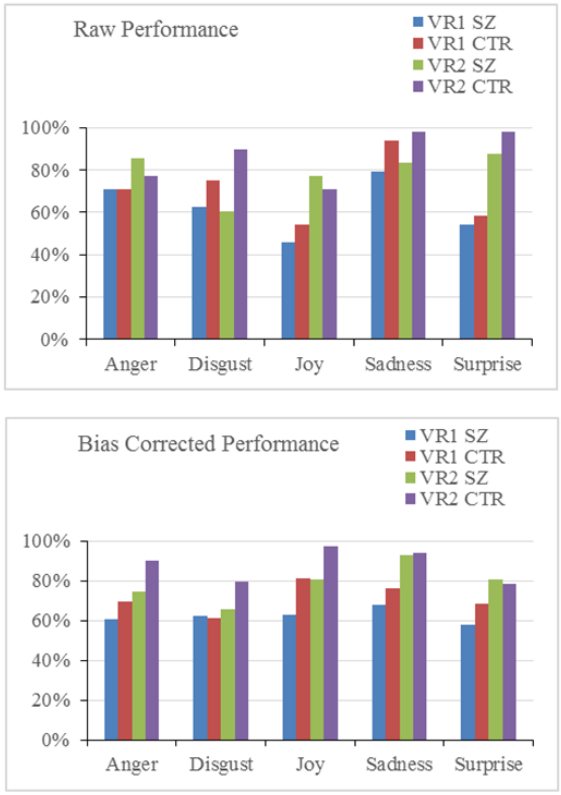


Fig. 41. Per emotion raw (biased) and bias-corrected performance.

U. Discussion and Conclusion

The objective of this research was to design and test a VR-based schizophrenia intervention system that would offer opportunities to analyze emotion recognition mechanism in patients with SZ that might not be possible either in static picture or isolated video clip based protocols. This novel system, VR-SAAFE, allows presentation of realistic facial emotional expressions in VR that can be precisely controlled both in terms of valence and arousal and can be embedded with contextual vignettes. Furthermore, VR-SAAFE is integrated and synchronized with both physiology and eye gaze data acquisition capabilities such that targeted physiology and gaze based analysis can be performed during the emotion recognition tasks. Since emotion recognition is a dynamic task, such abilities will likely have positive impact in understanding the process of emotion recognition in patients with SZ and eventually designing new intervention in the future. We are not aware of any similar systems in the literature.

We performed a pilot study to test the usefulness of VR-SAAFE. Eye tracking and physiological signals were collected during the VR task and analyzed offline. While these data are preliminary and based on one pilot study, they provide interesting findings and thus indicate the usefulness of such a system as VR-SAAFE.

The results from gaze and physiological feature level analysis show that they are viable indicators of internal emotional states of patients with SZ. The patient group overall responded more intensely to the positive emotion presentations than both the negative and neutral (baseline, in the case of VR) emotion conditions for almost all the features.

Both the IAPS and the VR systems were able to present the facial emotional expression trials successfully. Eye tracking and various physiological signals were collected and analyzed offline. The results from gaze and physiological feature level analysis show that they are viable indicators of internal emotional states of patients with SZ although their self-reporting can be biased by their emotion processing and understanding impairments.

The analysis of the eye gaze features indicated that the SZ group exhibited shorter fixation durations and scan-path lengths which corroborate previous eye tracking literature using static faces (Green et al., 2003; Manor et al., 1999; Williams, Loughland, Gordon, & Davidson, 1999). With shorter scan-path lengths and shorter fixation durations, individuals with SZ did not strategically visually scan the face in a manner that is conducive to perceiving emotional expression. As the visualizations show qualitatively, the SZ group showed an aberrant pattern of fixation allocation to the important features of the face. Additionally, the SZ group displayed reduced pupil dilation and blink rates while performing the task. Both of these indices are associated with engagement (Van Orden et al., 2001) and cognitive effort (Piquado, Isaacowitz, & Wingfield, 2010) during a task suggesting that the SZ group did not engage with the task to the same extent as the control group. This interpretation is tempered with the fact that the all individuals in the SZ group were currently taking antipsychotics when they participated and the association between dopamine and oculomotor function may have played a role in the group differences found (Karson et al., 1982).

The results from the physiological feature and cluster analyses suggest that – at a feature level – there are some indications of physiological differentiation based on emotional valence. Specifically, the control group showed differences between the positive and negative stimuli in their skin temperature and skin conductance. These two signals were also significantly different from the SZ group for both emotional valence conditions. These findings are in line with previous studies finding reduced tonic skin conductance responses in a subset of individuals with SZ (Bernstein et al., 1982; M. E. Dawson & Nuechterlein, 1984; Ikezawa et al., 2012). Interestingly, in the SZ group, there was a significant increase in the phasic skin conductance response rates, which was not seen in the control group. The increase in SCR during the social cognition task suggests that there was a failure to habituate (Gruzelier & Venables, 1972; Ikezawa et al., 2012). These indicators provide some initial evidence that they could be used in future social cognition interventions that take into account one's emotional state when interacting with others.

This conclusion is bolstered by the findings of the cluster-level analyses, which indicate that, across both groups, one's emotional state during the social cognition tasks is distinguishable from their baseline state. This suggests that there was a distinct change to participants' physiological states when performing the task and provides preliminary evidence for the potential use of this physiological state monitoring system in social cognitive intervention programs. To be able to track an individual's physiological state while interacting with the VR-based intervention system, it may be possible to tailor the interactions to best suit the particular individual. In this initial study, the interactions were fairly limited (e.g., presentation of emotional expressions accompanied by social vignettes) and still a difference in physiological state was found in the participants. Future studies may be able to capitalize on this and incorporate more complex social interactions, which may provide even richer physiological information to use in understanding the emotional and physiological states. Importantly, the current study did not find a significant effect of valence on the cluster-level analysis of the physiological data. This is perhaps to be expected given that GSR was one of the strongest predictors in the cluster analysis and GSR is believed to be more associated with arousal than valence (Bradley, Cuthbert, & Lang, 1996). Taken together, the current system provides initial evidence for its use in understanding emotional states of individuals via analysis of physiological signals when engaging in social cognitive tasks.

Finally, while there were no group differences in performance on the VR tasks, both groups performed better on the task wherein emotional expressions were presented without verbal social vignettes. As noted

above, this performance differential may have been due to the monotonous vocal intonation in which the vignettes were presented may have increased uncertainty in the participants. By removing the emotional prosody in the speech, the incongruence between auditory and visual cues may have made it more difficult for the participants to correctly identify the emotions expressed by the avatars. The fact that the control group reported significantly greater confidence in their performance on the VR task with just the emotional expressions supports this. Future investigations should take care to provide congruent prosodic presentations of social vignettes in order to better simulate the multimodal presentations of emotional and social information in social interactions.

This preliminary study could inform future adaptive VR applications for SZ therapy that could harness the inherent processing pattern of patients with SZ as captured from their gaze and body physiological signals. Such implicit mode of interaction is advantageous over performance-only interactions for objective, extensive, and natural interaction with the virtual social avatars. Despite several limitations related to the design of the emotional expressions in the VR system and limited interactivity in the current system, this initial study demonstrates the value of future adaptive VR-based SZ intervention systems. For example, the ability to subtly adjusting emotional expressions of the avatars, integrating this platform into more relevant social paradigms, and embedding online physiological and gaze data to guide interactions to understand psychological states of patients with SZ could be quite useful tools. We believe such capabilities will enable more adaptive, individualized and autonomic therapeutic systems eventually.

CHAPTER VI

MULTIMODAL ADAPTIVE SOCIAL INTERACTION IN VIRTUAL ENVIRONMENT (MASI-VR) FOR CHILDREN WITH AUTISM SPECTRUM DISORDERS (ASD)

The robot-mediated and virtual reality based assistive systems in ASD and SZ intervention demonstrated that intelligent systems hold promise in alleviating several limitations of traditional intervention. With the ability to control and modify interactions, incorporate implicit and explicit cues to change interaction, and measuring objective performance metrics, technology-assisted therapy could augment or eventually serve as self-contained therapeutic tools. The VR works specifically show variations in implicit cues such as body physiological signals and gaze processing patterns when the subjects were presented with isolated facial emotional expressions even in the absence of outward display of emotion or any performance differences.

Building on the momentum of results achieved of the use of multimodal interfaces in a VR environment, this chapter describes a more comprehensive VR-based intervention system for adaptive multimodal social interaction and emotion recognition in a social context that incorporates more multimodal interfaces than the completed systems. The current system also contains more improved interaction in the form of speech-based conversational dialog and allowing the subject to move freely around the virtual scene. This system not only expands the number of multimodal interaction interfaces for monitoring, but also, incorporates real-time gaze sensitive feedback and task modification based on gaze contingencies.

We administered the conversational paradigms as they are currently developed, in a game-format, utilizing the emotion recognition task from the previous facial expression protocol as a Pre- and Post-test measurement. Embedded within the social task, were emotion recognition questions. These were presented at the end of each conversation (i.e., upon successful completion of the mission). Hence regardless of the number of trials needed to be successful (1-3), the emotion was presented. If a participant is not successful in the conversation after 3 trials, the emotion recognition question was still displayed. Two separate systems were developed. One is based on purely subject performance and the other one is based on gaze contingency. Participants were randomly assigned to one of the two systems. Participants in the performance-based system received feedback simply based on their performance. Participants in the enhanced gaze-contingent system were unable to answer the emotion recognition questions until s/he spent time looking at certain regions of the face. Initially the face was occluded, but the participant was able to reveal core components of the face by where they were looking. Once core components of the face are revealed, they were able to answer the emotion identification questions.

Developing such multimodal and adaptive systems is a step in the direction of fully adaptive individualized intervention that is required to assist the therapist in the near future and emulate the traditional therapy-centric intervention in a long term.

V. System Design

Day-to-day social behaviors are expressions of one's attitude towards social situation and interaction and are manifested through verbal conversations and various non-verbal behavioral cues such as facial expressions, body postures and gestures, and vocal outbursts like laughter (Fig. 34 Top)(Vinciarelli, Pantic, & Bourlard, 2009). For a system to realistically model social communication in a simulated virtual environment, it needs to address these various aspects of social communication. The proposed system is endowed with capability

to capture most of these verbal and non-verbal cues to facilitate a more natural, individualistic and adaptive virtual social interaction.

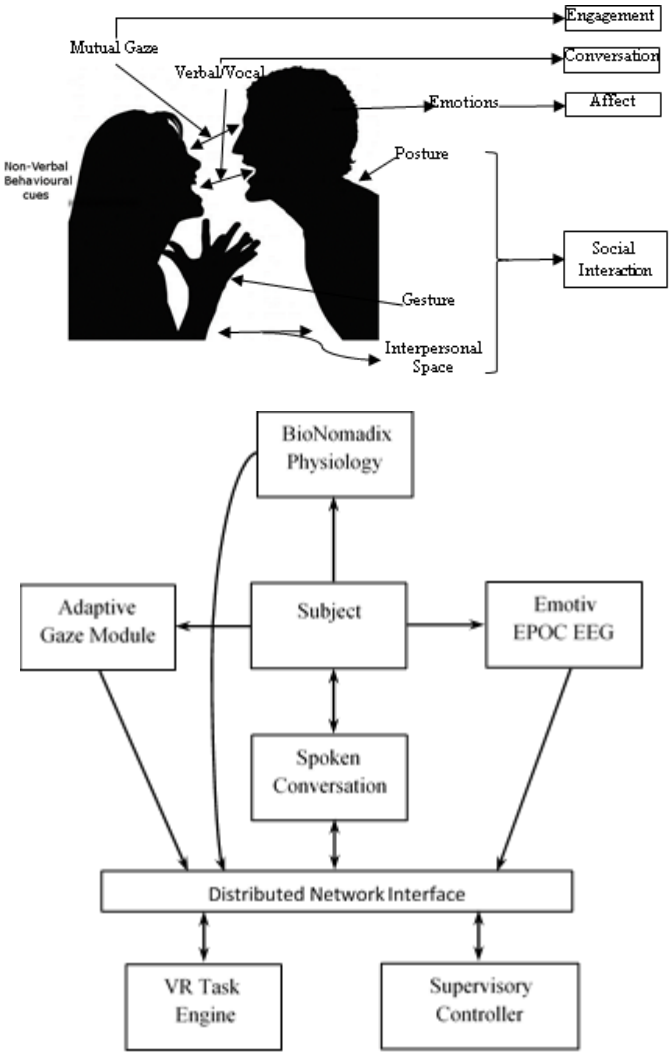


Fig. 42. (Top) Components of emotional social interaction and (bottom) system architecture of MASI-VR.

We have designed the VR system to model the different aspects of an emotional social interaction (Fig. 42 (top)). The VR system for adaptive multimodal social interaction is composed of 5 major components: (1) an adaptive social task presentation VR module, (2) a spoken conversation management module (Q/A-based dialog management module), (3) a synchronous physiological signal monitoring module, (4) a synchronous EEG monitoring, and (5) a synchronous eye tracking and online adaptive gaze feedback module (Fig. 1 (bottom)). All separate components of the system run independently in parallel while sharing data via light-weight network sockets message passing.

The overall system architecture is given in the bottom part of Fig. 42. The system is a distributed system with a central supervisory controller. Each peripheral interface components connects to the MASI-VR task engine using a distributed modular network interface. The peripheral components get events happening during the social interaction with a central timestamp via the supervisory controller.

The supervisory controller facilitates the event synchronization between the VR task presentation engine and the peripheral interfaces. In addition to the implicit cues collected from physiology, eye tracking and EEG, we have designed a spoken conversational dialog management module that interacts with the VR engine as one of the peripheral interfaces to provide speech recognition and dialog management services. In order to undertake naturalistic social interaction several components including conversational dialog, body language (gesture), facial emotional expressions and eye contact need to be considered. Conversational dialog is an important part of social interaction. Recently spoken conversational modules have been incorporated to VR systems to achieve more natural interaction instead of menu driven dialog management. Instead of large vocabulary, domain independent natural language understanding, limited vocabulary question-response dialog management, which is focused on the specific domain, has been shown to be effective (Kenny et al., 2007; Leuski et al., 2009). Such multimodal interaction helps in individualization and in cases of inaccessibility of trained therapists, it may serve as a self-contained therapeutic system. Proper facial emotional expressions recognition and appropriate gaze fixation pattern are considered to be an important building block for individuals with ASD to alleviate their overall social interaction impairment. Although isolated emotion recognition in VR was demonstrated earlier, proper processing of emotional faces in the presence of a social context, a dialog in this case, is of paramount importance and appropriate for the VR-based core skills training to generalize into real-world interactions. To aid in proper processing of emotional faces, we have designed a facial occlusion paradigm in which the subject sees an oval occlusion on the face and the occlusion gets revealed progressively as the subject scans the face appropriately by looking at context relevant regions of the face such as the eyes and the mouth. We developed a custom shader inside Unity to create the occlusion and continuous gaze feedback based revealing effects. Section 3 discusses the task and the protocol in detail.

The VR task presentation engine is built on top of the popular game engine Unity (www.unity3d.com) by Unity Technologies. The peripheral psychophysiological monitoring application was built using the software development kit (SDK) of the wireless BioNomadix physiological signals acquisition device by Biopac Inc. (www.biopac.com). The eye tracker application employed the Tobii X120 remote desktop eye tracker SDK by Tobii Technologies (www.tobii.com). The EEG monitoring is based on the SDK from the emotive EPOC EEG headset (www.emotiv.com).

The MASI-VR task presentation engine.

The VR environment is mainly built on and rendered in Unity game engine. However, various 3D software such as online animation and rigging service, Mixamo (www.mixamo.com), and Autodesk Maya were employed for character customization, rigging and animation. The venue for the social interaction task is a virtual school cafeteria (Fig. 43). This environment was chosen for the targeted age group (i.e. 13-17 year old), because it fosters various conversation and interactions for teenagers. The cafeteria was built using a combination of Google Sketchup and Autodesk Maya and was then imported into the Unity3D engine. A pack of 12 fully rigged virtual characters (10 teenagers and 2 adults) with 20 facial bones for emotional expressions and several body bones for various gestural animations were developed.

Fig. 43. The VR cafeteria environment for social task training. Dining area and food dispensary area. The two areas are constructed in separate rooms.

The characters used in this project were customized in mixamo to suit the 13-17 year age group targeted for the usability study. A total of 7 characters including four boys and three girls were selected and customized for the embedded facial expression display at the end of each conversational mission and another set of 12 characters for primary social interaction. Each character was rigged with a skeletal structure consisting of 94 bones. Twenty of these bones were involved with the face structure that was used for facial emotional expressions. Since the main focus of this project was displaying facial emotional expressions, greater emphasis was given to the face structure. The face rig was attached to a facial emotional controller using set driven keys. The rest of the body was provided with inverse kinematics (IK) controllers except the fingers in which case direct forward kinematics was employed. Besides the facial rig, due attention was given to the quality of the characters.

Each avatar's face was animated in Maya to generate Ekman's 7 universal emotional facial expressions (Ekman, 1993). Each of these emotions was generated with 4 degrees of arousal (e.g., low, medium, high and extreme). Fig. 36 shows some of the characters while displaying some of the gestural animations and the facial expressions they are capable of.

Fig. 44. Representative characters displaying example emotions and gestural animations.

As facial emotional expressions are major parts of the non-verbal communication cues in social interactions, we have completed the usability study of a VR-based facial emotional expression recognition using 10 typically developing children and 10 children with ASD. The results indicated that there exists inherent pattern difference in the way children with ASD processed the emotional faces and recognize them. Fig. 44 shows two examples of emotional expression and other gesture animations.

A total of 28 (7 emotional expression x 4 levels of each emotion) and 16 story lines lip-synced animations were created on one of the character. The rigs of each character were standardized to make transferring animations easier. All the animations done on one character were, then copied to all the remaining characters using attribute copying utility script. Once all the characters had all the lip-sync animations and all the emotional expression animations resulting in a total of 315 animations, they were batch exported using a utility script into Unity.

In addition to the facial expressions and gestural and some utility animations, seven phonetic viseme poses were also created for lip-syncing spoken audio by the avatars in the verbal conversation part of the social interaction. Also other gesture animations (shown in Table 15) were created for realistic social communication.

Table 15. Categories of Animations

Nonverbal Gestures	**Kinetic Actions**	**Idle Gestures**
Point Forward (L/R)	Eating	Arm Scratch
Point Side (L/R)	Sitting	Weight Shift
Air Quotes	Turning (L/R)	Looking at fingernails
Shoulder Shrug	Walking	Shoulder Scratch
Exaggerated Shoulder Shrug	Waving (L/R)	Carrying Tray Pose
Beckon	Tray Pickup	Standing Pose
Cross Arms	Tray Put Down	**Emotions**
Uncross Arms	Clapping	Joy
Head Shake	Give Object	Sadness
Exaggerated Head Shake	Take Object	Disgust
Head Nod	Card Swipe	Anger
Exaggerated Head Nod	Give Tray	Fear
Head Tilt	Take Tray	Contempt

The Spoken Dialog Management System.

All characters have all the lip-synced speech capabilities, several animations and the environment is populated with enough avatars to undertake the conversational missions (levels). Eventually the system will have both conversational and non-conversational (non-verbal) missions that would capture the social task components. The current spoken dialog management system uses already pre-defined conversational threads for each mission. We chose pre-defined conversational threads, at least for this development cycle, in order to keep this mechanism tractable while we could assess the reliability of the avatars and feedback mechanisms. The threads were designed by ASD therapists at Vanderbilt University Kennedy Center. We developed an application that automatically created XML grammar files for the speech recognition engine and serializes the conversation thread trees once the therapist created the thread intuitively using the application. To use a probabilistic question and answer dialog management with real natural language processing capabilities, however, requires a lot of vocabulary and speech corpus for training to achieve the level of accuracy that is sought for this task.

The verbal conversation component of the VR system creates context for the social interaction and emotion recognition in a social setting and is managed by a spoken dialog management module which was developed using the Microsoft speech recognizer from the speech API (SAPI) with domain specific grammar and semantics. The conversation module is based on question-answer dialog and it contained conversational threads for easy (level 1), medium (level 2) and hard (level 3) social tasks with each level having 4 conversational task blocks. Each block in a level has a mission that the participant is expected to accomplish. Each conversation block was represented by a tree of dialog with nodes representing each option and a particular branch in the tree representing the dialog alternative paths from the initial question to the final correct answer.

Failure and success is measured in each conversational block and there is a hierarchical scoring mechanism that keeps track of performance in conversation block level as well as mission level. Options in each block of conversation are presented to the participant using a list of items and the participant speaks out their choice through a microphone. Kinect is employed for this purpose as its microphones have superior sound localization and background noise cancelling features for a price of commodity hardware.

For all the conversation responses and feedback by the avatar, audio files were recorded by four separate people to randomly be lip-synced with the avatars response. Equal amount of lip sync animations were also created for realistic conversation. The mission files were created using an easy to use script editor software developed for this purpose. Fig. 45 shows the script creator software. A total 12 missions with 4 in each of the 3 levels of difficulty were created for this study. The missions in the easy levels of difficulty contain only one back and forth or one turn taking (see Fig. 46) while the missions in the medium level of difficulty contain two turn takings. The missions in the hard level contain three turn takings. Each turn taking is represented by the dialog tree shown in Fig. 38. Each level in the dialog tree is a trial in the conversation. For each turn taking, one file of conversation was created from the script creator together with 10 grammar files for the trials.

Overall performance, i.e., success/failure (S/F) is used to switch across missions (levels) as shown in Fig. 46 in the performance only version of the system. In the adaptive system, eye tracking-based engagement detection is employed to modify the interaction in addition to the overall performance of the participant for a gaze-sensitive system.

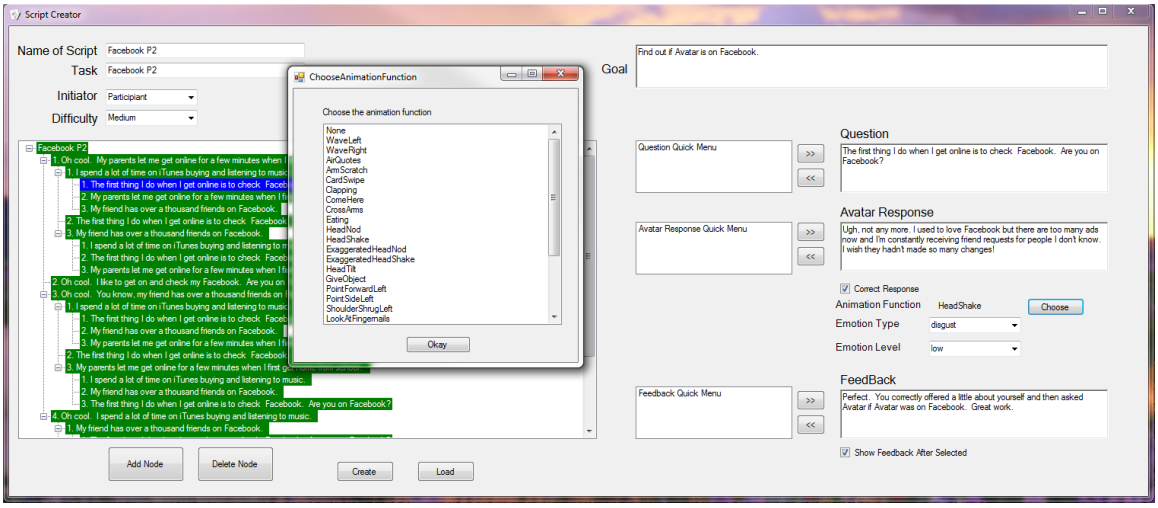


Fig. 45. The script creator program that was developed to generate the mission files.

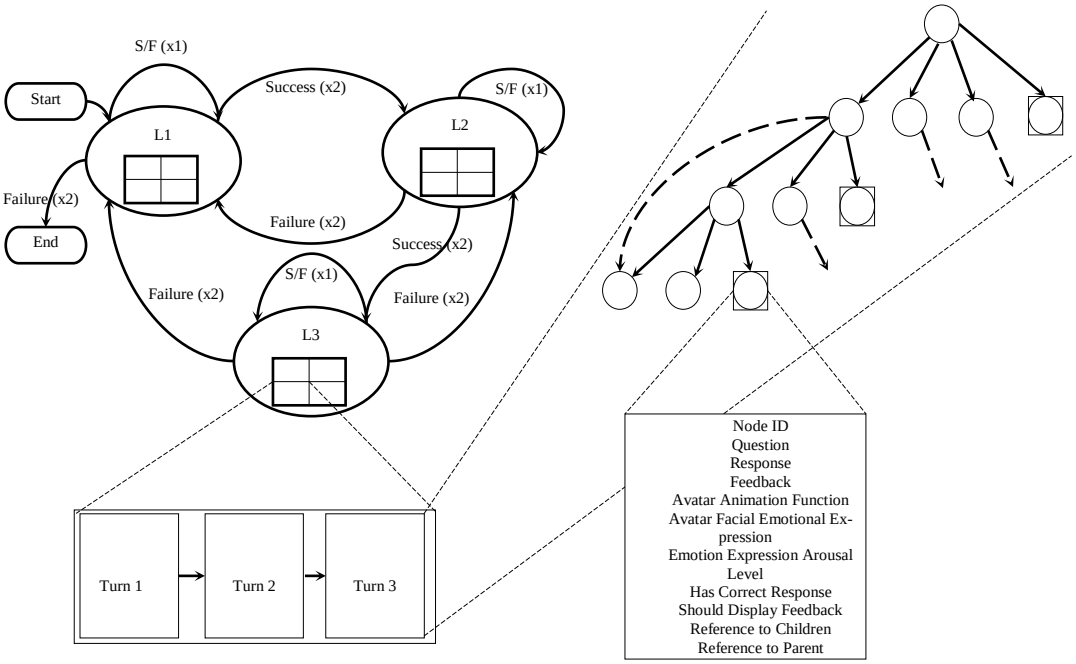


Fig. 46. Finite state diagram showing a simple example level switching logic.

Physiological Monitor.

Affective state recognition using peripheral physiological signals was formally introduced in (Picard, 2009; Picard, Vyzas, & Healey, 2001). Estimating the psychological affective states of subjects is important for technology-assisted therapy and it enables implicit and meaningful human-machine interaction. The physiological affective module for this study collected 8 channels of physiological signals for later offline analysis of affect. We collected labelled data for preliminary feature level analysis and future supervised online affect recognition module for interaction adaptation based on affective state of the subject. We utilized wireless sensors form Biopac Inc. (www.biopac.com) called BioNomadix. A subject wearing the body physiological signals is shown in Fig. 47.

The physiological monitoring application collects 8 channels of physiological data and was developed using the Biopac software development kit (SDK) and BioNomadix wireless physiological acquisition modules with a sampling rate of 1000 Hz. The physiological signals monitored are: electrocardiogram (ECG), pulse plethesymogram (PPG), skin temperature (SKT), galvanic skin response (GSR), 3 electromyogram (EMG), and respiration (RSP). Due to social communication impairments in adolescents with autism, they are not usually expressive of their internal affective states and these states often are not visible externally (C. Liu, Conn, K., Sarkar, N., Stone, W., 2008a; Picard, 2009). Physiological signals are, however, not affected by these impairments and can be useful in understanding the internal psychological states (Groden et al., 2005; Picard, 2009). Among the signals we monitored, GSR, PPG, and ECG are directly related to the sympathetic response of the autonomic nervous system (ANS) (Cacioppo et al., 2007). When there is increased sympathetic activity due to external factors and pressures, the heart rate, the blood pressure, and sweating are all elevated (Picard, 2009). Various features extracted out of these signals (see Appendix B) are used for supervised training of a machine learning algorithm for later affective state classification in the offline analysis stage.

Fig. 47. The peripheral physiological sensors

EEG Monitoring Module.

We also introduced an electroencephalography (EEG) monitoring module as part of the physiological monitoring. Fourteen channels of EEG signals were collected from the scalp of the subjects using the Emotive EPOC neuroheadset (www.emotiv.com) with a sampling rate of 128Hz. The channel locations were AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4, according to the international 10-20 system of electrode placement (Jasper, 1958). The common mode sense and driven right leg references were placed at locations P3 and P4 respectively. Similar to physiological monitoring, EEG monitoring module received trial and session markers from the task presentation engine via socket communication. For the purpose of this study EEG analysis is limited to offline feature level comparisons across groups and conditions. Preprocessing on continuous data includes slew rate limiting with a rising slew rate of 0.5 and a falling slew rate of -0.5, and band-pass filtering (0.2~43Hz). Then data are chopped into 1s epochs with 50% overlap. For each epoch, it is accepted if all the sensors maintain good contact with subject's scalp, there is no data point calculated by rising or falling rate, and less than 33% of the channels overcome the voltage threshold of 150 microvolts. EOG and EMG artifact correction algorithm are then applied on the remaining epochs. For now, the feature used is alpha asymmetry between F3 and F4. Six EEG features that related to affective stimuli processing were extracted from pre-processed EEG signals, including theta band (4-8Hz) power at right parietal area (P8), averaged theta2 band (6-8Hz) power at left anterior areas (AF3, F7, F3, and FC5), averaged alpha2 band (10-12Hz) power at anterior areas (AF3, F7, F3, FC5, FC6, F4, F8, and AF4), alpha2 band (10-12Hz) power at right parietal area, averaged beta1 band (12-18Hz) power at anterior areas, and gamma band (30-45Hz) power at right parietal area. Increase in power of theta ranges and high frequency activity, beta and gamma, were found to associate with processing of the affective stimuli, while the alpha2 bands were related to cognitive involvement during emotional stimuli processing (Aftanas, Reva, Varlamov, Pavlov, & Makhnev, 2004; Keil et al., 2001). Based on the approach-withdrawal model, relatively greater left frontal activity corresponds with experience of approach-related emotions whereas relatively greater right frontal activity corresponds with experience of withdrawal-related emotions (Demaree, Everhart, Youngstrom, & Harrison, 2005; Gotlib, 1998). Signals in these two channels are accepted if an epoch passes all four tests (REILLY & NOLAN, 2010): channel deviation test, variance test, amplitude range test, median gradient test. Then the mean value of F3 and F4 are removed respectively. After that, the alpha powers (8-13Hz) of F3 and F4 are calculated. Asymmetry measure is the difference between F3 and F4, which is log(F4) – log(F3).

Eye Gaze Monitoring and Online Adaptation.

The eye gaze tracker application was developed using Tobii SDK. The remote desktop eye tracker, Tobii X120, is used at 120 Hz frame rate that allows a free head movement of 30 x 22 x 30 cm (width x height x depth) at 70 cm distance. We run two applications: one to monitor the data visually as the experiment progresses and one to record, pre-process and pass the eye tracking data to engagement detection module. The main eye tracker application computed eye physiological indices (PI) such as pupil diameter (PD) and blink rate (BR) and behavioral indices (BI) (Lahiri, Warren, et al., 2012) such as fixation duration (FD) from raw gaze data. The FD is correlated with attention on a specific region of visual stimuli whereas the eye physiological indices PD and BR are indicative of sensitivity to emotion recognition and engagement (C.J. Anderson et al., 2006; Hsiao & Cottrell, 2008; Lahiri, Warren, et al., 2012; Libby Jr, Lacey, & Lacey, 1973). For each data point, gaze coordinates (X, Y), PD, BR, and FD were computed and logged together with the whole raw data, trial markers and timestamps in addition to being used as features for the rule-based engagement detection mechanism. The fixation duration computation was based on the velocity threshold identification (I-VT) algorithm (Salvucci & Goldberg, 2000). We chose the I-VT algorithm for its robustness and simplicity. The algorithm sets a velocity threshold to classify gaze points into saccade and fixation points. Generally, fixation points are characterized by low velocities (e.g.: < 100 deg/sec) (Salvucci & Goldberg, 2000). We used 35 pixels per sample (~ 60 deg/sec) as our velocity threshold. Spurious fixations processing was not considered in this online interaction phase. Offline post-processing rejected inadmissible fixations. The blink rate was computed using condition code returned from the eye tracker whereas the pupil diameter was averaged for both eyes when both eyes data were available and only one eye data when the other one was not in the tracking range.

The eye tracker application ran two separate network clients. One UDP client that accepted the raw tracking data from the eye tracker and a TCP (transmission control protocol) client that received trial beginning, trial end, and session end marker data from the task presentation engine. Fig. 48 shows details of the eye tracking application and its components in the online recording, pre-processing and logging stage.

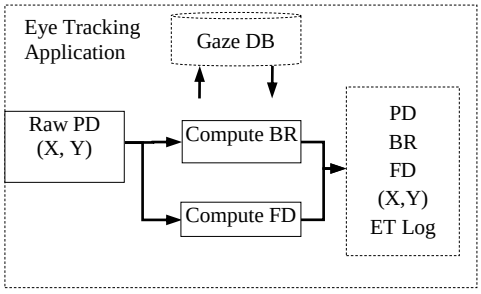


Fig. 48. Eye tracking application and its components

This application logs the raw data from the eye tracker and preprocessed data in the form of behavioral and physiological eye gaze indices such as fixation duration, blink rates, and pupil diameter are logged separately. Moreover, a dynamic eye gaze monitoring of where the participant is actually looking in the virtual environment is logged separately.

The online adaptation was after each task (which we call missions), the emotional expressions were presented with occlusions. If the subject looks at context relevant areas such as the eyes and mouth more than a certain threshold, the face reveals itself. See Section 5 for the details of the occlusion paradigm.

W. Data Analysis

The MASI-VR system allows collecting a host of important data for both online and offline analysis depending on the need of a particular intervention. In this section we present specific data analyses that were performed for the presented usability study (Section 3) where we used the MASI-VR for an occlusion-based facial emotion recognition paradigm.

Eye tracking data analysis.

The gaze data analysis was performed to determine behavioral viewing patterns of children with gaze group as compared to that of the control. The behavioral indices such as where they were looking in terms of screen coordinates were clustered into ROIs defined around the key facial bones. The clustering results were then averaged over trials for each subject and the aggregate results were used. The defined ROIs represented the following regions: forehead, eyes (left and right), nose, and mouth. The face region was modelled by a combination of an ellipsoid and a rectangular forehead region (Fig. 49). Facial regions outside of the 5 defined regions of interest were categorized as "other face regions" while all the background environment regions outside of the face regions were defined as "non-face regions". This gave a total of 7 regions into which all the gaze data points were clustered.

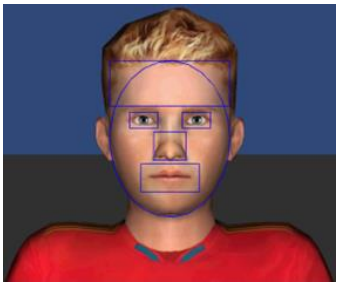


Fig. 49. The five facial ROIs defined on the face region.

The other behavioral index considered for analysis was the fixation duration. The raw fixation duration was computed for each gaze point during the online interaction. The raw data were first filtered to remove excessively small and large fixation durations. Typical fixation duration and saccades last between 200 and 600 ms and 30 and 120 ms, respectively [57]. The filtered fixation duration data were used to compute the average fixation duration (FDave). Another important behavioral eye index associated with fixation duration, called the total sum of fixation counts (SFC), was also computed from the filtered fixation duration data.

The eye physiological indices, i.e., the blink rate (BR) and the pupil diameter (PD) were also post processed. Missing PD data due to blinks and presence of noise was filtered from the collected PD data. First a threshold was established to segment missing noisy data from the actual PD. Then the missing data points were constructed by linearly interpolating the neighbouring data points. The BR data were also filtered based on typical blink ranges. Typical human blinks range between 100 and 200 ms [58]. The PD data were used to reject blinks at missing data points.

Physiological pattern analysis.

The collected physiological signals were analysed for feature level statistical comparisons to see any pattern differences between the pre and post conditions of the same group and between the two groups (See details of the experimental protocol in Section 3).

First, the signals were filtered to reject outliers and artefacts and smooth the data. Then, individual baseline mean was subtracted from the data to remove effects of individual variations. The signals were then standardized to be zero mean and unity standard deviation for further feature extraction. For ECG and PPG, the peaks were detected after baseline wander removal following the artefact removal.

Feature Extraction and feature level analysis.

From the 8 channels of physiological signals collected, 4 were selected for feature level analysis ECG, RSP, SKT and GSR, 16 features were extracted from these 4 channels. We selected 5 important features, i.e., heart rate (HR), skin temperature (SKT), respiration rate (RSPR), skin conductance rate (SCR) and skin conductance level (SCL) which were chosen because of their correlation with engagement and emotion recognition process as noted in psychophysiology literature [21, 33, 49] for the feature level analysis as shown in the results section. For example, cardiovascular activities such as inter-bit interval (IBI) represents the rate at which the cardiovascular activity changes and can be used to distinguish arousal levels of an emotion. Electrodermal activity as measured via GSR is indicative of response to external stimuli that might make the subject tense or anxious. The pulse transit time (PTT) is a measure of the time the blood takes to travel from the heart to the finger tips. This specific feature was computed using the peaks of both ECG and PPG signals.

EEG Data Analysis.

The logged EEG data will also be processed to see some pattern differences. The following analysis will be performed offline on the EEG data. Channel deviation test: Calculate the mean value of each channel and transform the result to its zscore form. If channel F3 or F4 have a zscore greater or equal to 3, reject this epoch. Variance test: Calculate the variance of each channel and transform the result to its zscore form. If channel F3 or F4 have a zscore greater or equal to 3, reject this epoch. Amplitude range test: Calculate the amplitude range (max(eeg)-min(eeg)) of each channel and transform the result to its zscore form. If channel F3 or F4 have a zscore greater or equal to 3, reject this epoch. Median gradient test: Calculate the median value of the changing rate of each channel and transform the result to its zscore form. If channel F3 or F4 have a zscore greater or equal to 3, reject this epoch.

X. Methods and procedure

A usability study was conducted to validate the system and to study the behavioral and physiological pattern difference of children with ASD that participated with a gaze-sensitive version of the system and control group that participated without the online gaze adaptation and occlusion paradigm. Subjects were randomly assigned to one of the two groups to control for bias.

The MASI-VR Protocol.

Each subject in the experimental protocol undergone through a typical 3 visits protocol. For some of the subjects that were unable to perform the pre and the post-tests together with the social task training, the sessions were spaced out to 5 visits. Table 16 describes what the subject was doing in each visit. First of all, the subject performs a pre-test of isolated emotion recognition task in VR without the conversational dialog context and a standardised "A Developmental NEuroPSYchological (NEPSY) Assessment" with emotion recognition components followed by the first session of the social task in visit 1.

Table 16. Summary of Visits

V1	Informed consent
	FE Pre-test + NEPSY Pre
	1st exposure to conversation paradigms - emotions at high valence
V2	2nd exposure to conversation paradigms - emotions at medium valence
V3	3rd exposure to conversation paradigms - emotions at low valence
	FE Post-test + NEPSY Post

FE: Facial Expression

Then, in visit 2, the subject performed the second session of the social task. Finally, in visit 3, the subject goes through the third session of the social task, the post-test isolated emotion recognition in VR which is the same as the pre-test, and the post-test of NEPSY.

The difference between the three sessions of the social task is the emotion intensity level of low, medium and high, respectively. The pre-post-test contains isolated emotion recognition of 28 trials whereas the social tasks have 12 conversational missions (goals). At the end of each mission, 2 emotional faces were presented with the occlusion paradigm for the gaze group and without for the control group for a total of 24 emotion recognition trials in the presence of social context.

Experimental setup.

The VR environment was run on Unity. Eye tracking and peripheral physiological monitoring were performed in parallel on separate applications that communicated with the Unity-based VR engine via a network interface as described in Section 3. The VR task was presented using a 24'' flat LCD panel monitor. The experiment was performed in a laboratory with two rooms separated by one-way glass windows for parent observation. The parents sat in the outside room. In the inner room, the subject sat in front of the task computer. A therapist was present in the inner room to monitor the process. The task computer monitor was also routed to the outer room for parent observation. The session was video recorded for the whole duration of the participation (Fig. 50).

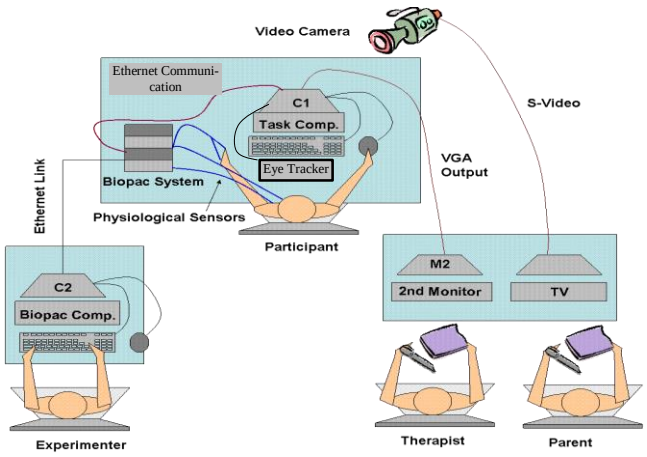


Fig. 50. The experimental setup.

Subjects.

A total of 6 high functioning subjects with ASD with no mental retardation record (Male: n=6,) of ages 13 – 17 (M=15.77, SD=1.87) and an age matched 6 (Male: n=6) controls of ages 13 – 17 y (M=15.20, SD=1.68) were recruited and participated in the usability study. All ASD subjects were recruited through existing clinical research programs and had established clinical diagnosis of ASD. All subjects with ASD fall well above the clinical threshold. The gold standard in clinical ASD diagnosis, the Autism Diagnostic Observation Schedule-Generic (ADOS-G) the new algorithm score and the severity score (ADOS-SS) were used to recruit the ASD subjects. IQ of the ASD subjects was obtained from existing clinical research database.

The control group were trained using the system without any online gaze feedback and without the occlusion paradigm. In addition to ADOS-G, we asked parents of the subjects to fill out the social responsiveness scale (SRS)[59] and the social communication questionnaire (SCQ) [60]. Parents of both groups have completed these forms. In addition, WASI [61] was used to measure IQ of the subjects. The IQ measures were used to potentially screen for intellectual competency to complete the tasks. Moreover, we used a measure of facial recognition called NEPSY together with an isolated facial expression recognition was administered to the subjects as pre and post measures before and after their repeated training with the social task. Again all the CTR subjects were well above the clinical cut-offs for the SRS and the SCQ as well as ADOS-G.

The MASI-VR Social Task.

Initially the subject is seated in front of the task computer. The eye tracker is calibrated and the peripheral interfaces are all connected via network. Once this is finished, the subject information is entered into the system. Then the subject is ready to start the task. At the start of the social task, the subject will go through a sequence of instructions on how to perform the task. Once the instruction is finished, the subject gets to move around the virtual cafeteria and approach a character to interact with. Each conversational character has a personal space which was implemented using Unity's collider triggers. Once the subject enters the invisible personal space of the character, the character invites the subject to interact by indicating it is ready to talk to the subject. If the subject chooses not to interact with that specific character, the message gets reset once the subject moves out of the personal space of the character. Once the subject chooses to interact with a specific character, then the subject is presented with the available 4 missions in low level. The subject can choose any of the 4 missions. Once the subject chooses the mission, the subject is asked to choose the conversational topic in order to measure if the subject understood the mission. Then the subject will enter the conversation dialog and the conversational dialog management engine goes through series of trials of conversation depending on the subject's success and failure as shown in Fig. 51.

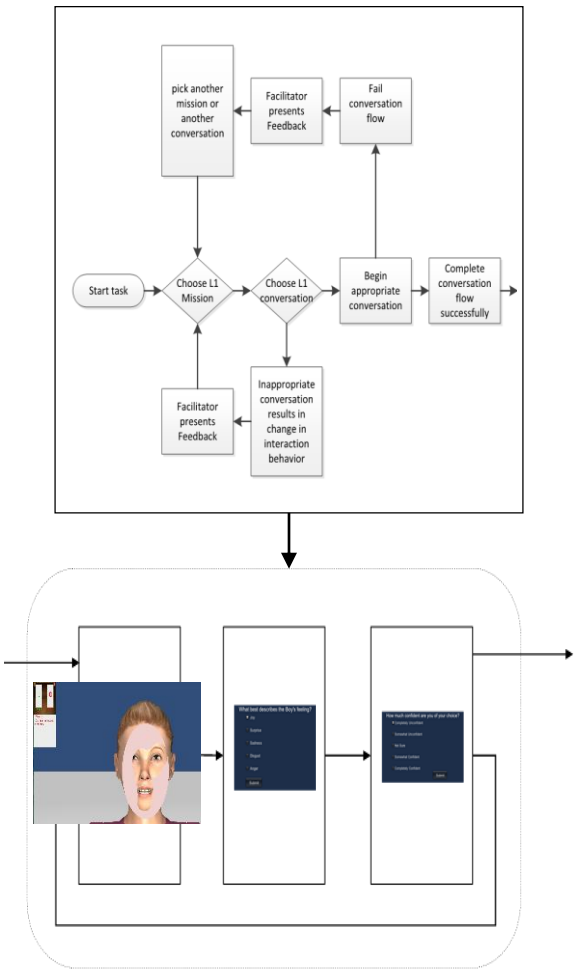


Fig. 51. The spoken conversation and the emotion recognition

The VR-based facial emotional recognition in the presence of conversational dialog system presented a total of 12 conversational dialog missions. At the end of each dialog mission two facial expressions were presented with the face occluded with oval occlusion as shown in Fig. 9. As the subject scans the face, the occlusion erodes by the gaze of the subject to give an online adaptive gaze feedback. If the subject pays attention to the context relevant areas of eyes and mouth beyond a threshold, the face reveals with the emotion and the subject gets to choose what the emotion was. If the subject was not successful in revealing the face in 15s, the face reveals itself and the process continues normally.

The emotions in the pre-post-tests consisted of a total of 28 trials corresponding to the 7 emotional expressions with each expression having 4 levels. Each trial was 30-45 s long. For the first 25-40 s, the character narrated a lip-synced context story that was linked to the emotional expression that followed for the next 5 s. The character exhibited a neutral emotional face during story telling. Subjects were expected to rate the emotions based on the last 5 seconds of interaction. The story was used to give context to the displayed emotions. The context of the stories ranged from incidents at school to interactions with families and friends that were suitable for the targeted age group. However, since this VR task was used as purely a pre and post measure, the story was not interactive and the audios of the stories were recorded with monotonous tone so the subject gets to decide the emotion solely based on the isolated expression and not get influenced by the story.

A typical laboratory visit was approximately one hour long. During the first 15 minutes, a trained therapist read approved ascent and consent documents to the subject and the parent, and explained the procedures. Once the subject finished signing the ascent document, he/she began the task. While the parent completed the SRS and SCQ forms, the subject wore the wearable physiological sensors with the help of a researcher. Before the task began the eye tracker was calibrated. The calibration was a fast 9 points calibration that took about 10-15 s. At the start of the task, a welcome screen greeted the subject and described what was about to happen and how the subject was to interact with the system. Immediately after the welcome screen, the trials started. At the end of each trial, questionnaires popped up asking the subject what emotion he/she thought the character displayed and how confident he/she was in his/her choice. The total participation time was about 20-25 minutes. The emotional expressions presentations were randomized for each subject across trials to avoid ordering effects. To avoid other compounding factors arising from the context story, the story was recorded with monotonous tone and there was no facial expression displayed by the character during that period.

Usability Study.

Participants will be assigned to one of the two groups. The first group will be evaluated solely based on performance without any occluded face and the second group will be evaluated with the proper gaze-contingent system with facial occlusions. Group 1 will be performance based system (PBS) whereas group 2 is presented with the gaze-based system (GBS). In PBS, the participant is presented with animated face of the avatar without any occlusion displaying looped presentation of the emotion animation for 15 seconds. Then, the participant will be presented with the emotion identification questions. Feedback is provided solely on performance. In the GBS, animated face of the avatar is presented displaying looped animation of the emotion with an opaque screen occluding the facial expression. As the participant gazes on the crucial ROIs of the face, the occlusion starts to reveal the face behind. Participant is required to spend time looking at certain regions of the face. Participant gaze on these regions will reveal core components of the face. If participant successfully reveals core components of the face, participant receives questions and continues. (System guards against premature selection of the emotion.) If participant is unsuccessful in revealing core component of the face within 15 seconds, the face is displayed and the participant receives the question and continues. Feedback is presented based on performance and the next mission continues.

The overarching goal of this system is to teach a participant to properly process emotional faces in the presence of social context while logging useful physiological, EEG and eye tracking data for an offline processing and pattern analysis. The goal of the system is to provide information on how to teach a participant to process a face. Aims of the system is that it will train the user to look more frequently at the faces that we have, with the idea being that there could be performance differences between the two groups. If there are not differences in performance, we may be able to show differences in how they are processing the information. Are participants?

- Performing more quickly: In both systems we will test the latency of response. The emotion recognition question will be available for the user to make his/her selection when:
  PBS: automatically. Revealed face appears and user can make selection as soon as choice is determined.
  GBS: once gaze targets key components of the face and full face is revealed. At that point, the emotion recognition question appears and the user can make choice when ready. System will guard against making emotion choice before face is fully revealed.
- Spending more time on the core components of the face.
- Feeling more confident/comfortable identifying emotions.

The idea is not that everything would be revealed by eye-gaze toward the face. Participants would have to identify key regions of the face. Participants will have to understand what is happening at the mouth and the eyes. And we have the ability in this environment where the occluded face can be dynamically displaying what's happening, not just a flat picture, but actively displaying the emotion. Therefore, when participants look at those regions, those active regions of the face become visible. Once a participant looks at all facial ROIs the full face (actively displaying the emotion) becomes visible. After a certain amount of seconds, the face would appear whether or not all regions were revealed, and the user would be able to answer the question. This way, users would learn that they will be rewarded by looking quickly at key regions of the face in order to fully reveal the face. If one looks at the key regions quickly enough s/he gets to move on. If one is struggling, then it takes longer but eventually the system reveals it. Hence the system would have an internal reinforcement mechanism for quick processing.

Y. Results and Discussion

Performance.

Performance of the subjects in identifying the emotions in the five sessions (three social task training sessions, S1, S2, and S3, and pre and post isolated facial emotional recognition sessions) was measured using three metrics. Their total score was gauged as a percentage of the total trials in each session. The subjects also indicated how confident they were in their choices immediately after they made their choices. The time they took to indicate their choices was also considered as a performance metric and calculated as latency of response in seconds.

Fig. 52 shows that both the gaze and the control group were able to improve their performance from pre to post while the gaze group improves by 3% more than the control group's improvement. However, the performance score differences were not statistically significant. The confidence metrics indicated that gaze

group subjects were relatively as confident as the ones in the control group while taking slightly more time across all the sessions.

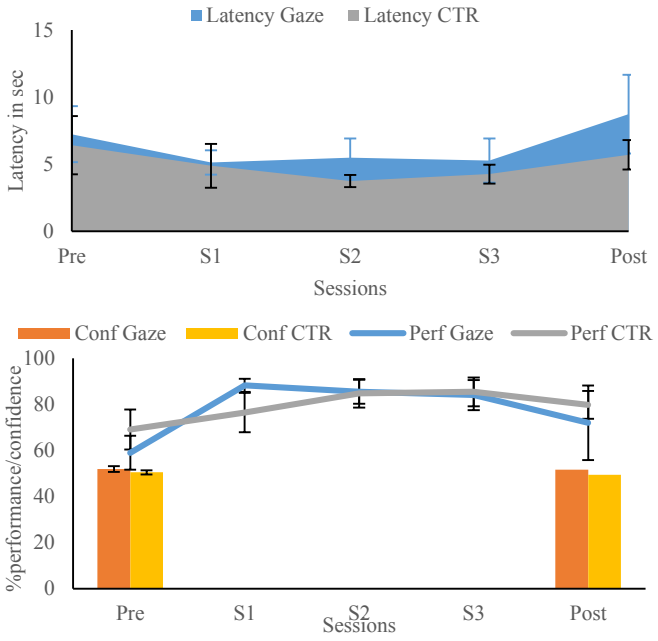


Fig. 52. Generalized performance metrics

To further control for the effect of choice bias, i.e. subjects choosing a specific emotion selection regardless of stimuli, we computed the bias for each emotion and corrected the raw performance by removing the bias. Fig. 53 shows the bias for each emotion across subjects for the pre and post sessions with the expected bias, which is 14.29% for 7 emotions, for reference. We can clearly see that subjects were more biased for some emotions such as fear over the others. The bottom figure shows the raw performance and the bias corrected performance overlaid on top as lines.

In summary, the results indicated that MASI-VR gaze-sensitive system was effective in making the gaze group close the performance difference by more than 3% from pre to post without significant difference in confidence and latency with the control group.

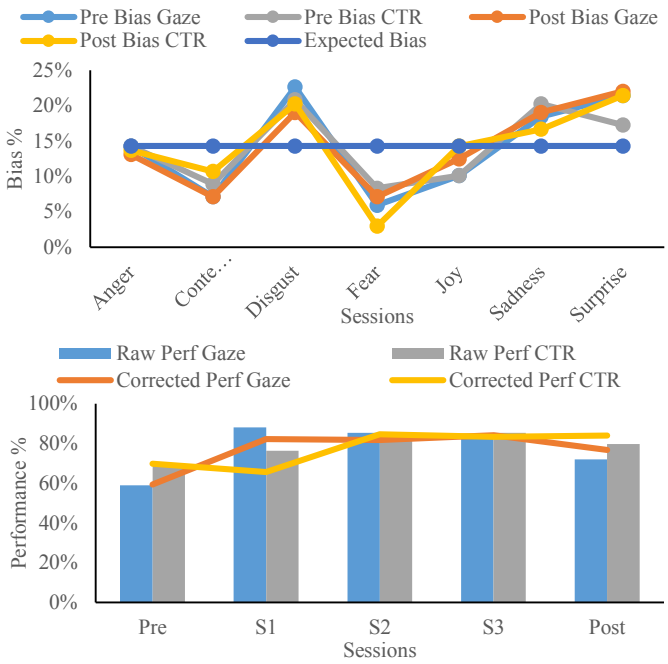


Fig. 53. (Top) bias per emotion, and (bottom) raw and bias corrected performance

Gaze towards ROIs.

This is the core gaze analysis to determine if subjects improved due to the social emotion recognition training by MASI-VR.

ROIs and gaze analysis.

Fig. 54 presents the gaze towards the various ROIs that were defined for the VR task as shown in Fig. 49. The plot shows time spent in that specific ROI as a percentage of total trial time.

Data was averaged across trials for each subject and across subjects in each group. In the context relevant areas such as the eyes, both groups increased paying attention from pre to post and decreased gaze towards the mouth ROI from pre to post with the gaze group decrease of 10%, $p<0.05$, while the control group decreased gaze to the mouth by 5%. In a similar manner the gaze towards other parts of the face also increased as the forehead ROI. Since there are many facial bones that were animated on the forehead that move during the emotional expression trials, the increase in the forehead was not surprising. The gaze group actually increased more than the control group towards the forehead ROI. There were no statistically significant differences between the two groups. Most of the significant changes were within group changes of the gaze group towards the mouth and the control group towards the eyes.

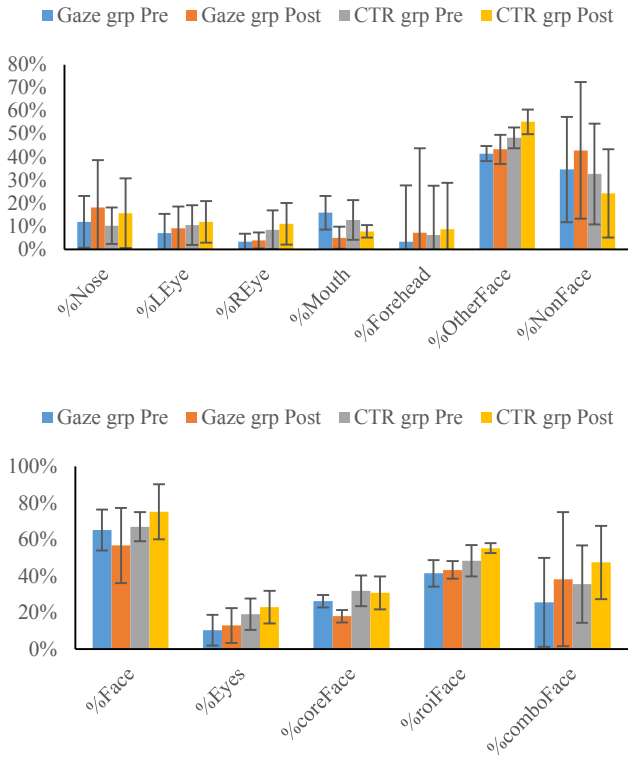


Fig. 54. (top) Gaze towards ROIs defined in Fig. 49 and gaze towards combined ROIs.

To further enhance the differences, we have combined several ROIs. The right and left eyes were collapsed into one measure towards the eyes, all the face region, and the core face ROIs including eyes and mouth. Further we combined all the ROIs on the face into a single ROI, roiFace and the ROIs that might interfere with the eyes such as the nose and the forehead combined together with the eyes as comboFace. Both groups increased in the eyes, roiFace and the comboFace ROIs. The gaze group increase in the comboFace was specifically larger than the control group. The total face area ROIs were combined and compared with non-face ROIs as well. The gaze group decreased on the face while increasing on the non-face ROIs while the control group displayed the opposite behaviour.

Gaze visualizations

We have generated heat map and masked map visualizations to qualitatively compare the differences from pre to post in both groups collapsed across all trials and all subjects. The heat map was smoothed using a Gaussian filter after the fixation duration was accumulated and computed for each region to remove unnecessary discontinuities and minimize the effect of noise on the visualization.

Fig. 55 shows representative gaze pattern towards a character across trials and across all subjects in the gaze group. The gaze group increased their gaze towards the eyes with a more symmetric gaze in the post test (bottom) than the pre-test. The change in symmetry in the gaze processing pattern is consistent over all 7 characters the subjects identified the emotions with.

The control group also had a more symmetric and slightly increased gaze towards the eyes in the pre-test than the gaze group. However, the slightly decreased and the gaze slightly faded towards the eyes (Fig. 56).

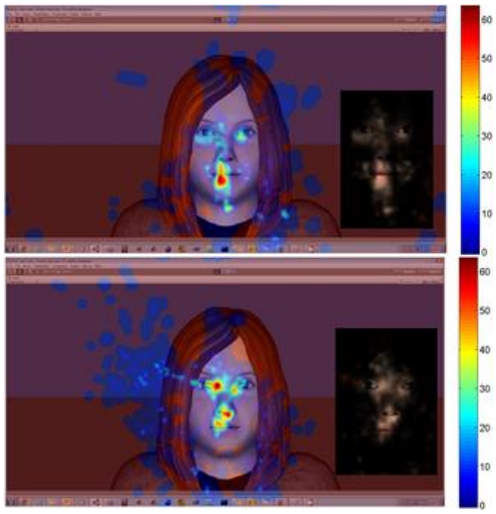


Fig. 55. Gaze Group Pre (top) and Post (bottom).

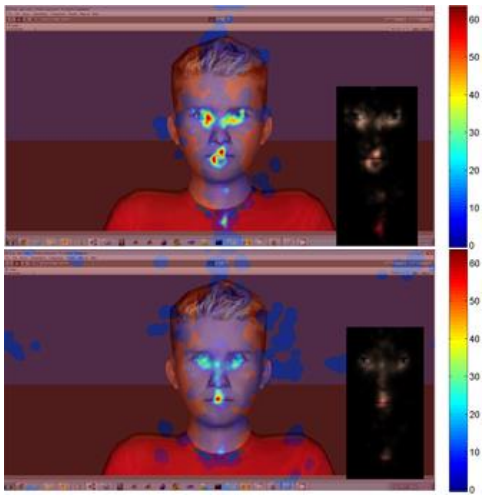


Fig. 56. Control (CTR) Group Pre (top) and Post (bottom).

Analysis of Eye Features.

In addition to the gaze towards ROIs and qualitative visualizations, we analysed 5 features from the eye tracking data. There were: the pupil diameter (PD), the average fixation duration (FDave), sum of fixation counts (SFC), blink rate (BR), and saccade path length (SPL). These are measures of behavioral viewing patterns in this study. Section 2 describes how these indices were computed. These behavioral indices are indicative of engagement to particular stimuli and are correlated with social functioning for individuals with autism [62, 63]. Generally, children in the gaze group had slightly lower FD and slightly higher SFC than the control group in the pre and post sessions. However, the variations were not as such statistically significant and hence we computed correlation between these features in order to come up with a combined gaze engagement index. Fig. 57 shows the correlation matrix and it is evident that features 3, 4, and 5 are highly correlated. So, we choose these highly correlated features and the fixation duration as it is directly correlated with engagement. Then we combined the three together with inverse and add it to the normalized fixation duration. From the figure it is evident that the three features are inversely correlated with fixation duration.

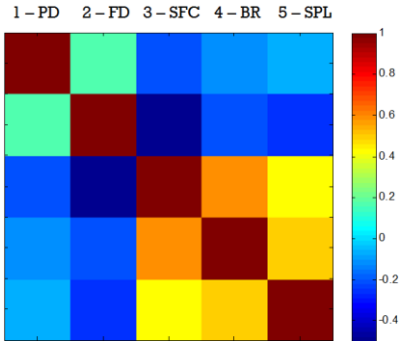


Fig. 57. Correlations of physiological eye indices.

The physiological patterns of the eyes of the subjects were represented by the average pupil diameter (PDave) and the average blink rates (BRave). PD is indicative of how engaged a subject is and literature suggests that there are variations of these indices among individuals with ASD [24] given the same stimuli. Individuals with autism were shown to have abnormal eye blink conditioning compared [64]. Generally, children in the gaze group exhibited lower pupil diameter and blink rates in all the sessions.

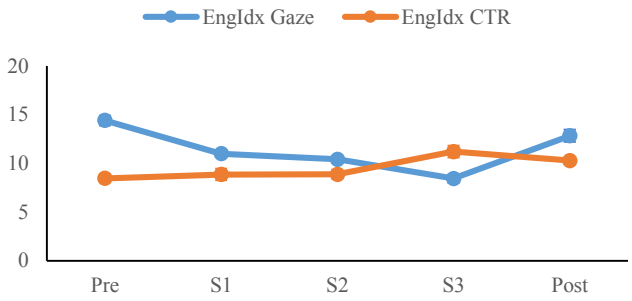


Fig. 58. Comparisons of physiological eye indices.

Fig. 58 shows that with the combined engagement index the gaze group engagement was slightly lower in the post-test compared to the pre-test and the control group was the opposite. However, the system was able

to maintain engagement in both groups across all the sessions and ending up with almost similar pattern of engagement in the post-test with the gaze group with higher engagement than the control group.

Physiological Features Analysis.

The physiological data of the children in both groups were processed and five features were extracted as described in Section 2 in the data analysis section. These were the heart rate (HR), the skin temperature (SKT), the respiration rate (RSPR), the galvanic skin conductance rate (SCR), and the skin conductance level (SCL).

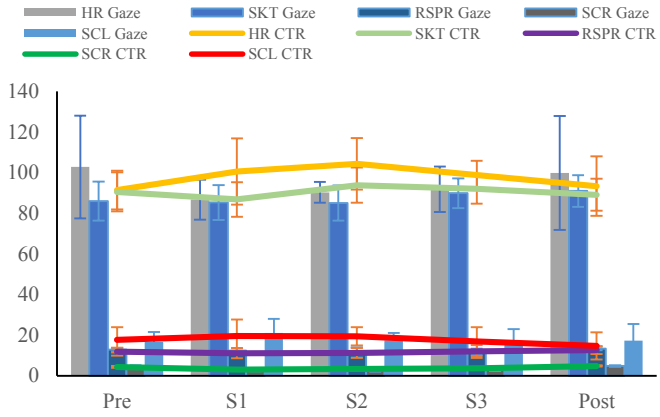


Fig. 59. Comparisons of physiological eye indices.

Fig. 59 shows that the general trend in both groups (the bars are for the gaze group whereas the line plots are for the control group for clarity) decreased activity from pre-to-post with the control group slightly higher than the gaze group. However, none of these changes were statistically significant. Lower heart rate, lower skin conductance and respiration rate are all indicative of lower emotion reflection activity. These seem to be consistent with the gaze features as well.

EEG Features Analysis.

Table 17. EEG features for the VR session

			RPT	LAT2	AA2	RPA2	AB1	RPG
Pre	Gaze	M	0.84	0.83	0.88	0.91	0.94	0.96
		SD	0.13	0.17	0.15	0.10	0.23	0.23
	CTR	M	0.65	0.70	0.84	0.81	0.80	1.16
		SD	0.12	0.23	0.35	0.24	0.15	0.87
Post	Gaze	M	0.61	0.76	0.76	0.67	0.86	0.85
		SD	0.24	0.23	0.15	0.30	0.09	0.23
	CTR	M	0.90	1.18	1.31	1.20	0.99	1.01
		SD	0.24	0.66	0.97	0.46	0.17	0.24

RPT: Right parietal theta, LAT2: Left anterior theta2, AA2: Anterior alpha2, RPA2: Right parietal alpha2, AB1: Anterior beta1, RPG: Right Parietal Gamma

EEG features extracted from each trial were divided by corresponding baseline measures to remove individual variations. Table 17 listed the mean and standard deviation of each individual feature for gaze and control group during pre and post-tests. For gaze group, the values of all the features decreased. For control group, all but right parietal gamma increased. Right parietal theta was significantly different (t-test, $P < 0.05$) between two groups for pre, while for post right parietal alpha2 was significantly different (t-test, $P < 0.05$). Within the control group, right parietal theta increased significantly from pre to post (t-test, $P < 0.05$). Since the correlation coefficients among six features were all positive, we normalized each feature to range [0, 1] and combined them together as one single feature. Fig. 60 shows the change of the combined feature for the five sessions. From pre to post, the combined feature decreased for gaze group and increased for control group. Even though the gaze group had higher feature value for pre, it was not statistically significant. After training with social task, the gaze group had lower feature value and it was statistically significant (t-test, $P < 0.05$). As these features relate to the subjects ability to process emotion and cognitive task load, the gaze group was found to be less engaged in the emotion recognition mental process. This result also goes with the eye tracking features in which the gaze group decreased engagement from pre to post-test although the change was not as significant.

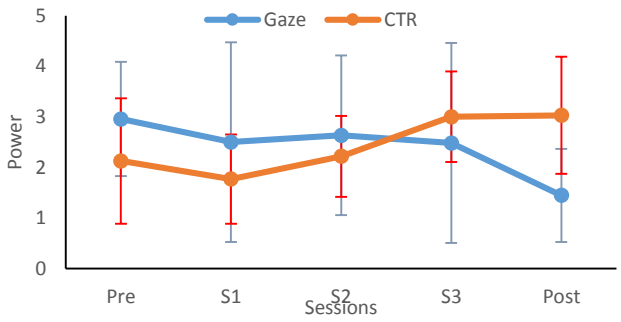


Fig. 60. Combined EEG feature

Z. Overall Conclusion

The main contribution of this work is to present the development of a realistic multimodal VR-based social interaction platform that can be used for ASD intervention. The uniqueness of this platform relies on its ability to gather objective eye gaze and physiology data while a participant is engaged in a closed-loop VR-based adaptive social interaction.

In this chapter, we have presented the design, development and usability study of a multimodal adaptive social interaction VR environment (MASI-VR). The system was able to collect eye tracking, peripheral psychophysiological and EEG data while the subjects were involved in the emotion recognition training and pre and post isolated emotion recognition tasks over 3 visits over a period of time. Such capabilities are expected to be useful in understanding the underlying deficits individuals with ASD and, in turn, will hopefully help developing new intervention paradigms to improve such impairments. The 3 social task training sessions in between the pre and post-tests were specifically designed to allow the training of proper gaze processing

pattern with an online adaptive and individualized gaze feedback mechanism. Moreover, it also incorporated a spoken conversational dialog management for more natural social emotion recognition. These two are the pillars of the system in an effort to teach children with ASD the ability to recognize faces in a social context. A usability study involving 12 children with ASD was performed to evaluate the efficacy of the system as well as to study behavioural and physiological pattern differences. Half of them were randomly assigned to the gaze-sensitive part of the system while the remaining half were used as controls without the occlusion paradigm and hence without the online gaze feedback mechanism to evaluate the effectiveness of the dynamic system as a whole.

The system successfully performed the adaptive social task as well as the pre and post isolated facial expression tasks and collected the synchronized eye gaze, EEG and physiological data.
The results of the usability study indicated that the gaze-sensitive system enabled the gaze group to close the performance gap that existed in the pre-test with the control group by more than 3% points while maintaining relatively similar engagement as depicted by the various modalities with that of the control group. There were several limitations of the system in its current form. Although the system was equipped with measuring EEG as well as physiology signals, only the gaze was used for online adaptation. The gaze group subjects had difficulty in understanding the occlusion paradigm at first as explicit guidance was not given so as to not bias the outcome. However, over the course of the 3 visit social task training, they picked up the importance of the occlusion and what they are supposed to do to reveal the faces. This initial confusion might have contributed to the eventual non-pronounced engagement differences.

Despite these limitations, the system proved that the controllability, ease of interaction without information overload and the game nature of the interaction were useful in training core deficit areas of children with ASD for eventual better social functioning. These preliminary findings will be used to build a more robust adaptive VR-based social interactive environment that enables online adaptation not only by gaze but also using all the multimodal inputs using some form of decision level fusion for children with ASD to improve their emotion recognition abilities and eventual social functioning.

CHAPTER VII

POTENTIAL CONTRIBUTION

This chapter briefly describes the potential contributions of the dissertation in terms of technological advancement, advancement in the technology-assisted intervention and overall long term contributions to the society. The main contribution of this dissertation research is the design, development, and user evaluation of intelligent assistive technology for ASD and SZ intervention. The developed systems have capabilities to adapt to the user in an individualized way while maintaining objectivity and modifiability in a scale that is not possible in a traditional human centric intervention and technology-assisted systems that do not integrate implicit cues.

AA. Technological Contribution

The main technological contribution of this dissertation is the development of a self-contained, adaptive robotic intervention architecture for young children with ASD and two specific virtual reality-based technology-assisted platforms for individuals with ASD and SZ. These systems not only respond to individuals' response performance, but also collect and partially respond to implicit and explicit cues integrated in the overall framework and platforms.

Although, there are preliminary works in both robot-mediated (K. Dautenhahn, 2003; K. Dautenhahn, Werry, I., 2004; Diehl et al., 2011; E. S. Kim, Paul, Shic, & Scassellati, 2012; Kozima, 2005; B. Robins, Dautenhahn, et al., 2004; Scassellati, Henny Admoni, & Matarić, 2012; Werry, 2001) and virtual-reality-based (Harris & Reid, 2005; S. Parsons & Cobb, 2011; T. D. Parsons, Rizzo, Rogers, & York, 2009; Riva, 2005; Standen, 2005; Wang & Reid, 2011) therapies for ASD (Billard et al., 2006; K. Dautenhahn et al., 2002; D. Feil-Seifer & Mataric, 2008; Josman et al., 2011; Kozima, 2005; Michaud, 2002; Millen, Cobb, & Patel, 2010; S. Parsons, Mitchell, P., 2002; B. Robins, Dickerson, et al., 2004; Strickland, Marcus, Mesibov, & Hogan, 1996) and schizophrenia(Dyck, Winbeck, Leiberg, Chen, & Mathiak, 2010; Freeman, 2008; Gutiérrez-Maldonado, Rus-Calafell, Márquez-Rejón, & Ribas-Sabaté, 2012; Park et al., 2011; Rus-Calafell, Gutiérrez-Maldonado, & Ribas-Sabaté, 2014; Ruse et al., 2014; Tsang & Man, 2013), they have fundamental technological limitations as discussed in previous chapters. In particular, they solely depend on task performance and are, in general, unable to incorporate other interaction modalities and implicit cues for individualized adaptation.

Most of the robot-mediated projects use the robot as therapeutic partner and mostly assess the focus of the participants on the robot itself and less on the adaptive capabilities of the robot (Billard et al., 2006; K. Dautenhahn et al., 2009; Kozima, 2005). Our work on Robot-mediated joint attention task proposed a new architecture called Adaptive Robot-mediated Intervention Architecture (ARIA) that focused primarily on the adaptive feedback capability of the system. In this architecture, the robot is embedded within an intelligent learning environment that could measure one's approximate gaze in real-time and allow individualized feedback. As a part of this system, we developed and tested a standalone wearable IR head tracker and integrated it with the robotic system for online feedback. We showed in the pilot study that adaptive robotic therapeutic system that responds to user's cues in real-time hold potential in individualizing the intervention. The performance of the ARIA system was comparable to that of a human therapist. It also attracted children's attention more than the therapist, which indicates its potential for skill training.

In addition to the main contribution, the design, implementation and user testing of ARIA resulted in other technological tools. The development of the innovative IR-light based wearable head tracker that was used in the pilot study was effective enough for the pilot study and is currently used as a basis for remote non-contact based gaze estimation system. Beside the head tracker, we have developed a distributed system based on light-weight client-server architecture for multiple application and multiplatform communication. Most of the robot modules and the camera processing modules were developed and run remotely using SOAP RPC to offload the robot processor for the basic robotic actions and communicating via the network interface. The

socket interface library written for this project was extended and used across various projects in our lab afterwards including the virtual reality projects.

Current virtual reality work on ASD and SZ therapy focus on the virtual reality as a tool to simulate interaction scenarios that is driven only by performance rather than adaptively interactive platform for individualized therapy. We have developed two virtual reality systems for isolated and contextual facial emotional expression identification tasks. The completed and proposed systems integrate monitoring and adjusting interaction in real-time (gaze only) based on implicit cues such as physiological, EEG and eye gaze in addition to performance within the system. These systems incorporated facial emotional presentation, a speech-based pre-scripted hierarchical dialog management system and gaze-sensitive occlusion paradigm that guides the user to look more on context relevant regions of interest of a face while processing facial emotional expressions in the presence or absence of social contextual cues.

In addition to the integration of multimodal implicit sensing and the development of a holistic VR systems endowed with social interaction and emotion presentation, the virtual reality systems resulted in the development of generic and multi-language light weight socket communication library, improved signal processing and feature extractors for the physiological signals and eye trackers, physiological and eye tracking baseline and task specific datasets for affect recognition, offline affect recognition and clustering, generic event management system and state-transition finite state automata system.

BB. Scientific Contribution

The existing intelligent technology assisted intervention literatures in ASD and SZ are both limited in technology and mostly represents proof of concept systems that rarely focus on core deficit areas of the respective disorders. The presented ARIA robot mediated architecture for joint attention task and the virtual reality facial expression presentation platforms enabled individualization by making the technology sensitive to individual cues. As a result, these systems allow micro and macro within system interaction adjustments and fine-tuning that is tailored to the individual user. Their autonomous capability of administering the therapy with minimal or no human intervention with objective and new kinds of measures paves the way for designing new intervention paradigms that may not be possible in the traditional human-centric intervention.

The other contribution of this dissertation in terms of advancing intervention science is by focus on core deficit areas of the disorders. Most technology assisted work in ASD and SZ focus on the technology per se and put less emphasis on the need to evaluate the systems in real world clinical studies with controlled groups focusing on core deficit areas of the disorders to see if the technological systems actually translate into meaningful improvements highly relevant to these disorder. We specifically put great emphasis on designing intelligent systems focused on meaningful neurobehavioral processing for these populations and evaluated these systems in well controlled clinical studies. Although, the number of subjects involved in the studies and our study timelines somewhat limit our ability to draw conclusions on the ultimate impact of these system related change, the studies proved efficacy of these systems for potential large scale tests in clinical settings. The robot-mediated system (ARIA) was focused on joint attention task, which is considered a major foundation skill in the children with ASD in early development and affects later social communication skills (C. Kasari et al., 2006; C. Kasari, Paparella, T., Freeman, S., Jahromi, L.B., 2008; MacDonald et al., 2006; P. Mundy, Block, J., Delgado, C., Pomares, Y., Van Hecke, A.V., Parlade, M.V., 2007; P. Mundy, Rebecca Neal, A., 2000; P. Mundy & Stella, 2000; Navab, 2011; Poon, 2011). Our virtual reality work is also focused on core deficit, namely facial emotional expression recognition, in both ASD and SZ. There is limited evidence in the role of virtual reality based intervention for this core deficit in both disorders. In addition to adding to the body of literature in the field, the virtual reality based systems advance intervention science from traditional human-centric interventions by providing repeatable and objective performance metrics, adaptive behavior for individualization and measures of multimodal implicit cues such as physiological states and gaze sensitivity that are difficult to measure and combine in a traditional human-centric intervention.

These implicit cues hold vital importance especially for determining and modeling the internal psychological state of the subjects.

CC. Societal Contributions

With the average lifetime cost of care for individual with autism estimated around $3.2 million, average medical expenditures for individuals with ASD 4.1–6.2 times greater than for those without ASD (Ganz, 2007), and with the ever increasing alarming prevalence figures designing more powerful treatments and therapies for the disorder is often considered a public health priority ("Interagency Autism Coordinating Committee Strategic Plan for Autism Spectrum Disorder Research," 2009). Similarly, SZ affecting about 1% of the population, costing more than $100 billion annually in the USA, and with behavioral and pharmacological interventions of somewhat limited effectiveness, finding a more flexible assistive intervention paradigms is of great importance. Evidence based clinical technology-assisted intervention for ASD and SZ hold promise in this respect. The systems developed in this dissertation potentially pave the ways for using intelligent technology for long term therapy of both ASD and SZ. The overarching eventual goal of such systems is not to make the patients dependent on the technology itself. Rather, long term skill trainings with these systems could potentially and eventually translate into real world skills for individuals with ASD and SZ. With the prevalence of ASD increasing at an alarming rate and the cost of both ASD and SZ therapy rapidly increasing, effective technology-based intervention systems could also hold future potential for alleviating the apparent lack of trained therapists in these fields.

APPENDIX A

Details of the Camera Processing Module (CPM)

Camera Processing Module (CPM): These modules are equipped with light-weight contour-based image processing in the XYZ color space to detect the LEDs. Formulated based on the human eye's photo receptors (cones), the XYZ space is characterized by identifying brightness (short, medium, and long wavelength) cues better than other color spaces which is a suitable property for detection of LEDs in the IR spectrum. The tristimuli values were experimentally determined as shown below in (A.1) (Smith & Guild, 1931).

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{0.17697} \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.812401 & 0.01063 \\ 0.00 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{A.1}$$

The detection of the LEDs is in the projective (perspective) image plane with the camera matrix and the focal length is estimated using the well-known Levenberg- Marquardt optimization algorithm.

A three dimensional (3D) approach with only top LEDs array and a stereo camera pair for 3D line reconstruction was first attempted. However, the error was found to be not acceptable for our task. A small depth difference between the two extreme points of the LEDs array creates a large angular measurement error. Subsequently, a two-view two dimensional (2D) solution (which is similar to multi-view scene reconstruction in computer graphics) with 2 side and one top LEDs arrays and multiple cameras was developed to track head movement.

The idea of a pinhole camera model and the 2D projective geometry in the image plane (Hartley & Zisserman, 2003) has been extended to govern the projection of targets (flat LCD monitors in the scene which are not directly observable by the cameras). It is assumed that the perspective projection of the targets is translated and rotated on the projective image plane using the perspective projection of the coordinates from each target to the camera center.

The projections of the arrays to the top and side perspective projective image plane is shown in Fig. 61 together with the respective rays produced from one of the points with lower *y*-coordinate and extended indefinitely. When the child moves his/her head, the LEDs arrays move with the head and so do their 2D projections. Hence the ray will move around the projection plane. The intersections of rays with the projections of the targets in the top and side projection planes give the *x* and *y*-coordinates of the gaze point, respectively. Therefore, a ray-line segment intersection is performed to compute the intersection coordinate in the respective dimensions as shown in Fig. 62.

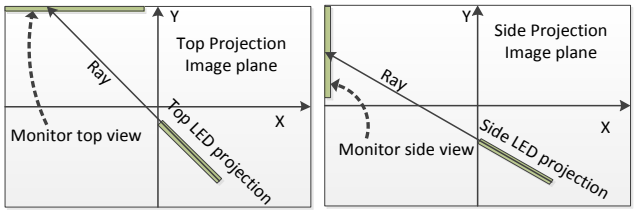


Fig. 61. Top and side perspective projections of the LED arrays and the targets (LCD monitors)

Let P_0P_1 be the projection of one of the targets on the projection plane and Q_0Q_1 be the ray formed from the LEDs projection as described above. Then, we define $\boldsymbol{u} = (u_1 + \boldsymbol{j}u_2)$, and $\boldsymbol{v} = (v_1 + \boldsymbol{j}v_2)$ to be the directional vectors for the target projection, and the LED projection ray, respectively. Vector $\boldsymbol{w} = (w_1 + \boldsymbol{j}w_2)$

is a vector from Q_0P_0. When the ray $(P(s)-Q_0)$ intersects the line extended from the line segment P_0P_1, the vector $\boldsymbol{v}^{\perp}$ is perpendicular to the vector $(P(s_I)-Q_0)$. This is equivalent to the perpendicular product condition (Brandt & Schneider, 2005):

$$\boldsymbol{v}^{\perp}\cdot(\boldsymbol{w}+s\boldsymbol{u})=0 \tag{A.2}$$

This condition can be solved to give the equation for the intersection fractions, s_I, as follows.

$$s_I=\frac{-\boldsymbol{v}^{\perp}\cdot\boldsymbol{w}}{\boldsymbol{v}^{\perp}\cdot\boldsymbol{u}}=\frac{v_2w_1-v_1w_2}{v_1u_2-v_2u_1} \tag{A.3}$$

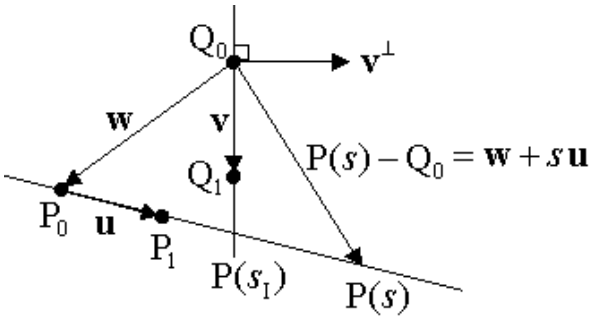


Fig. 62. Ray-line segment intersection for one target (P_0P_1) and a LEDs ray ($Q_0P(s)$)

This algorithm determines the point of intersection in each view (top and side) to get the (*x*,*y*) coordinates of intersection of the gaze direction vector (**v)** with the line segment (target projection to the 2D projective plane). These image projective plane coordinates are projected back to the actual lengths to each LCD monitor reference frame. This system measures the roll and the yaw angles of the head.

APPENDIX B

Physiological features extracted for the social task

Table 18. Physiological Features extracted for the social task

Physiological signal	Feature extracted	Label used	Unit of measurement
Electrocardiogram (ECG/EKG)	Sympathetic power	power_sym	Unit/s^2
	Parasympathetic power	power_para	Unit/s^2
	Very low-frequency power	power_vlf	Unit/s^2
	Ratio of powers	para_vlf	No unit
		para_sym	
		vlf_sym	
	Mean Interbeat Interval (IBI)	mean_ibi_ekg	ms
	Std. of IBI	std_ibi_ekg	Standard deviation(no unit)
Photoplethysmogram (PPG)	Mean and std. of amplitude of the peak values	ppg_peak_mean	µV
	Mean and std. of heart rate variability	ppg_peak_std	No unit
		hrv_mean	ms
		hrv_std	No unit
Electrodermal activity (EDA)	Mean and std. of tonic activity level	SCL_mean	µS
		SCL_sd	µS/s
	Slope of tonic activity	SCL_slope	µS
	Mean and std. of amplitude of skin conductance response (phasic activity)	SCR_mean	
		SCR_sd	µS
	Rate of phasic activity		
	Mean and std. of rise time	SCR_rate	Response peaks/s
		tRise_mean	
	Mean and std. of recovery time	tRise_std	
		tHRecovery_mean	
		tHRevovery_sd	
Electromyographic Activity (EMG)	Mean of Corrugator, Zygomaticus and Trapezius activities	Cemg_mean	µV
		Zemg_mean	
		Temg_mean	
	Std. of Corrugator, Zygomaticus and Trapezius activities	Cemg_std	No unit
		Zemg_std	
		Temg_std	
	Slope of Corrugator, Zygomaticus and Trapezius activities	Cemg_slope	µV/s
		Zemg_slope	
		Temg_slope	
	Number of burst activities per minute of Corrugator, Zygomaticus and Trapezius	Cburst_count	/min
		Zburst_count	
	Mean of Corrugator, Zygomaticus and Trapezius burst activities	Tburst_count	
		Cburst_mean	mS
		Zburst_mean	
	Std. of Corrugator, Zygomaticus and Trapezius burst activities	Tburst_mean	
		Cburst_std	No unit
		Zburst_std	
	Mean and Median frequency of Corrugator, Zygomaticus and Trapezius	Tburst_std	
		Cfreq_mean	Hertz
		Cfreq_med	
		Zfreq_mean	
	Mean of the amplitude of Corrugator, Zygomaticus and Trapezius burst activities	Zfreq_med	
		Tfreq_mean	
		Tfreq_med	
		Cburst_amp_mean	µV
		Zburst_amp_mean	
		Tburst_amp_mean	
Respiration (RSP)	Mean amplitude	RSP_mean	No unit
	Std. of amplitude	RSP_std	
	Subband spectral entropy		

		RSP_subbandSpectralEnt	
	Minimum and maximum difference	ropy(1,2,3)	
	Change rate	RSP_minmax_diff	
	Power spectrum density	RSP_rate	
		RSP_low_power	
		RSP_high_power	
	Std. of Poincare plot geometry	RSP_firstOrder_std	
		RSP_poincare_SD1	
	Mean and std. of peak valley magnitude	RSP_poincare_SD2	
	Mean and std. of breath per minute	PVM_mean	
		PVM_std	
		RRI_mean	
		RRI_std	
Peripheral temperature (SKT)	Mean temperature	temp_mean	F
	Slope of temperature	temp_slope	F/s
	Std. of temperature	temp_std	No unit

REFERENCES

1. Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive neuroscience, 13*(2), 232-240.
2. Aftanas, L., Reva, N., Varlamov, A., Pavlov, S., & Makhnev, V. (2004). Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neuroscience and behavioral physiology, 34*(8), 859-867.
3. Agrawal, P., Liu, C., Sarkar, N. (2008). Interaction between human and robot An affect-inspired approach. *9*(2), 230-257.
4. Almeida, O., Zhang, M., & Liu, J.-C. (2007). *Dynamic fall detection and pace measurement in walking sticks.* Paper presented at the High Confidence Medical Devices, Software, and Systems and Medical Device Plug-and-Play Interoperability, 2007. HCMDSS-MDPnP. Joint Workshop on.
5. Anderson, C. J., Colombo, J., & Jill Shaddy, D. (2006). Visual scanning and pupillary responses in young children with autism spectrum disorder. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1238-1256.
6. Anderson, C. J., Colombo, J., & Shaddy, D. J. (2006). Visual scanning and pupillary responses in young children with autism spectrum disorder. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1238-1256.
7. Andreasen, N. C. (1983). Scale for the assessment of negative symptoms. *University of Iowa, Iowa City*.
8. Andreasen, N. C. (1984). Scale for the assessment of positive symptoms. *Iowa City: University of Iowa*.
9. Andreasen, N. C., Pressler, M., Nopoulos, P., Miller, D., & Ho, B.-C. (2010). Antipsychotic dose equivalents and dose-years: a standardized method for comparing exposure to different drugs. *Biological psychiatry, 67*(3), 255-262.
10. Annaz, D., Campbell, R., Coleman, M., Milne, E., & Swettenham, J. (2012). Young children with autism spectrum disorder do not preferentially attend to biological motion. *Journal of autism and developmental disorders, 42*(3), 401-408.
11. Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry, 28*(3), 289-304.
12. . Autism Spectrum Disorders Prevalence Rate. (2011): Autism Speaks and Center for Disease Control (CDC).
13. Baio, J., Autism, Network, D. D. M., Control, C. f. D., & Prevention. (2012). *Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008*: National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention (CDC), US Department of Health and Human Services.
14. Battiti, R. (1992). First-and second-order methods for learning: between steepest descent and Newton's method. *Neural computation, 4*(2), 141-166.
15. Bauminger, N., Goren-Bar, D., Gal, E., Weiss, P., Kupersmitt, J., Pianesi, F., . . . Zancanaro, M. (2007). *Enhancing social communication in high-functioning children with autism through a co-located interface*. Paper presented at the Multimedia Signal Processing. IEEE 9th Workshop on.
16. Bekele, E., Lahiri, U., Davidson, J., Warren, Z., & Sarkar, N. (2011). *Development of a novel robot-mediated adaptive response system for joint attention task for children with autism*. Paper presented at the 20th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN, , Atlanta, GA.
17. Bekele, E., Young, M., Zheng, Z., Zhang, L., Swanson, A., Johnston, R., . . . Sarkar, N. (2013). A step towards adaptive multimodal virtual social interaction platform for children with autism *Universal Access in Human-Computer Interaction. User and Context Diversity* (pp. 464-473): Springer.
18. Bekele, E., Zheng, Z., Swanson, A., Crittendon, J., Warren, Z., & Sarkar, N. (2013). Understanding How Adolescents with Autism Respond to Facial Expressions in Virtual Reality Environments. *IEEE Transactions on Visualizations and Computer Graphics PP (to appear)*((special issue)).

19. Bekele, E., Zheng, Z., Swanson, A., Crittendon, J., Warren, Z., & Sarkar, N. (2013). Understanding How Adolescents with Autism Respond to Facial Expressions in Virtual Reality Environments. *Visualization and Computer Graphics, IEEE Transactions on, 19*(4), 711-720.
20. Bellani, M., Fornasari, L., Chittaro, L., & Brambilla, P. (2011). Virtual reality in autism: state of the art. *Epidemiology and psychiatric sciences, 20*(03), 235-238.
21. Berenbaum, H., & Oltmanns, T. F. (1992). Emotional experience and expression in schizophrenia and depression. *Journal of abnormal psychology, 101*(1), 37.
22. Bernard-Opitz, V., Sriram, N., Nakhoda-Sapuan, S. (2001). Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. *Journal of Autism and Developmental Disorders, 31*(4), 377-384.
23. Bernstein, A. S., Frith, C. D., Gruzelier, J. H., Patterson, T., Straube, E., Venables, P. H., & Zahn, T. P. (1982). An analysis of the skin conductance orienting response in samples of American, British, and German schizophrenics. *Biological Psychology, 14*(3), 155-211.
24. Billard, A., Robins, B., Nadel, J., & Dautenhahn, K. (2006). Building robota, a mini-humanoid robot for the rehabilitation of children with autism. *RESNA Assistive Technology Journal, 19*(1), 37-49.
25. Bird, G., Leighton, J., Press, C., Heyes, C. (2007). Intact automatic imitation of human and robot actions in autism spectrum disorders. *Proceedings of the Royal Society B: Biological Sciences, 274*(1628), 3027–3031.
26. Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: a revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3*(2), 129-136.
27. Bleuler, E. (1911). Dementia praecox or the group of schizophrenias. *International University Press, New York*.
28. Blocher, K., Picard, R.W. (2002). Affective social quest: emotion recognition therapy for autistic children: In Socially Intelligent Agents-Creating Relationships with Computers and Robots, Citeseer.
29. Bölte, S., Hubl, D., Feineis-Matthews, S., Prvulovic, D., Dierks, T., & Poustka, F. (2006). Facial affect recognition training in autism: can we animate the fusiform gyrus? *Behavioral neuroscience, 120*(1), 211.
30. Bourkeφ, A., Scanaillφ, C. N., Culhaneφ, K., O'Brien, J., & Lyonsφ, G. (2006). An optimum accelerometer configuration and simple algorithm for accurately detecting falls.
31. Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1996). Picture media and emotion: Effects of a sustained affective context. *Psychophysiology, 33*(6), 662-670.
32. Brandt, J., & Schneider, K. (2005). Using three-valued logic to specify and verify algorithms of computational geometry. *Formal Methods and Software Engineering*, 405-420.
33. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery, 2*(2), 121-167.
34. Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Handbook of psychophysiology*: Cambridge Univ Pr.
35. Capps, L., Yirmiya, N., & Sigman, M. (1992). Understanding of Simple and Complex Emotions in Non-retarded Children with Autism. *Journal of Child Psychology and Psychiatry, 33*(7), 1169-1182.
36. Castelli, F. (2005). Understanding emotions from standardized facial expressions in autism and normal development. *Autism, 9*(4), 428-449.
37. Celani, G., Battacchi, M. W., & Arcidiacono, L. (1999). The understanding of the emotional meaning of facial expressions in people with autism. *Journal of autism and developmental disorders, 29*(1), 57-66.
38. Chasson, G. S., Harris, G. E., & Neely, W. J. (2007). Cost comparison of early intensive behavioral intervention and special education for children with autism. *Journal of Child and Family Studies, 16*(3), 401-413.
39. Cheung, C., Yu, K., Fung, G., Leung, M., Wong, C., Li, Q., . . . McAlonan, G. (2010). Autistic disorders and schizophrenia: related or remote? An anatomical likelihood estimation. *PloS one, 5*(8), e12233.

40. Cohen, A. S., & Minor, K. S. (2010). Emotional experience in patients with schizophrenia revisited: meta-analysis of laboratory studies. *Schizophrenia Bulletin, 36*(1), 143-150.
41. Conn, K., Liu, C., Sarkar, N., Stone, W., Warren, Z. (2008). *Affect-sensitive assistive intervention technologies for children with autism: An individual-specific approach*. Paper presented at the The 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN, Munich, Germany, .
42. Constantino, J., & Gruber, C. (2002). The social responsiveness scale. *Los Angeles: Western Psychological Services*.
43. Couture, S. M., Granholm, E. L., & Fish, S. C. (2011). A path model investigation of neurocognition, theory of mind, social competence, negative symptoms and real-world functioning in schizophrenia. *Schizophrenia research, 125*(2), 152-160.
44. Dautenhahn, K. (2003). Roles and functions of robots in human society: implications from research in autism therapy. *21*(4), 443-452.
45. Dautenhahn, K., Billard, A. (2002). Games children with autism can play with Robota, a humanoid robotic doll. University of Cambridge, cambridge, pp. 179-190: Proceedings of 1st Cambridge workshop on Universal access and assistive technology, Springer.
46. Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A., & Blow, M. (2009). KASPAR–a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics, 6*(3-4), 369-397.
47. Dautenhahn, K., Werry, I., Rae, J., Dickerson, P., Stribling, P., & Ogden, B. (2002). *Robotic playmates: Analysing interactive competencies of children with autism playing with a mobile robot*. pp. 117-124: Socially Intelligent Agents – Creating Relationships with Computers and Robots, Dordrecht, Kluwer Academic Publishers.
48. Dautenhahn, K., Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition, 12*(1), 1-35.
49. Dawson, G., Webb, S. J., Carver, L., Panagiotides, H., & McPartland, J. (2004). Young children with autism show atypical brain responses to fearful versus neutral facial expressions of emotion. *Developmental Science, 7*(3), 340-359.
50. Dawson, G., Webb, S. J., & McPartland, J. (2005). Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental neuropsychology, 27*(3), 403-424.
51. Dawson, M. E., & Nuechterlein, K. H. (1984). Psychophysiological dysfunctions in the developmental course of schizophrenic disorders. *Schizophrenia Bulletin, 10*(2), 204-232.
52. Demaree, H. A., Everhart, D. E., Youngstrom, E. A., & Harrison, D. W. (2005). Brain lateralization of emotional processing: historical roots and a future incorporating "dominance". *Behavioral and cognitive neuroscience reviews, 4*(1), 3-20.
53. Demchak, M. A. (1990). Response prompting and fading methods: A review. *American Journal on Mental Retardation, 94*(6), 603-615.
54. Developmental, D. M. N. S. Y., & Investigators, P. (2014). 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), 63*, 1.
55. Di, P., Huang, J., Sekiyama, K., & Fukuda, T. (2011). *Motion control of intelligent cane robot under normal and abnormal walking condition.* Paper presented at the RO-MAN, 2011 IEEE.
56. *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the diagnostic criteria from DSM-IV-TR*. (2000). Washington, DC: American Psychiatric Association, Amer Psychiatric Pub Incorporated.

57. Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2011). The clinical use of robots for individuals with Autism Spectrum Disorders: A critical review. *Research in Autism Spectrum Disorders, 6*(1), 249-262. doi: 10.1016/j.rasd.2011.05.006
58. Doop, M. L., & Park, S. (2006). On knowing and judging smells: identification and hedonic judgment of odors in schizophrenia. *Schizophrenia research, 81*(2), 317-319.
59. Duquette, A., Michaud, F., Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots, 24*(2), 147-157.
60. Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., & Mathiak, K. (2010). Virtual faces as a tool to study emotion recognition deficits in schizophrenia. *Psychiatry research, 179*(3), 247-252.
61. Earnst, K. S., & Kring, A. M. (1999). Emotional responding in deficit and non-deficit schizophrenia. *Psychiatry research, 88*(3), 191-207.
62. Eikeseth, S. (2009). Outcome of comprehensive psycho-educational interventions for young children with autism. *Research in Developmental Disabilities, 30*(1), 158-178.
63. Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.
64. Fabri, M., Elzouki, S. Y. A., & Moore, D. (2007). *Emotionally expressive avatars for chatting, learning and therapeutic intervention*. Paper presented at the Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments.
65. Feil-Seifer, D., & Mataric, M. J. (2008). *B3IA: A control architecture for autonomous robot-assisted behavior intervention for children with Autism Spectrum Disorders*. Paper presented at the The 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN.
66. Feil-Seifer, D., Mataric, M. (2011). *Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features*. Paper presented at the In Proceedings of the 6th international conference (ACM/IEEE) on Human-robot interaction, New York, NY: ACM Press.
67. Feil-Seifer, D., Matarić, M. (2009). *Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders*. Paper presented at the Experimental Robotics, 54.
68. Freeman, D. (2008). Studying and treating schizophrenia using virtual reality: a new paradigm. *Schizophrenia Bulletin, 34*(4), 605-610.
69. Gal, E., Bauminger, N., Goren-Bar, D., Pianesi, F., Stock, O., Zancanaro, M., & Weiss, P. L. (2009). Enhancing social communication of children with high-functioning autism through a co-located interface. *AI & Society, 24*(1), 75-84.
70. Ganz, M. L. (2007). The lifetime distribution of the incremental societal costs of autism. *Archives of Pediatrics and Adolescent Medicine, Am Med Assoc, 161*(4), 343-349.
71. Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *Journal of autism and developmental disorders, 40*(3), 269-279.
72. Golan, O., & Baron-Cohen, S. (2006). Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. *Development and psychopathology, 18*(02), 591-617.
73. Goodwin, M. S. (2008). Enhancing and Accelerating the Pace of Autism Research and Treatment. *Focus on Autism and Other Developmental Disabilities, 23*(2), 125-128.
74. Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders, 37*(4), 613-627.
75. Gotlib, I. H. (1998). EEG alpha asymmetry, depression, and cognitive functioning. *Cognition & Emotion, 12*(3), 449-478.
76. Green, M. J., Williams, L. M., & Davidson, D. (2003). Visual scanpaths to threat-related faces in deluded schizophrenia. *Psychiatry research, 119*(3), 271-285.

77. Groden, J., Goodwin, M. S., Baron, M. G., Groden, G., Velicer, W. F., Lipsitt, L. P., . . . Plummer, B. (2005). Assessing cardiovascular responses to stressors in individuals with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 20*(4), 244-252.
78. Gruzelier, J., & Venables, P. (1972). SKIN CONDUCTANCE ORIENTING ACTIVITY IN A HETEROGENEOUS SAMPLE OF SCHIZOPHRENICS: Possible Evidence of Limbic Dysfunction. *The Journal of nervous and mental disease, 155*(4), 277-287.
79. Gutiérrez-Maldonado, J., Rus-Calafell, M., Márquez-Rejón, S., & Ribas-Sabaté, J. (2012). Associations Between Facial Emotion Recognition, Cognition and Alexithymia in Patients with Schizophrenia: Comparison of Photographic and Virtual Reality Presentations. *Stud Health Technol Inform, 181*, 88-92.
80. Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on, 5*(6), 989-993.
81. Harris, K., & Reid, D. (2005). The influence of virtual reality play on children's motivation. *Canadian Journal of Occupational Therapy, 72*(1), 21-30.
82. Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*: Cambridge university press.
83. Healey, K. M., Pinkham, A. E., Richard, J. A., & Kohler, C. G. (2010). Do we recognize facial expressions of emotions from persons with schizophrenia? *Schizophrenia research, 122*(1), 144-150.
84. Hempel, R. J., Tulen, J. H., van Beveren, N. J., Mulder, P. G., & Hengeveld, M. W. (2007). Subjective and physiological responses to emotion-eliciting pictures in male schizophrenic patients. *International Journal of Psychophysiology, 64*(2), 174-183.
85. Hempel, R. J., Tulen, J. H., van Beveren, N. J., van Steenis, H. G., Mulder, P. G., & Hengeveld, M. W. (2005). Physiological responsivity to emotional pictures in schizophrenia. *Journal of psychiatric research, 39*(5), 509-518.
86. Herbener, E. S., Song, W., Khine, T. T., & Sweeney, J. A. (2008). What aspects of emotional functioning are impaired in schizophrenia? *Schizophrenia research, 98*(1), 239-246.
87. Hobson, R. P. (1986). The autistic child's appraisal of expressions of emotion. *Journal of Child Psychology and Psychiatry, 27*(3), 321-342.
88. Hobson, R. P., Ouston, J., & Lee, A. (1988). Emotion recognition in autism: Coordinating faces and voices. *Psychological Medicine, 18*(4), 911-923.
89. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks, 2*(5), 359-366.
90. Hsiao, J. H., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science, 19*(10), 998-1006.
91. Huang, J., Di, P., Wakita, K., Fukuda, T., & Sekiyama, K. (2008). *Study of fall detection using intelligent cane based on sensor fusion.* Paper presented at the Micro-NanoMechatronics and Human Science, 2008. MHS 2008. International Symposium on.
92. Ikezawa, S., Corbera, S., Liu, J., & Wexler, B. E. (2012). Empathy in electrodermal responsive and nonresponsive patients with schizophrenia. *Schizophrenia research, 142*(1), 71-76.
93. . Interagency Autism Coordinating Committee Strategic Plan for Autism Spectrum Disorder Research. (2009): USDHHS Document, http://iacc.hhs.gov/strategic-plan/2009/, U.S. Department of Health and Human Services.
94. Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.
95. Johnson, K. L., & Tassinary, L. G. (2005). Perceiving sex directly and indirectly meaning in motion and morphology. *Psychological Science, 16*(11), 890-897.
96. Josman, N., Ben-Chaim, H. M., Friedrich, S., & Weiss, P. L. (2011). Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism. *International Journal on Disability and Human Development, 7*(1), 49-56.

97. Kandalaft, M. R., Didehbani, N., Krawczyk, D. C., Allen, T. T., & Chapman, S. B. (2012). Virtual Reality Social Cognition Training for Young Adults with High-Functioning Autism. *Journal of autism and developmental disorders*, 1-11.
98. Karson, C. N., Bigelow, L., Kleinman, J., Weinberger, D., & Wyatt, R. (1982). Haloperidol-induced changes in blink rates correlate with changes in BPRS score. *The British Journal of Psychiatry, 140*(5), 503-507.
99. Kasari, C., Freeman, S., & Paparella, T. (2006). Joint attention and symbolic play in young children with autism: a randomized controlled intervention study. *Journal of Child Psychology and Psychiatry, 47*(6), 611-620.
100. Kasari, C., Gulsrud, A.C., Wong, C., Kwon, S., Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. *Journal of autism and developmental disorders, 40*(9), 1045-1056.
101. Kasari, C., Paparella, T., Freeman, S., Jahromi, L.B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. *Journal of consulting and clinical psychology, 76*(1), 125–137.
102. Keil, A., Müller, M. M., Gruber, T., Wienbruch, C., Stolarova, M., & Elbert, T. (2001). Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical neurophysiology, 112*(11), 2057-2068.
103. Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion, 13*(5), 467-480.
104. Kenny, P., Parsons, T., Gratch, J., Leuski, A., & Rizzo, A. (2007). *Virtual patients for clinical therapist skills training*. Paper presented at the Intelligent Virtual Agents.
105. Kim, E. S., Paul, R., Shic, F., & Scassellati, B. (2012). Bridging the research gap: making HRI useful to individuals with autism. *Journal of Human-Robot Interaction, 1*(1).
106. Kim, J., & Ande, E. (2008). Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30*(12), 2067-2083.
107. Kim, K. H., Bang, S., & Kim, S. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing, 42*(3), 419-427.
108. Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry, 59*(9), 809.
109. Klin, A., Lin, D.J., Gorrindo, P., Ramsay, G., Jones, W. (2009). Two-year-olds with autism orient to nonsocial contingencies rather than biological motion. *Nature, 459*(7244), 257–261.
110. Knapp, M., Romeo, R., & Beecham, J. (2009). Economic cost of autism in the UK. *Autism, 13*(3), 317-336.
111. Kozima, H., Nakagawa, C., Yasuda, Y. (2005). *Interactive robots for communication-care: A case-study in autism therapy*. Paper presented at the IEEE International Workshop on Robot and Human Interactive Communication, ROMAN, Nashville, TN.
112. Kring, A. M., Kerr, S. L., & Earnst, K. S. (1999). Schizophrenic patients show facial reactions to emotional facial expressions. *Psychophysiology, 36*(02), 186-192.
113. Kring, A. M., Kerr, S. L., Smith, D. A., & Neale, J. M. (1993). Flat affect in schizophrenia does not reflect diminished subjective experience of emotion. *Journal of abnormal psychology, 102*(4), 507.
114. Kring, A. M., & Moran, E. K. (2008). Emotional response deficits in schizophrenia: insights from affective science. *Schizophrenia Bulletin, 34*(5), 819-834.
115. Kring, A. M., & Neale, J. M. (1996). Do schizophrenic patients show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion? *Journal of abnormal psychology, 105*(2), 249.

116. Lacava, P. G., Rankin, A., Mahlios, E., Cook, K., & Simpson, R. L. (2010). A single case design evaluation of a software and tutor intervention addressing emotion recognition and social interaction in four boys with ASD. *Autism*.
117. Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., & Sarkar, N. (2012). Design of a Virtual Reality based Adaptive Response Technology for Children with Autism. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society, PP (early access)*(99), 1.
118. Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., Sarkar, N. (2011). *Design of a Virtual Reality Based Adaptive Response Technology for Children with Autism Spectrum Disorder*. Paper presented at the Affective Computing and Intelligent Interaction (ACII, 2011), Springer, Memphis, TN.
119. Lahiri, U., Warren, Z., & Sarkar, N. (2012). Design of a Gaze-Sensitive Virtual Social Interactive System for Children With Autism. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*(99), 1-1.
120. Lahiri, U., Warren, Z., Sarkar, N. (2011). Design of a Gaze-Sensitive Virtual Social Interactive System for Children With Autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 19*(4), 443-452.
121. Lan, M., Nahapetian, A., Vahdatpour, A., Au, L., Kaiser, W., & Sarrafzadeh, M. (2009). *SmartFall: an automatic fall detection system based on subsequence matching for the SmartCane.* Paper presented at the Proceedings of the Fourth International Conference on Body Area Networks.
122. Landry, R., & Bryson, S. E. (2004). Impaired disengagement of attention in young children with autism. *Journal of Child Psychology and Psychiatry, 45*(6), 1115-1122.
123. Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Technical manual and affective ratings: Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.
124. Leuski, A., Patel, R., Traum, D., & Kennedy, B. (2009). *Building effective question answering characters.* Paper presented at the Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue.
125. Libby Jr, W. L., Lacey, B. C., & Lacey, J. I. (1973). Pupillary and cardiac activity during visual attention. *Psychophysiology, 10*(3), 270-294.
126. Liu, C., Agrawal, P., Sarkar, N., Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *25*(6), 506-529.
127. Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. *Robotics, IEEE Transactions on, 24*(4), 883-896.
128. Liu, C., Conn, K., Sarkar, N., Stone, W. (2007a). *Affect recognition in robot assisted rehabilitation of children with autism spectrum disorder*. Paper presented at the Proc. of the 15th IEEE Intl. Conf. on Robotics and Automation.
129. Liu, C., Conn, K., Sarkar, N., Stone, W. (2007b). *Online Affect Detection and Adaptation in Robot Assisted Rehabilitation for Children with Autism*. Paper presented at the The 16th IEEE International Symposium on Robot and Human interactive Communication, RO-MAN, Jeju Island, Korea
130. Liu, C., Conn, K., Sarkar, N., Stone, W. (2008a). Online affect detection and robot behavior adaptation for intervention of children with autism. *Robotics, IEEE Transactions on, 24*(4), 883 - 896.
131. Liu, C., Conn, K., Sarkar, N., Stone, W. (2008b). Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *International Journal of Human-Computer Studies, Elsevier, 66*(9), 662-677.
132. Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., . . . Rutter, M. (2000). 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, 30*(3), 205-223.

133. MacDonald, R., Anderson, J., Dube, W. V., Geckeler, A., Green, G., Holcomb, W., . . . Sanchez, J. (2006). Behavioral assessment of joint attention: a methodological report. *Research in developmental disabilities, 27*(2), 138-150.
134. Manor, B. R., Gordon, E., Williams, L. M., Rennie, C. J., Bahramali, H., Latimer, C. R., . . . Meares, R. A. (1999). Eye movements reflect impaired face processing in patients with schizophrenia. *Biological psychiatry, 46*(7), 963-969.
135. Mattes, R., Schneider, F., Heimann, H., & Birbaumer, N. (1995). Reduced emotional response of schizophrenic patients in remission during social interaction. *Schizophrenia research, 17*(3), 249-255.
136. Michaud, F., Théberge-Turmel, C. (2002). *Mobile robotic toys and autism*. Paper presented at the Socially Intelligent Agents: Creating Relationships with Computers and Robots.
137. Millen, L., Cobb, S., & Patel, H. (2010). *Participatory design with children with autism*. Paper presented at the Proc. 8th Intl Conf. Disability, Virtual Reality & Associated Technologies, Valparaíso, Chile.
138. Mitchell, P., Parsons, S., Leonard, A. (2007). Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders. *Journal of autism and developmental disorders, 37*(3), 589-600.
139. Moore, D., McGrath, P., Thorpe, J. (2000). Computer-aided learning for people with autism-a framework for research and development. *Innovations in Education and Training International, 37*(3), 218-228.
140. More, J. (1978). The Levenberg-Marquardt algorithm: implementation and theory. *Numerical analysis: Lecture Notes in Mathematics, 630*, 105-116.
141. Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A.V., Parlade, M.V. (2007). Individual differences and the development of joint attention in infancy. *Child development, 78*(3), 938-954.
142. Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., & Seibert, J. (2003). *Early Social Communikation Scales (ESCS)*. Coral Gables, FL: University of Miami.
143. Mundy, P., Rebecca Neal, A. (2000). Neural plasticity, joint attention, and a transactional social-orienting model of autism. *International review of research in mental retardation, Autism, 23*, 139-168.
144. Mundy, P., & Stella, J. (2000). Joint attention, social orienting, and nonverbal communication in autism. *Autism spectrum disorders: A transactional developmental perspective, 9*, 55-77.
145. Navab, A., Gillespie Lynch, K., Johnson, S.P., Sigman, M., Hutman, T. (2011). Eye Tracking as a Measure of Responsiveness to Joint Attention in Infants at Risk for Autism. *Journal of Infancy, International Society on Infant Studies (ISIS), 17*(4), 416-431. doi: 10.1111/j.1532-7078.2011.00082.x
146. Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological reports, 10*(3), 799-812.
147. Park, K.-M., Ku, J., Choi, S.-H., Jang, H.-J., Park, J.-Y., Kim, S. I., & Kim, J.-J. (2011). A virtual reality application in role-plays of social skills training for schizophrenia: a randomized, controlled trial. *Psychiatry research, 189*(2), 166-172.
148. Parsons, S., & Cobb, S. (2011). State-of-the-art of Virtual Reality technologies for children on the autism spectrum. *European Journal of Special Needs Education, 26*(3), 355-366.
149. Parsons, S., Mitchell, P. (2002). The potential of virtual reality in social skills training for people with autistic spectrum disorders. *Journal of Intellectual Disability Research, 46*(5), 430-443.
150. Parsons, S., Mitchell, P., Leonard, A. (2004). The use and understanding of virtual environments by adolescents with autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 34*(4), 449-466.
151. Parsons, T. D., Rizzo, A. A., Rogers, S., & York, P. (2009). Virtual reality in paediatric rehabilitation: A review. *Developmental Neurorehabilitation, 12*(4), 224-238.
152. Peacock, G., Amendah, D., Ouyang, L., & Grosse, S. D. (2012). Autism spectrum disorders and health care expenditures: the effects of co-occurring conditions. *Journal of Developmental & Behavioral Pediatrics, 33*(1), 2-8.
153. Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders, 32*(4), 249-261.

154. Picard, R. W. (1997). *Affective computing*: MIT Press, Cambridge.
155. Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies, 59*(1), 55-64.
156. Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3575-3584.
157. Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23*(10), 1175-1191.
158. Pierno, A. C., Mari, M., Lusher, D., Castiello, U. (2008). Robotic movement elicits visuomotor priming in children with autism. *Neuropsychologia, 46*(2), 448-454.
159. Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560-569.
160. Platt, J. C. (1998). 12 Fast Training of Support Vector Machines using Sequential Minimal Optimization.
161. Ploog, B. O., Scharf, A., Nelson, D., & Brooks, P. J. (2012). Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders. *Journal of autism and developmental disorders*, 1-22.
162. Poon, K. K., Watson, L.R., Baranek, G.T., Poe, M.D. (2011). *To What Extent Do Joint Attention, Imitation, and Object Play Behaviors in Infancy Predict Later Communication and Intellectual Functioning in ASD?* Paper presented at the Journal of Autism and Developmental Disorders.
163. . Prevalence of Autsim Spectrum Disorders-ADDM Network. (2009). pp. 1-20: MMWR Weekly Report, 58, United States, Centers for Disease Control and Prevention [CDC].
164. Rani, P., Liu, C., Sarkar, N., & Vanman, E. (2006). An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis & Applications, 9*(1), 58-69.
165. Rani, P., Sarkar, N., Smith, C. A., & Kirby, L. D. (2004). Anxiety detecting robotic system–towards implicit human-robot collaboration. *Robotica, 22*(01), 85-95.
166. REILLY, R., & NOLAN, H. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection.
167. Riedmiller, M., & Braun, H. (1993). *A direct adaptive method for faster backpropagation learning: The RPROP algorithm*.
168. Riva, G. (2005). Virtual reality in psychotherapy: review. *Cyberpsychology & behavior, 8*(3), 220-230.
169. Robins, B., Dautenhahn, K., & Dickerson, P. (2009). *From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot*. Paper presented at the Advances in Computer-Human Interactions, ACHI '09. Second International Conferences on.
170. Robins, B., Dautenhahn, K., Te Boekhorst, R., & Billard, A. (2004). *Effects of repeated exposure to a humanoid robot on children with autism*. Paper presented at the Presented at Cambridge Worshop Universal Access and Assistive Technology (CWUAAT).
171. Robins, B., Dautenhahn, K., Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies, 7*(3), 509-542.
172. Robins, B., Dickerson, P., Stribling, P., & Dautenhahn, K. (2004). Robot-mediated joint attention in children with autism: A case study in robot-human interaction. *Interaction Studies, 5*(2), 161-198.
173. Rogers, S. J., & Ozonoff, S. (2005). Annotation: What do we know about sensory dysfunction in autism? A critical review of the empirical evidence. *Journal of Child Psychology and Psychiatry, 46*(12), 1255-1268.
174. Ruble, L. A., & Robson, D. M. (2007). Individual and environmental determinants of engagement in autism. *Journal of autism and developmental disorders, 37*(8), 1457-1468.
175. Rus-Calafell, M., Gutiérrez-Maldonado, J., & Ribas-Sabaté, J. (2014). A virtual reality-integrated program for improving social skills in patients with schizophrenia: A pilot study. *Journal of behavior therapy and experimental psychiatry, 45*(1), 81-89.

176. Ruse, S. A., Harvey, P. D., Davis, V. G., Atkins, A. S., Fox, K. H., & Keefe, R. S. (2014). Virtual reality functional capacity assessment in schizophrenia: Preliminary data regarding feasibility and correlations with cognitive and functional capacity performance. *Schizophrenia Research: Cognition, 1*(1), e21-e26.
177. Rutter, M., Bailey, A., Lord, C., & Berument, S. (2003). Social communication questionnaire. *Los Angeles, CA: Western Psychological Services*.
178. Salvucci, D. D., & Goldberg, J. H. (2000). *Identifying fixations and saccades in eye-tracking protocols*. Paper presented at the Proceedings of the 2000 symposium on Eye tracking research & applications.
179. Sandall, S. R. (2005). DEC recommended practices: a comprehensive guide for practical application in early intervention/early childhood special education.
180. Scassellati, B., Henny Admoni, & Matarić, M. (2012). Humanoid Robots for Use in Autism Diagnosis/Research. *Annual Review of Biomedical Engineering, 14*(1), 275-294.
181. Schmidt, C., & Schmidt, M. (2008). *Three-dimensional virtual learning environments for mediating social skills acquisition among individuals with autism spectrum disorders*. Paper presented at the Proceedings of the 7th international conference on Interaction design and children.
182. Sears, L. L., Finn, P. R., & Steinmetz, J. E. (1994). Abnormal classical eye-blink conditioning in autism. *Journal of autism and developmental disorders, 24*(6), 737-751.
183. Shiffman, H. (2001). Fundamental visual functions and phenomena *sensation and perception: An Integrated Approach* (pp. 89-115): John Welsly and Sons, New York.
184. Smith, T., & Guild, J. (1931). The CIE colorimetric standards and their use. *Transactions of the Optical Society, 33*(3), 73.
185. Standen, P. J., Brown, D.J. (2005). Virtual reality in the rehabilitation of people with intellectual disabilities: review. *Cyberpsychology & behavior, 8*(3), 272-282.
186. Strickland, D., Marcus, L. M., Mesibov, G. B., & Hogan, K. (1996). Brief report: Two case studies using virtual reality as a learning tool for autistic children. *Journal of autism and developmental disorders, 26*(6), 651-659.
187. Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters, 9*(3), 293-300.
188. Tapus, A., Mataric, M., Scassellati, B. (2007). The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine, 14*(1), 35-42.
189. Tsang, M. M., & Man, D. W. (2013). A virtual reality-based vocational training system (VRVTS) for people with schizophrenia in vocational rehabilitation. *Schizophrenia research, 144*(1), 51-62.
190. Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(1), 111-121.
191. Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 2): Wiley New York.
192. Venables, P., & Wing, J. (1962). Level of arousal and the subclassification of schizophrenia. *Archives of general psychiatry, 7*(2), 114-119.
193. Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing, 27*(12), 1743-1759.
194. Volz, M., Hamm, A. O., Kirsch, P., & Rey, E.-R. (2003). Temporal course of emotional startle modulation in schizophrenia patients. *International Journal of Psychophysiology, 49*(2), 123-137.
195. Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of abnormal psychology, 97*(4), 487.
196. Wagner, J., Kim, J., & André, E. (2005). *From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification*.
197. Wang, M., & Reid, D. (2011). Virtual Reality in Pediatric Neurorehabilitation: Attention Deficit Hyperactivity Disorder, Autism and Cerebral Palsy. *Neuroepidemiology, 36*(1), 2-18.

198. Warren Z. E., Veenstra-VanderWeele J, & Stone W. (2011). Therapies for Children with Autism Spectrum Disorders. *Comparative Effectiveness Review*. Rockville, MD: AHRQ Publication: Agency for Healthcare Research and Quality.
199. Wechsler, D. (2008). *Wechsler Abbreviated Scale of Intelligence® – Fourth Edition (WASI®-IV)*. San Antonio, TX: Harcourt Assessment, The Psychological Corporation.
200. Weeks, S. J., & Hobson, R. P. (1987). The salience of facial expression for autistic children. *Journal of Child Psychology and Psychiatry, 28*(1), 137-152.
201. Welch, K., Lahiri, U., Liu, C., Weller, R., Sarkar, N., Warren, Z. (2009). *An Affect-Sensitive Social Interaction Paradigm Utilizing Virtual Reality Environments for Autism Intervention*. Paper presented at the Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction.
202. Welch, K., Sarkar, M., Sarkar, N., Liu, C. (2010). Affective Modeling for Children with Autism Spectrum Disorders: Application of Active Learning. In L. Berhardt (Ed.), *Advances in Medicine and Biology* (Vol. 8, pp. 297-310). New York: Nova Science Publishers.
203. Welch, K. C., Lahiri, U., Sarkar, N., Warren, Z., Stone, W., Liu, C. (2010). Affect-Sensitive Computing and Autism. In G. Y. Didem Gökçay (Ed.), *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives* (pp. 325-343): Information Science Reference.
204. Werry, I., Dautenhahn, K., Harwin, W. (2001). *Investigating a robot as a therapy partner for children with autism*. Paper presented at the 5th European Conference for the Advancement of Assistive Technology: added value to the quality of life (AAATE '01).
205. Wiens, J., & Guttag, J. (2010). *Patient-adaptive ectopic beat classification using active learning*.
206. Williams, L. M., Loughland, C. M., Gordon, E., & Davidson, D. (1999). Visual scanpaths in schizophrenia: is there a deficit in face recognition? *Schizophrenia research*, *40*(3), 189-199.
207. Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G. R., & Vogeley, K. (2010). It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, nsq024.
208. Wolf, K., Mass, R., Kiefer, F., Wiedemann, K., & Naber, D. (2006). Characterization of the facial expression of emotions in schizophrenia patients: preliminary findings with a new electromyography method. *Canadian journal of psychiatry, 51*(6), 335.
209. Yoder, P. J., McDuffie, A.S. (2006). Treatment of responding to and initiating joint attention. In T. Charman & W. Stone (Eds.), *Social and communication development in autism spectrum disorders: Early identification, diagnosis, and intervention* (pp. 117-142). New York, NY: Guilford Press
210. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(1), 39-58.