Use of a Wearable Camera System in Conversation: Toward a Companion Tool for Social-Emotional Learning in Autism

by

Alea Chandler Teeters

Bachelor of Science in Electrical Engineering
Massachusetts Institute of Technology, 2001

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements of the degree of
Master of Science
at the Massachusetts Institute of Technology

September 2007

Signature of Author ____

Program in Media Arts and Sciences
August 10, 2007

Certified by ____

Rosalind W. Picard
Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ____

Prof. Deb Roy
Chairperson, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences

Use of a Wearable Camera System in Conversation: Toward a Companion Tool for Social-Emotional Learning in Autism

by

Alea Chandler Teeters

Bachelor of Science in Electrical Engineering
Massachusetts Institute of Technology, 2001

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning, on August 15, 2007
in partial fulfillment of the requirements of the degree of
Master of Science
at the Massachusetts Institute of Technology

September 2007

ABSTRACT

Self-Cam is a wearable camera system that allows a person to collect video and audio from the movements of her own head and face. Like looking in a mirror, live feedback or video playback from the camera can be used to experience and learn how you look to others. Video playback and analysis tools can also be used to review and learn how others wearing the Self-Cam express themselves to you. We are developing a tool to teach the kind of facial analysis that an empathetic person might notice in interpretation of everyday interaction. Because it works from real life experience from a single person's point of view, we hypothesize that this analysis of self and immediate social environment will help the process of generalization of facial expression and mental state inference for that person, giving them a better understanding of the significance of facial movements and improvement in recognition of social cues. As steps toward this investigation, this thesis constructs a wearable camera system, designs a process of expression collection and analysis, and develops and implements a video test to evaluate the recognition abilities of study subjects throughout the investigation. Preliminary results show a great difference in ability between individual autistic subjects, some of whom approach the abilities of well scoring neurotypical individuals.

Thesis supervisor: Rosalind Picard
Title: Professor of Media Arts and Sciences, Co-Director, Things That Think

Use of a Wearable Camera System in Conversation: Toward a Companion Tool for Social-Emotional Learning in Autism

by

Alea Chandler Teeters

Advisor

Rosalind W. Picard
Professor of Media Arts and Sciences
Co-Director, Things That Think
Program in Media Arts and Sciences, MIT

Reader

Janet Sonenberg
Professor of Music and Theater Arts
Chairman of Music and Theater Arts, MIT

Reader

Ted Selker
Associate Professor of Media Arts and Sciences
Benesse Career Development Professor of Research in Education
Program in Media Arts and Sciences, MIT

ACKNOWLEDGMENTS

Thank you to everyone who has helped me get through this process as friends, family, technical experts, or simply a pair of healthy hands. I could not have done it without you.

Thank you to my advisor, Rosalind Picard, who is always trying to understand things better and who shares my interest in questioning the status quo.

Thank you Readers. Janet Sonenberg for your high level comments, stories, and for looking for the work as a whole. Ted Selker for sharing your personal experience with autism and for all your great feedback on my physical creations.

Thank you Mia Shandell for being my hands through to the end, for jumping in at a moments notice, and for all the help in writing, researching, and doing hours of laborious computer tasks, complete with thesis stress that was not your own. Thank you Seth Raphael for constant encouragement, for laughter, for fun, and for reminding me of all the things in life more important than thesis. Thanks for setting up the software intricacies and for always being there to help with anything. Thank you Jay and Jodi for being there when I needed you the most, for all of your love. Thank you Jim for all your crazy projects and excitement and for always being around.

Thank you Rana el Kaliouby for many conversations and helping to point me in the right directions, and for being a good collaborator throughout this project. Thank you Matthew Goodwin for your excitement and extensive knowledge in making this project real, and for the opportunity to work directly in an environment where my work could become meaningful. Thanks Clayton for taking care of the server and the interface, always with good spirits.

Thanks to the rest of my group and their families, past, present, and future, including Shani, Layla, Julian, Hoda, Hyungil, Selene, Jen Lee, Andy, and Barry.

Thanks to my roommates, Rikky and Rebecca, and to my friends in JP for giving me a relaxing and fun place to come home to. Thanks Andrew, Emma, Laurie, Annie, Seth, and everyone else.

Thank you Matthew Lerner, Kelley and the Asperger's group, and my subjects at the Groden Center.

And thank you first and last to my family for all of their love and support – Mom, Dad, Kelsa, Jayna, David, Barbara, Gram. You are always with me.

This work is supported by the National Science Foundation (NSF) SGER Award IIS-0555411, by the sponsors in the MIT Media Laboratory Things That Think and Digital Life consortia, and by Seagate. Any opinions, findings and conclusions or recommendations expressed in this thesis are those of the author and do not necessarily reflect the views of NSF or MIT.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

When you speak with someone closely involved in the life of a person with an autism spectrum condition (ASC), the issue that always comes up is difficulty with communication. This parent, caretaker, teacher, or other individual has trouble understanding the autistic[1] person in his or her life and also has trouble teaching this person the more commonly accepted and practiced forms of communication, including facial expressions, gestures, and a conversation partner's expression of emotional recognition.

As part of a larger body of work aimed at real-time assistance with reading faces in conversation, this thesis develops a video test to evaluate the understanding of real-life social cues by autistic individuals as compared to typically developing individuals. The development of this video test involves collecting videos in real-life situations, manually dividing the videos by social cue content, labeling the resulting video clips, and sorting the video clips by difficulty. These are all tasks we hope to automate in the future, using technologies such as the Mindreader software (see section 2.1.3). This automation would reduce the burden on caretakers who work with autistic individuals, allowing them to spend more time on developing a personal relationship with the individual and working on more complex issues instead of the tedious, repetitive tasks that are better left to a computer. In particular, this work is related to a method called social indexing in which a caretaker narrates the social cues in everyday life and suggests appropriate responses ("Look, Susie just fell down and now she is crying. Let's go over there and see if she is okay."). With real-time, automatic recognition of social cues and with the help of a caretaker to focus on appropriate social skills and responses for a particular individual, a wearable computer could prompt the autistic individual to respond appropriately to social cues as they occur in real life. These social cues could be saved to review at a later time in order to give the person space and time to understand and then to build recognition skills that are not dependent on the technology tool. The automation of producing labeled videos of social cues would allow the autistic individual to work on understanding complex social cues with a tool that is extremely consistent, systematic, logical, predictable, and inherently non-social – all characteristics that autistics are drawn to and which might simplify learning.

[1] For an explanation of our respectful use of the term "autistic" instead of "person with autism", please see Sinclair, J. (1993). "Don't mourn for us." Our Voice **1**(3).

The video test developed herein is part of a pre- and post- assessment that is to be used to test a technology tool that is the first step toward the automation described above. The video test was developed to measure the progress of social cue understanding over time. It breaks down complex interactions into pieces that are repeatable in smaller segments and presents these segments (the video clips) in association with a set of social cues. The videos recorded and used represent live, naturally occurring conversations in the lives of people with high-functioning autism. The video test contains clips to gauge a person's generalization of reading face and head gestures and specific abilities with faces of self, peers, familiar adults, strangers, and unknown actors. Through this video test, we hope to better understand the abilities of autistics to read others in natural situations and to find any trouble spots as individuals and as a group.

Throughout this thesis, I will be using the term "social cue" where others might sometimes use "emotion," "mental state," or "affect." This is a deliberate choice, as this thesis works exclusively with social facial expressions and head gestures that take place in the presence of direct communication with another person. "Affect" is a general term that encompasses many social cues, but is not exclusive to communication. The term "emotion" has a narrower definition that often excludes expressions such as thinking, which is less of a clear internal feeling while "mental state" excludes signals like acknowledgement.

This thesis begins with an overview of autism and some existing technology related to collecting real-life video and to recognizing and learning social cues. It then describes wearable cameras created for this thesis, including Self-Cam (a video system used to collect and process videos of one's own face in natural situations). Next, after describing the larger study to contextualize the video test, it talks about the video collection, division, and labeling and putting together the video test itself. This is followed by results obtained from administering the video test and the discussion, conclusions, and future work.

2 Background

2.1 Related Research

There is increasing evidence that autistics are comfortable using technology for learning (Moore 2000) and communication. In this thesis, I explore the use of technology to improve social-emotional understanding and test its use for high-functioning autistics. This thesis extends previous work on technologies for teaching social-emotional skills by bringing social-emotional learning into the natural environment. We use technology to capture the day-to-day social interactions and pull out social cues that may be overwhelming or may otherwise go unnoticed. This section describes work on wearable video camera systems and social sensors as well as two recent autism technologies.

2.1.1 Wearable and Affective Technologies

Other head-mounted cameras and personal life-recording devices have been created for a wide variety of applications. One example is a very robust face-centered camera developed for the television show, "Fear Factor." This system is made of a crash helmet with a camera on an adjustable goose-neck extension. The camera system appears to be made for capturing the facial expressions of the reality show's contestants as they perform fear-inducing tasks. Steve Mann developed wearable camera systems that create an augmented reality for the wearer (Mann 1997) (Starner, Mann, Rhodes et al. 1997). His cameras face out into the world and mediate the wearer's vision and experience via image processing by computers worn in backpacks and other places about the person. My Life Bits (Gemmell, Bell, Lueder et al. 2002) is a project conducted by Gordon Bell to store his entire life digitally. After converting all paper records to digital form, he has begun keeping records of as much live information as possible, including phone calls, IM transcripts, television, and radio. He intends to be able to do various audio, visual, and text searches of the events and interaction he has personally experienced. Sensecam (Hodges, Williams, Berry et al. 2006) is one tool used in the My Life Bits project. Sensecam is a wearable still camera with a fish-eye lens that autonomously captures pictures of the world from the user's position. A timer and a number of different sensors can be used to trigger the photos and to capture additional data centered on the user. Startlecam is a video recording system created by Jennifer Healey and Rosalind Picard (Healey and Picard 1998). The video from the camera is buffered so that a strong response from a sensor that detects the user's skin conductance (a startle response) will trigger the wearable computer to store the video during and immediately preceding the startle response. This video may capture the user's own facial expression during this time and/or the surrounding events, which may be the cause of the startle. LAFCam (Lockerd and Mueller 2002) is a system made by Andrea Lockerd and Floyd Mueller that was able to analyze the cameraperson's interest in what they are video taping by analyzing skin conductance and the subtle vocalizations that the cameraperson might make.

2.1.2 Mind Reading

The Mind Reading DVD (Baron-Cohen and Golan 2004) was developed to teach recognition of emotional expression to individuals on the autism spectrum through video and audio clips made by professional actors. The DVD consists of an extensive collection of acted video clips lasting 3-7 seconds each, which are based on Baron-Cohen's taxonomy of emotions. In addition to the clips, the DVD contains a set of games with still picture faces, lessons and quizzes with the videos, and a section of acted contextual situations. Golan and Baron-Cohen (Golan and Baron-Cohen 2006) evaluated the use of the DVD to teach a group of high-functioning autistics emotional and mental states and found that, while participants improved on recognizing expressions from the DVD, additional methods would be needed for generalization.

2.1.3 Mindreader

Mindreader (el Kaliouby 2005) (el Kaliouby and Robinson 2005) is a computer vision based system that attempts to recognize a person's affective and cognitive state based on face and head movements. Mindreader was originally developed by Rana el Kaliouby for her PhD thesis at Cambridge University, UK where she collaborated with Simon Baron-Cohen, co-Director of the Autism Research Centre at Cambridge, UK. Using a commercial facial tracker (Nevenvision - recently acquired by Google), the software takes in carefully framed video and tracks the movements of the face and head. The face features identified are largely based on Ekman and Friesen's Facial Action Coding System (Ekman and Friesen 1978), a characterization of the muscle movements of the face that map to particular facial expressions. The software maps the changes in facial features and gestures to the following set of labels: thinking, agreeing, disagreeing, concentrating, unsure, and interested (Figure 1). Mindreader is trained on the Mind Reading DVD described in section 2.1.2. In a preliminary test of the system's generalization accuracy in recognizing the six labels beyond the DVD training examples, a total of 18 participants (50.0% male, 50.0% female) between the ages of 19 and 28 chose the most applicable label for a set of videos also scored by Mindreader. The system performed as well as the top six percent of these people. It classified videos of expressions acted by participants in a computer-vision conference (el Kaliouby and Robinson 2005).

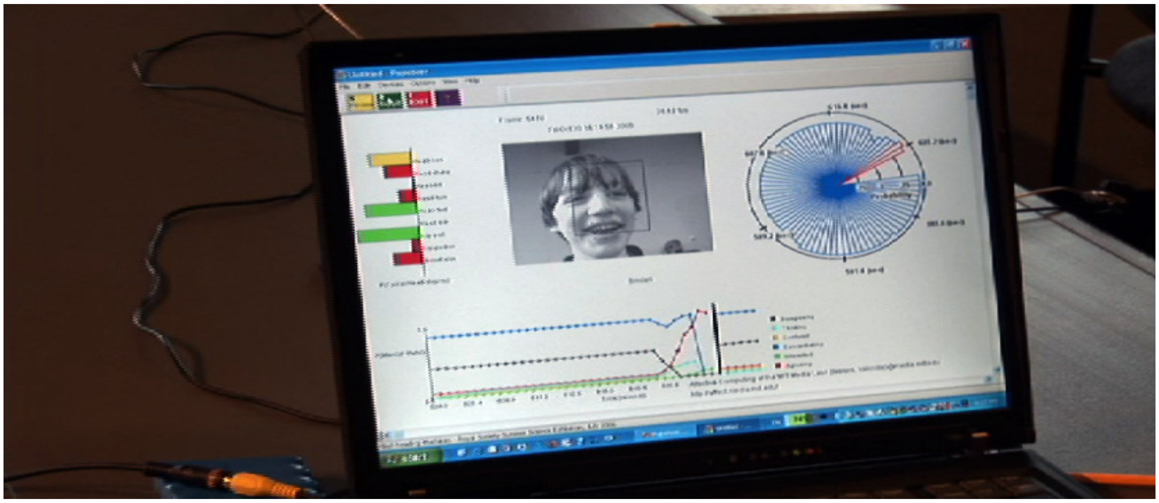


Figure 1: Mindreader Interface (el Kaliouby and Robinson 2005)

2.2 Autism

Autism is a neurodevelopmental condition that is specified by a combination of symptoms, including some or all of the following: qualitative impairment in communication skills, a qualitative impairment in social interaction skills, repetitive and stereotyped patterns of behavior, and delays or abnormal functioning in creative or imaginary play (American Psychiatric Association 2000). Autism is usually diagnosed in childhood, with onset of symptoms before the age of three. Although there are many treatments available for people with autism, there is no proven cause or cure. Autism is diagnosed behaviorally, and includes a large range of functionality, from those who cannot speak and rarely interact with people, to very high functioning people who can support themselves and participate independently in greater society. Autism is often co-diagnosed with a number of other mental disorders that affect a person's abilities and make it difficult to separate the causes and effects related to autism. In addition there is a related diagnosis of Asperger syndrome (AS), which is similar to autism in the difficulties in social interaction and unusual or repetitive patterns of interest and behavior, but without the delay in speech (American Psychiatric Association 2000). Asperger syndrome is often diagnosed in adolescence or adulthood.

This thesis describes work with young adults, ages 18-20, who are diagnosed with high functioning autism. These young adults are students at the Groden Center, which is a non-profit school in Providence, RI that provides community-based, evaluative, therapeutic, and educational programs for children and adults who have moderate to severe behavioral and emotional difficulties, including ASC. They are verbal and communicative and are working on developing their social skills and interaction through home, school, and/or focused individual or group sessions. In the development of this project, we also worked with a group of young adults, ages 14-21, who have Asperger syndrome (Figure 2). This group attends monthly social pragmatics sessions at the Groden Center.

Figure 2: Group with Asperger's syndrome at the Groden Center

3 Wearable Camera Systems

This section describes wearable and situated video camera systems that were developed and used in collection of videos for this thesis and future work. These camera systems are platforms for potential applications of this work, including feedback from automated social cue recognition, video segmentation and labeling, and video clip playback and review.

3.1 Self-Cam

Self-Cam is a tool for reflection upon one's own self-expression (Teeters, el Kaliouby and Picard 2006). The tool allows people to see themselves from an outside point of view while they are engaged in an interaction. Self-Cam also allows a wearer to reflect on his or her communications in relation to the source of feelings and emotions internally and to the response of interaction partners.

Self-Cam is a chest-mounted camera that is coupled to a digital recorder or laptop computer (Figure 3). The camera faces inward and centers on the wearer's face. Designed for use with the Mindreader software, Self-Cam uses an undistorted lens at arm's length from the face, supported by a thin wire, the combination weighing about 2.6 ounces (75 grams) in total. While visually awkward when used initially, Self-Cam is light and comfortable to wear in a research setting (Figure 4 and Figure 5).

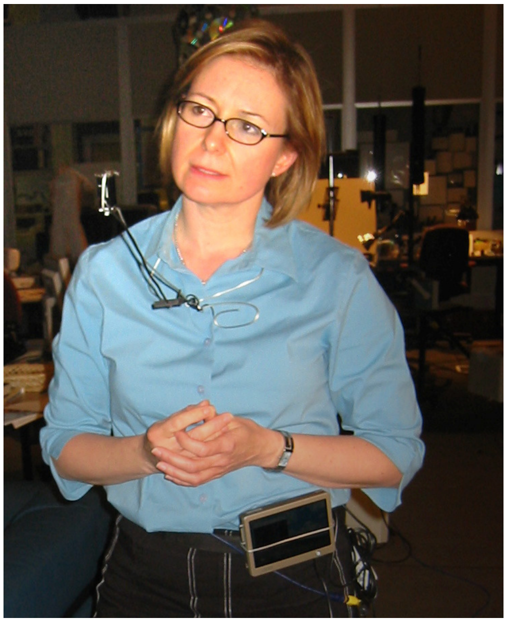


Figure 3: Rosalind Picard Wearing Self-Cam and Recording with the OQO 01+ Handtop Computer

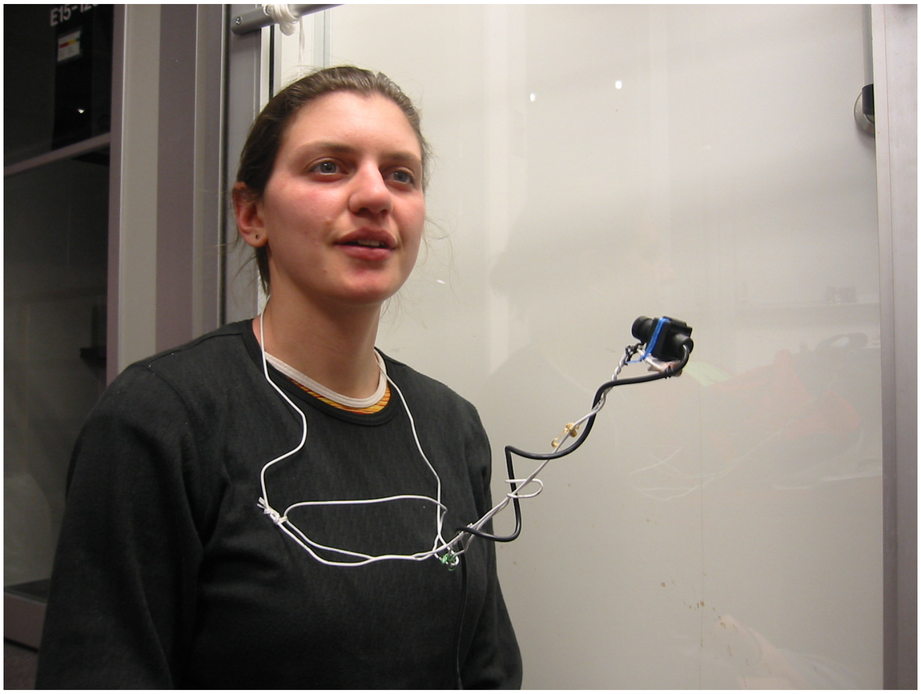


Figure 4: Author wearing prototype of Self-Cam

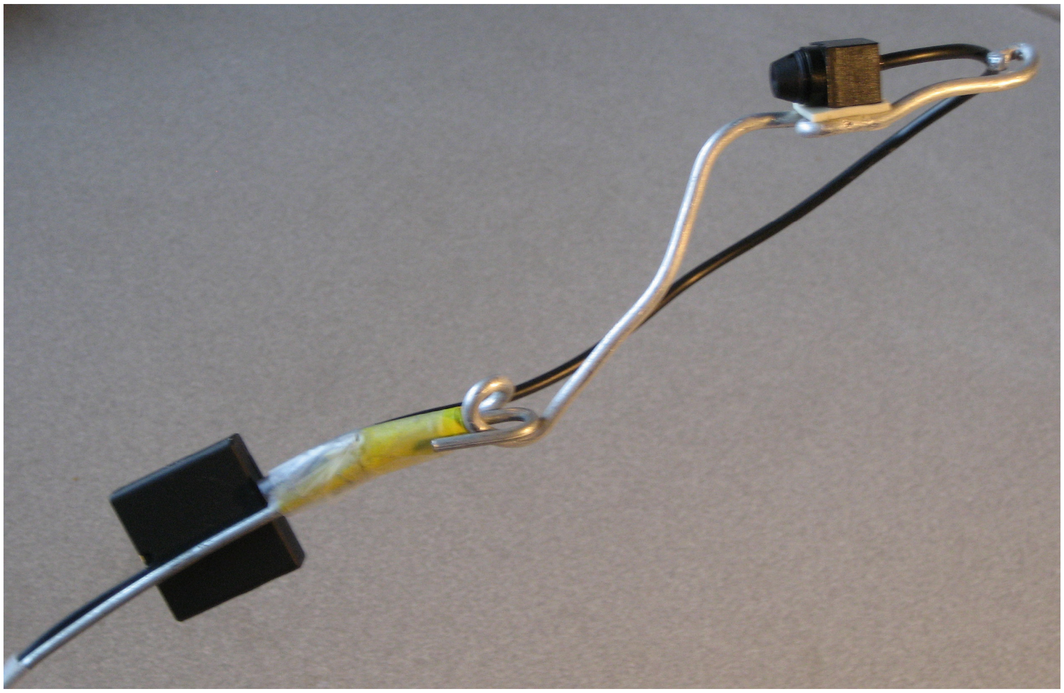


Figure 5: Closeup of Self-Cam with PC223XP Color CCD Micro Camera.

The first conception for a wearable camera system to assist individuals with autism in understanding the faces of others was the Emotional Hearing-Aid (el Kaliouby and Robinson 2005), which later developed into the Emotional-Social Intelligence Prosthesis. The Emotional-Social Intelligence Prosthesis combined the Mindreader software with an apparatus such as Hat-Cam (see Section 0). However, this presented two major difficulties: tracking and privacy. The current physical design of the Self-Cam system was developed in response to these problems as well as the basic requirements of wearability. First, the system needs to be portable. In order to function in a natural conversation, the system has to be flexible, light, and small enough to be worn while seated, standing, or walking around. Second, the Mindreader software has to function as reliably as possible. Mostly, this means a configuration that will allow the tracker to find and maintain its position on the face for enough time to obtain consistent mental state readings from the software. Additionally, the camera has to be positioned to obtain head movements for use in mental state recognition. Third, the system should maintain the privacy of any person who might appear in video recordings or analysis.

The physical system consists of soft 12 gauge aluminum wire that is bent into a form that rests on the chest (to capture head movements as well as facial movements). The camera is held away from the body but at a minimum distance from the wearer in order to simulate the view of the face by an interaction partner while creating as little intrusion as possible into the interaction. The image captured is of the person's own face because the tracking remains functional when it is grounded on the body that it is measuring (difficult-to-track movement is minimized). In addition, capturing and collecting data on one's own image allows the wearer to maintain control of their own data and minimizes the appearance of people who have not agreed to be captured in the video recordings or data collection.

The camera used is a PC223XP Color CCD Micro Camera, an analog camera that is .45 by .45 by .75 inches and 380 lines of resolution. It runs on a 12-volt rechargeable battery pack. The signal is digitalized at 30 frames per second using a KWorld Xpert DVD Maker USB 2.0 Video Capture Device. The camera is attached to a small, portable computer that can be carried on a waist holder or small backpack. Several devices are used in this capacity. An Archos digital video recorder is a waist-worn computer used to capture videos for later processing (Figure 6)

An OQO hand-top computer model 01+ (1GHz processor, running Windows XP) is used for capturing or processing the videos in real time, but the processing speed reduces accuracy of the face tracking and social cue recognition. For better performance, a 13-inch MacBook with a 2.16GHz Intel Core 2 Duo processor is used to capture and process the videos in real time. The use of additional sensors, such as a microphone input for the video recorder or a skin conductance sensor with a separate data storage and transfer device will also maintain the privacy and portability of the system through use of devices that will record data without identity information given off by non-wearers. In consideration of social awkwardness and wearability, the apparatus was made as small and light as possible. While not functional in all natural settings, this apparatus will serve to run experiments in a private school, classroom, home, or lab in order to assess the accuracy of the given hypotheses.

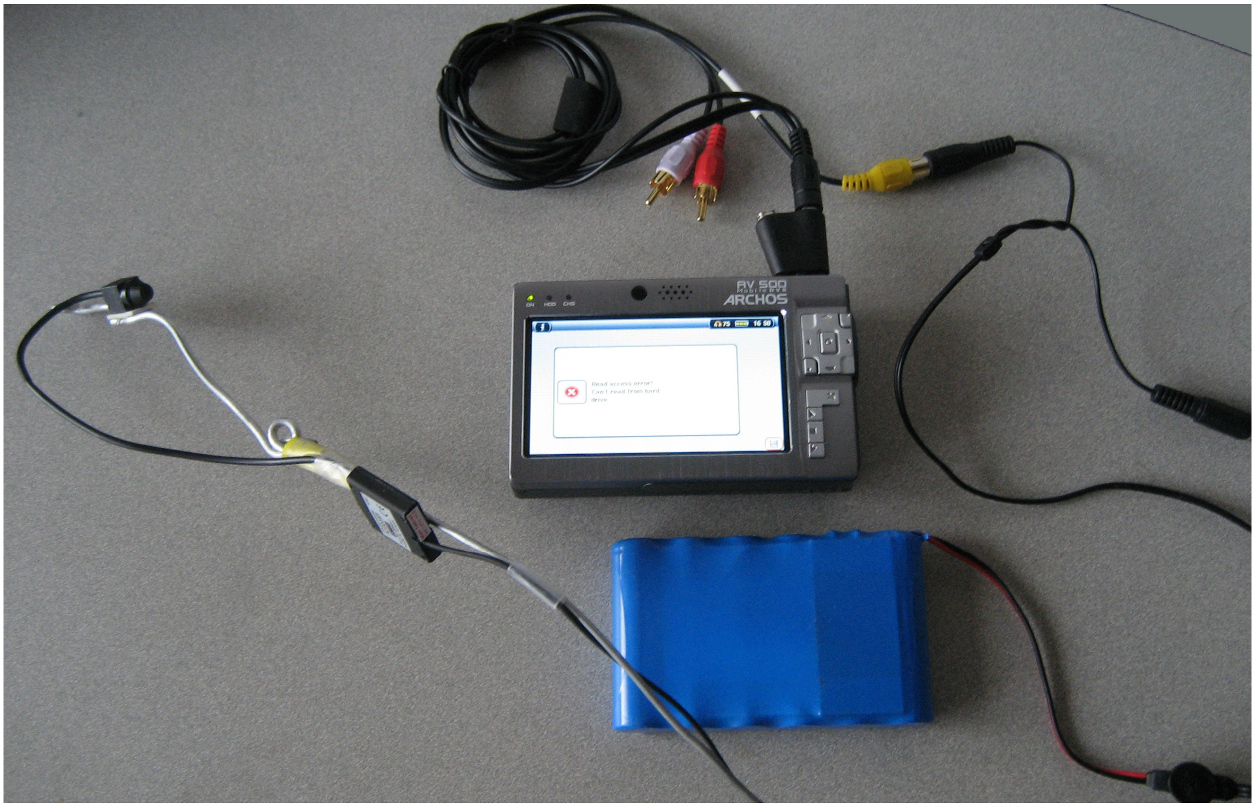


Figure 6: Self-Cam Setup with Archos Video Recorder

3.2 Multiple Self-Cams

Using multiple Self-Cams, we can view side-by-side, synchronized videos of two people in conversation. This allows an observer to find cases of mirroring, where one interaction partner imitates the gestures of the other. It also allows the observer to find any nonverbal signaling that has taken place between the two parties. The concept can be expanded to any number of conversational participants, such as a classroom of students interacting with a teacher or an audience interacting with a performer. In such cases where the interaction is one to many, the one has an opportunity to see the reaction of the entire crowd both individuals and as a group post performance (See Appendix 10.1.5).

3.3 Hat-Cam and Eye-Jacking

Hat-Cam is a small camera mounted facing outward on the brim of a baseball cap (Figure 7). While not used in the Emotional Social Prosthesis system for reasons stated above, the Hat-Cam becomes an interesting tool when used in conjunction with software where the human is the sensor. This is the case with Eye-Jacking. In Eye-Jacking, a remote party is given access to the live video from one or more people wearing the outward facing camera. This allows the remote person to see what is happening at the location of the wearer and also allows the person to see where the wearer is looking. As a remote viewer lacks the non-visual and peripheral-vision cues present in the local environment, the wearer allows this person to "jack in" to his or her visual field. Social cues seen by the wearer are brought into focus for the remote viewer. Having access to the social cues seen by the wearer allows the remote viewer to react appropriately. For instance, a person gesturing for a turn to speak while the remote viewer is speaking can be seen by the remote viewer when the wearer points the camera towards the person who is gesturing. The remote viewer now sees the person gesturing, then can choose to acknowledge the gesture and give the person the opportunity to speak. When used by multiple people in video conferencing, Eye-Jacking allows the remote party to follow the attention of an entire group of people. Local cues that convey a desire to speak, clarifying gestures, and disturbances may be communicated to the remote party by the wearers, simply by direction of gaze. As we have experienced in our own group meetings, seeing these social cues allows the remote party to participate in a meeting in a more socially appropriate manner.

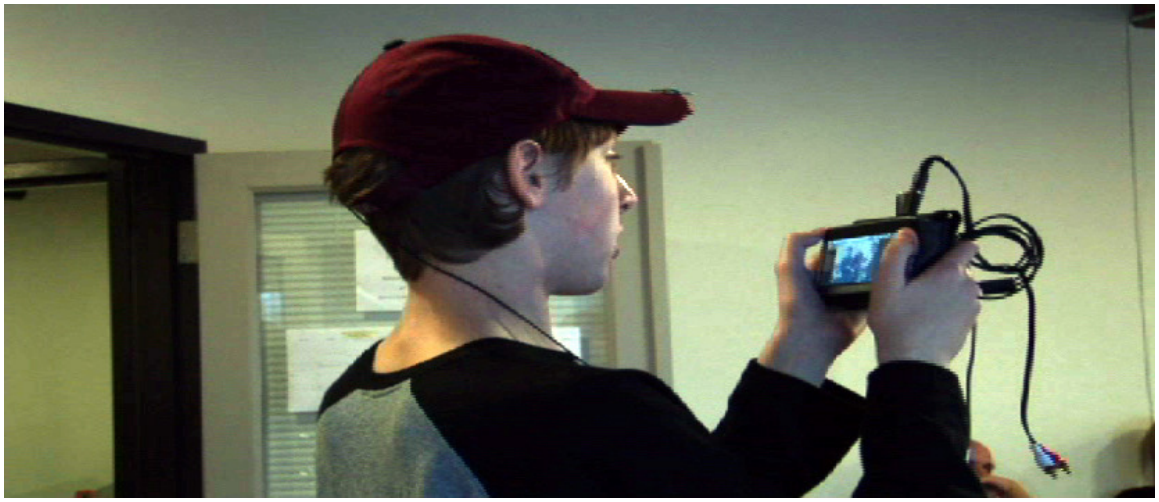


Figure 7: Young adult using Hat-Cam to record video at the Groden Center

3.4 Built-in Cameras

While not wearable, the cameras built in to modern laptops are small and portable. They sit at the appropriate distance to get a clear view of the face and are in a position to capture the image of a person working over long periods of time. Cameras like these are often used in video conferencing between remote colleagues and friends, where the output of your own video capture device has a significant presence on the computer screen.

Increasingly, cell phones have a camera next to the display to photograph the owner and a camera on the back to photograph what the owner is looking at. These cameras function much like a low resolution, handheld version of the equipment we are using in our research, but without the power to do real-time processing.

4 Piloting at the Groden Center

The Groden Center is a non-profit school if Providence, Rhode Island that provides community-based, evaluative, therapeutic, and educational programs for children and adults who have moderate to severe behavioral and emotional difficulties, including ASC. This section overviews the study at the Groden Center and details the pilot for which the video test in this thesis was developed. The video test is the pre- and post-assessment for the pilot and is implemented alongside the first stage of the pilot, which involves the same steps of video collection and processing. The subjects, the physical apparatus, the video development interfaces, and the trials and procedures for eliciting social cues from subjects are described in full.

4.1 The Study

The purpose of the study is to evaluate improvement of social cue recognition when training on naturally-situated videos of familiar faces compared with training on acted expressions of unfamiliar faces. The video test assesses generalization at three levels of familiarity: the participants' own face, familiar faces, and faces of strangers. The familiar faces include both autistic peers and neurotypical adults. The stranger videos include social cues from natural interactions as well as explicitly acted emotions.

The study participants are split into three groups: a neurotypical control group that is evaluated but does not participate in an intervention, an autism control group that works with videos from the Mind Reading DVD, and an autism group that participates in the Self-Cam/Mindreader intervention while reviewing videos of their own faces as well as the faces of their interaction partners. While we do not plan to run the intervention with the neurotypical group in this study, it would be interesting future work to compare their learning to that of the two groups of autistic students. Each of the three groups participates in a video collection phase where videos are taken of social interactions between participants and a Groden Center staff member. The videos are divided into social cue clips for the video test. These clips are then evaluated by a group of neurotypical adults for agreement on the expressed social cue. The three groups of study participants then go through pre-assessment, which includes the video test and other evaluations. The intervention groups will complete a 10-week intervention followed by a post-assessment that is identical to the pre-assessment (Figure 8). This thesis begins the pilot for the study described and consists of a video collection phase, designing a video test and administering a pre-assessment.

Group	Video Collection	Pre(T1)	Intervention	Post(T2)	Analysis: T2-T1
					(Hypotheses)
Group 1	Self-Cam	Social Responsiveness Scale	Self=Cam	Social Responsiveness Scale	(=/+)
n=10	+	Social Interaction Survey	+	Social Interaction Survey	(=/+)
|	Panels of	% accurate Self videos	Mindreader	% accurate Self videos	(+)
ASC Random Assignment	raters	% accurate Partner videos	software	% accurate Partner videos	(+)
(age-sex-IQ matched)		% accurate Others videos		% accurate Others videos	(?)
|					
Group 2	Self-Cam	Social Responsiveness Scale	Mind Reading	Social Responsiveness Scale	(=)
n=10	+	Social Interaction Survey	DVD	Social Interaction Survey	(=)
	Panels of	% accurate Self videos		% accurate Self videos	(?)
	raters	% accurate Partner videos		% accurate Partner videos	(?)
		% accurate Others videos		% accurate Others videos	(=)
Typically	Self-Cam	% accurate Self videos	N/A	N/A	>ASC Groups
Developing	+	% accurate Partner videos			
n=10	Panels of	% accurate Others videos			
	raters				

Figure 8: Groden Study Outline

4.2 The Pre- and Post- Assessment

The pre- and post- assessment measures the ability to recognize social cues with tasks of varying generalization levels and measures general social abilities, assessed before and after the intervention. The methods used in the assessment include a behavioral report called the Social Responsiveness Scale (SRS) (Constantino, Davis, Todd et al. 2003), direct observation and video analysis (interventions will be video-taped), self-report questionnaires, and the video test, which consists of video clips of participants, strangers, and actors from the Mind Reading DVD (Figure 15). The SRS is a survey that assesses the level of autism spectrum conditions in a child overall as well as assessing social awareness, social cognition, social communication, social motivation, and autistic mannerisms. The SRS is given by Groden Center staff who have been trained in formal assessment of autism characteristics.

4.3 Human Subjects

The subjects for the pilot are a group of six young adult males, ages 18-20, with high-functioning autism. All participate in the day program at the Groden Center. A summary of subject data is given in Figure 9.

Figure 10 gives subject scores on the five sections of the Social Responsiveness Scale (SRS), a test of social skills and behaviors that classifies an individual's range on the autism spectrum. A score of 76T or higher on the SRS is the severe range, 60T through 75T is mild to moderate, and a score of 59T or less is mild to not present. The TAS-20 is a test of alexithymia (scores shown in Figure 11), a condition described by three factors. Factor 1 is "difficulty identifying feelings," factor 2 is "difficulty describing feelings," and factor 3 is "externally-oriented thinking." A score of 61 or higher specifies high alexithymia while a score of 51 or lower is low alexithymia. Both the SRS and the TAS-20 were completed by trained staff members at the Groden Center who work closely with the subjects at school.

Subject Number	Age	Diagnosis	IQ (range of disability)	Overall SRS	Total TAS-20	Video Test Scores (%)
1	19	Autism Spectrum Disorder	Stanford-Binet Composite: 55 (Mild to Moderate)	74 (Mild to Moderate Autism)	76 (High Alexithymia)	63
3	20	Autism Spectrum Disorder; Down Syndrome	Stanford-Binet Composite: 65 (Mild)	53 (Mild Autism)	57 (Medium Alexithymia)	57
4	18	Autism Spectrum Disorder; Generalized Anxiety Disorder	Stanford-Binet Composite: 50 (Mild to Moderate)	56 (Mild Autism)	70 (High Alexithymia)	74
5	18	Autism Spectrum Disorder; ADHD	WISC-III - FSIQ: 42 (Moderate)	49 (Mild Autism)	76 (High Alexithymia)	82
6	19	Autism Spectrum Disorder; ADHD	Stanford-Binet Composite: 64 (Mild to Moderate)	52 (Mild Autism)	69 (High Alexithymia)	80

Figure 9: Subject Statistics

Subject Number	Overall T-Score	Social Awareness	Social Cognition	Social Communication	Social Motivation	Autistic Mannerisms
1	74T	67T	73T	71T	62T	81T
3	53T	52T	58T	53T	48T	53T
4	56T	57T	62T	55T	50T	55T
5	49T	56T	42T	48T	57T	49T
6	52T	52T	56T	51T	50T	53T

Figure 10: Social Responsiveness Scale Scores

Subject Number	TAS-20 Total	Factor 1	Factor 2	Factor 3
1	76	28	20	28
3	57	16	12	29
4	70	26	19	25
5	76	26	21	29
6	69	22	20	27

Figure 11: TAS-20 Alexithymia Scores

4.4 Apparatus

The recording system consists of a MacBook connected to an analog camera via an Analog-to-Digital converter (Figure 12). The rechargeable battery is pocket-sized and the camera is attached to the Self-Cam (described above). Loaded with Boot Camp (software that allows us to install Windows on a Macintosh: http://www.apple.com/macosx/bootcamp/), the MacBook runs WindowsXP and records video with applications including Windows Media Player, Mindreader, and PVR Plus. At each recording session, two participants sit across a desk from each other in conversation with their own recording system in front of them.

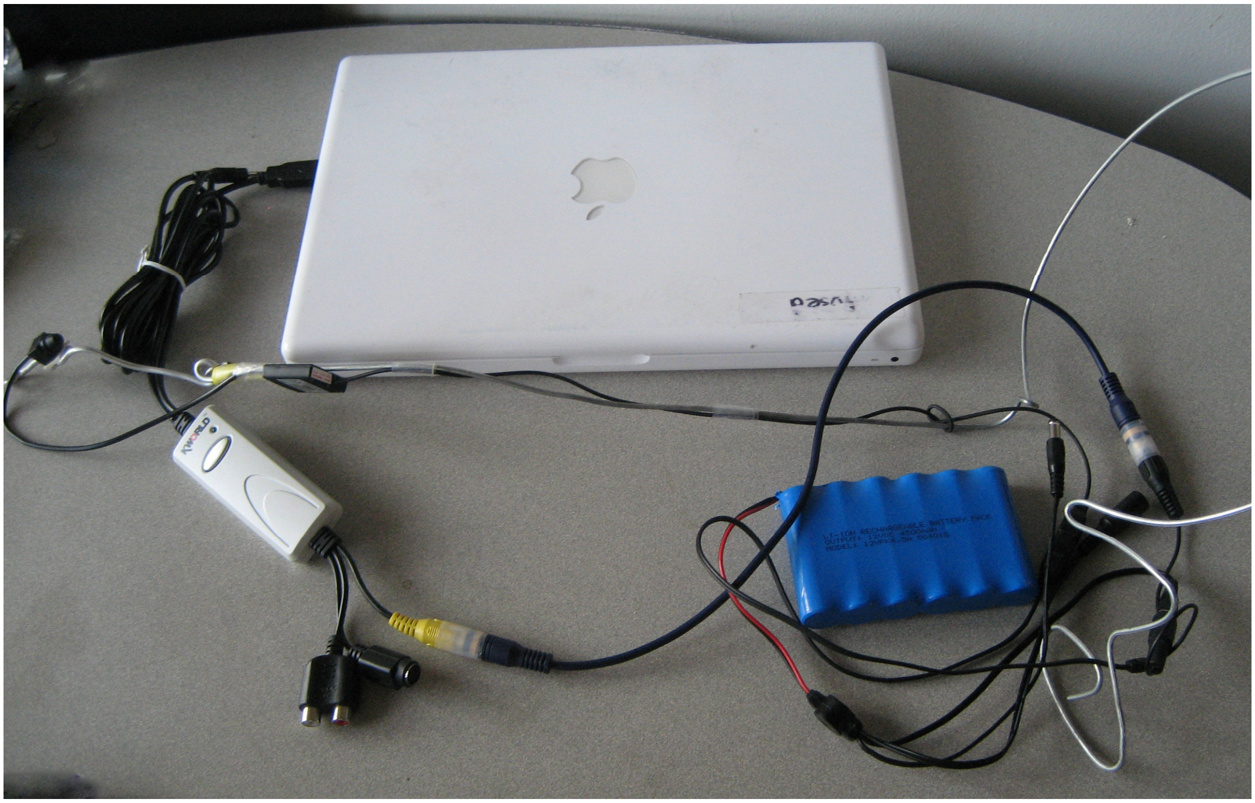
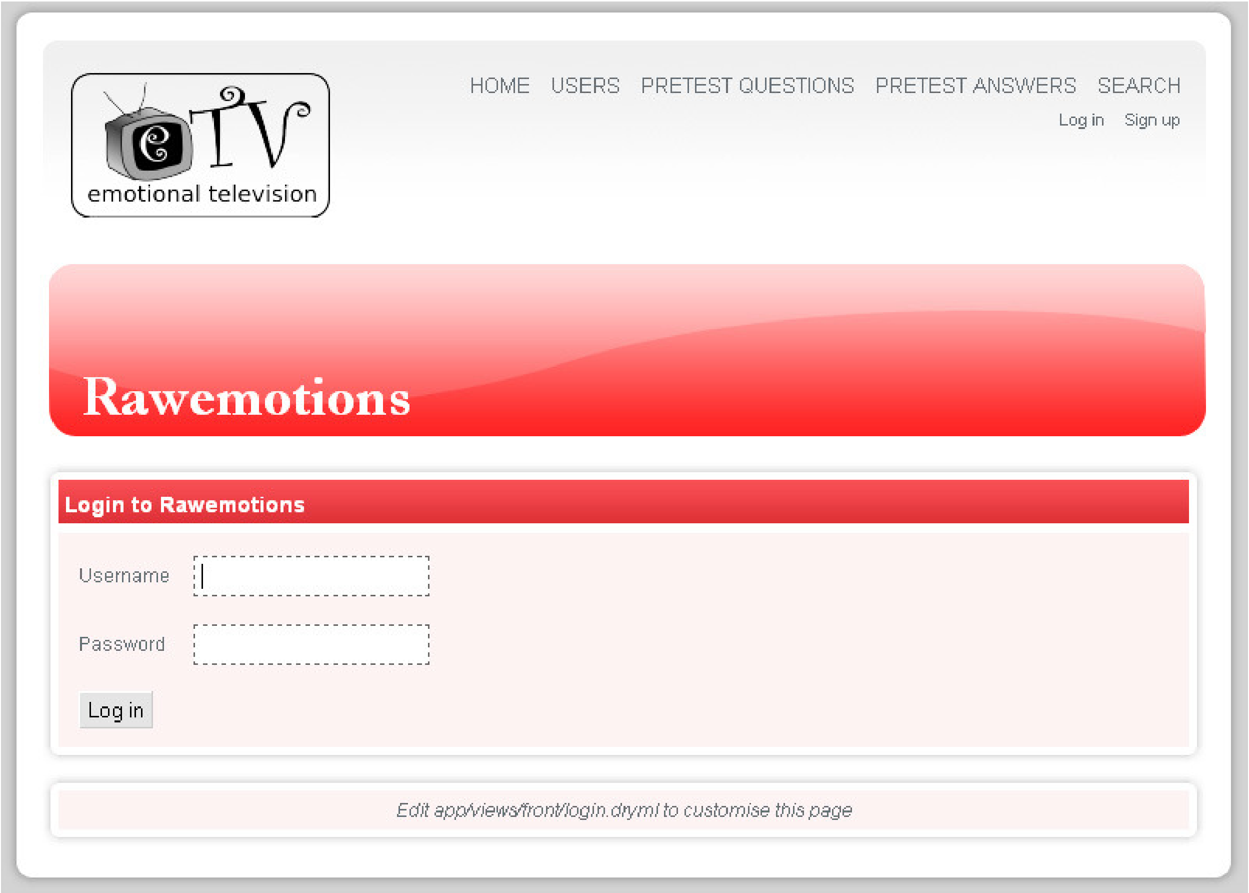


Figure 12: Self-Cam Setup with MacBook

The computer interface consists of three parts: the rating interface, the pre/post-test interface, and the intervention interface. The rating interface allows independent raters to give labels to the videos collected in the study. The video test interface plays pairs of videos to assess the understanding of social cues by subjects, raters, and controls. The intervention interface is designed to teach social cues as portrayed in videos and, in this study, will be used by the subjects. All of the interfaces are written in the Flex programming language and also have components written in the Ruby programming language and use a MySQL database. They are accessed through a single login interface (Figure 13).



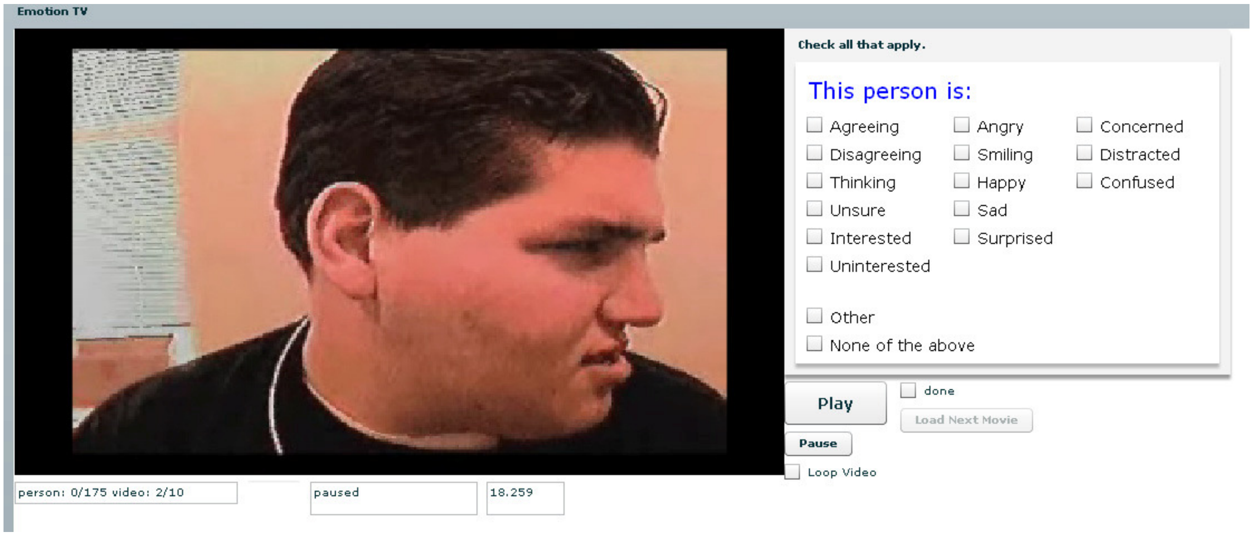
Rawemotions

Login to Rawemotions

Edit app/views/front/login.dryml to customise this page

Figure 13: Pre-test Login

In the rating interface (Figure 14), the raters are presented with video clips one at a time and in order by recording session. The raters can replay the clip as many times as they like. To the right of the video are 16 choices. This list was a constantly evolving set of labels that continued to change throughout the study, but were consistent for the full set of ratings. For those ratings, the labels were: agreeing, disagreeing, thinking, unsure, confused, interested, uninterested, angry, smiling, happy, sad, surprised, concerned, distracted, other, and none of the above. Agreeing, disagreeing, thinking, unsure, and interested are taken from the Mindreader software (el Kaliouby 2005). Angry, happy, sad, and surprised were added in order to include some basic emotions that might be present in the videos. These were chosen from Ekman and Friesen's six basic emotions that were found to be universal across cultures (Ekman and Friesen 1971). Disgusted and afraid were included briefly but removed because they were observed infrequently in the acquired videos, which were all taken of amiable conversation. Smiling was added as a confidence builder as many of the video clips appear difficult to read; including smiling not only introduces the identification of smiling but also the discernment of the more subtle social cue of smiling when not happy. Uninterested was added because it seemed to be present in the videos and it is a useful cue to understand in conversation. Concerned, distracted, and confused were suggested by several of the raters after a test run of the rating interface. None of the above is meant to be used as when the video doesn't seem to display any of the listed emotions, and the "other" option is for when there is a clear emotion displayed in the video but it is not on the list. The raters are asked to choose as many of the options as they see in the video. When they are finished with a video, they check a "done" box and click a button to go to the next video.



Check all that apply.

This person is:

☐ Agreeing
☐ Disagreeing
☐ Thinking
☐ Unsure
☐ Interested
☐ Uninterested

☐ Angry
☐ Smiling
☐ Happy
☐ Sad
☐ Surprised

☐ Concerned
☐ Distracted
☐ Confused

☐ Other
☐ None of the above

☐ done
☐ Loop Video

person: 0/175 video: 2/10

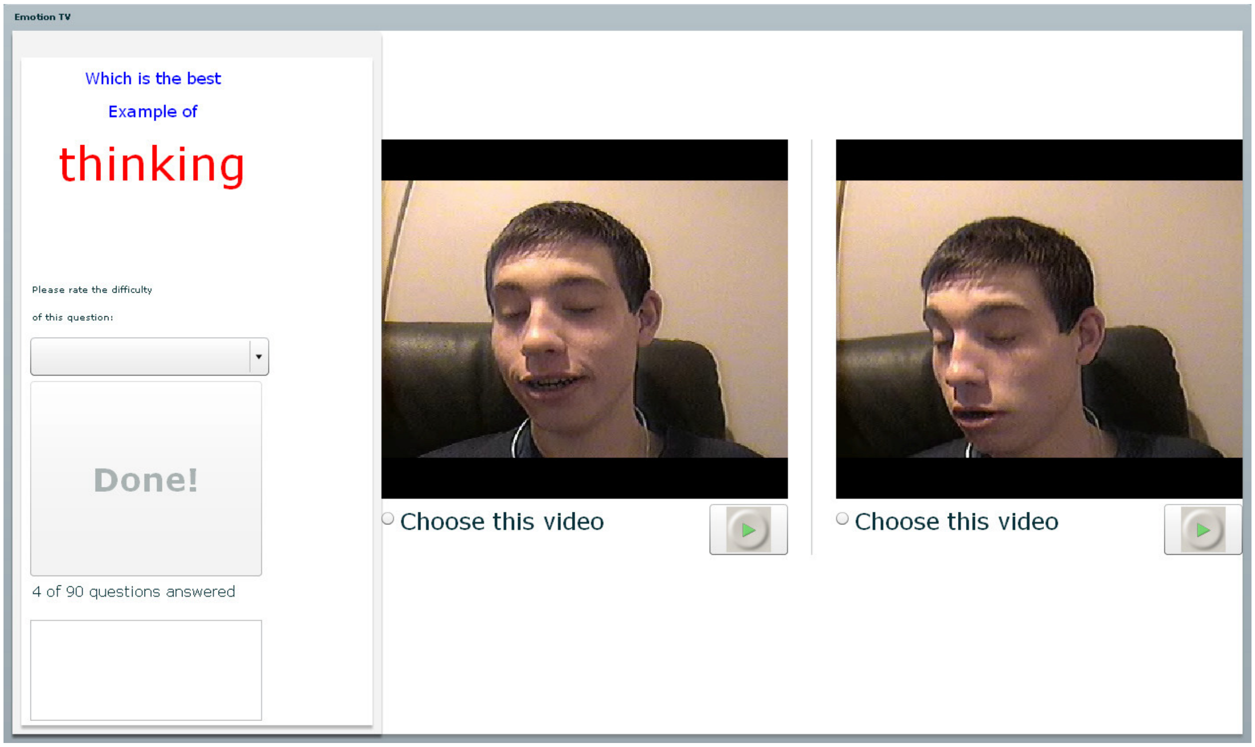
paused

18.259

Figure 14: Original Video Rating Interface (Used by Neurotypical Raters)

The video test interface (Figure 15) reverses these options, starting with a single social cue and asking the subject to choose the video that best fits the social cue. The subject is given two videos to choose from and the questions vary in difficulty.

The intervention interface is much like the rating interface, but it is designed for the autistic subjects. The subjects are given five social cue choices, instead of 16, and they are given immediate audio feedback when they choose the social cue(s) correctly or incorrectly.


Emotion TV

Which is the best
Example of
thinking

Please rate the difficulty
of this question:

4 of 90 questions answered

Figure 15: Video Test Interface

4.5 Eliciting Expressions Through Games and Theater

With help from Matt Lerner of the Spotlight Program (an after school program to help teens develop social skills through theater exercises), we piloted several interaction scenarios with the subjects at the Groden Center. These scenarios were an attempt to elicit conversational social cues to be recorded and used in the experiment. All of the practice scenarios were part of the recording sessions and included use of the Self-Cam or the MacBook camera. The games included Emotion Ball, Sausage, Mirroring, Thumbs up/down, and "Yes and...." In addition, we added some exercises including mental math, storytelling, and straight discussion. These games were used for 6-10 recording sessions, but were eventually dropped in favor of a more spontaneous conversation directed by the Groden Center Staff member (see section 4.6). The games are described below:

4.5.1 Emotion Ball

An object such as a ball or stuffed toy is passed back and forth between players to specify the turn to speak. The first person chooses an emotion, names it out loud, and displays that emotion on her face. The "ball" is then passed to the other person who says the emotion, and then shows how he would display that emotion. Once this iteration is complete, the receiver of the "ball" becomes the new sender and chooses a new emotion, names it out loud, displays it, and passes back to the first person.

4.5.2 Sausage

The object of this game is to get the "sausage sayer" to laugh. The "sausage sayer" is only allowed to say the word "sausage," but can vary other characteristics of speech such as intonation and speed. The other player(s) ask questions in hopes of triggering the humor of the "sausage sayer." Once the "sausage sayer" laughs, whoever asks the winning question becomes the new "sausage sayer" and play continues. In our pilots, this game was only somewhat successful and was limited to the subject being the "sausage sayer" and the Groden Center Staff asking questions.

4.5.3 Mirroring

In this game, there is a leader and a follower. The leader is free to move her face into various expressions and contortions while the follower plays along and copies those movements as exactly as possible.

4.5.4 Thumbs up/down

Before play begins, a subject is chosen which the speaker must discuss according to the indications of the non-speaker. The non-speaker has only her thumb to direct the discussion. An upturned thumb (thumbs up!) means the speaker must say only good things about the chosen subject. A down turned thumb (thumbs down!) means that the speaker must only say bad things about that subject.

4.5.5 Yes and...

In this game, each player must agree with the previous player, no matter how true or false the statement, and then add a new statement to the play. For example, the first player might say, "it's cloudy today." The second player would say, "yes it's cloudy today and the sky is purple!" The first person would continue play and say, "yes the sky is purple and...."

4.5.6 Mental Math

In this activity, the subject is asked to agree or disagree with a series of simple numerical statements. For example, "two plus three equals seven."

4.5.7 Storytelling

In this activity, the subject is the listener, and the Groden Center Staff member is the storyteller. The subject is unspecified, but the story is a semi-active attempt to elicit the facial expressions that indicated Mindreader's six mental states as well as Paul Ekman's six basic emotions.

4.5.8 Straight Discussion

Fairly self explanatory, straight discussion was unspecified conversation between the student and the Groden Center staff member. The discussion incorporated the events of the past week or the past day and was often directed toward topics of interest to the student such as music, the Red Sox, or other sports. The student was also given the opportunity to discuss anything else he had on his mind.

4.6 Eliciting Facial Expressions

The goal of developing these theater exercises and other activities is to elicit natural facial expressions - especially the six mental states that Mindreader is trained to recognize: thinking, interested, unsure, concentrating, agreeing, and disagreeing. The theater exercises as used in the Spotlight program, however, are more an intervention than elicitation. The exercises are intended to teach the skills of communication and many explicitly provoke facial expressions or concentrate on mental reasoning skills for social situations. The students at the Groden center are diagnosed with Autism and are, as a group, at a much lower functioning level than the Asperger's kids in the Spotlight program. Most of the Groden center students have a difficult time following the exercises as intended. Therefore, the exercises were simplified and did not employ the full effect of the complex skill development for which they were intended. We tried the exercises and various simplified versions for the different levels of understanding among the 6 subjects, recorded videos, and qualitatively assessed the results. It was determined that we should continue the exercises that were best understood by the students while trying out other possible exercises to evoke the simplest subset of emotions that could be consistently obtained. This would allow us to obtain the most natural expressions possible.

Since the recognition of agreement and disagreement were two of the most easily recognized expressions using the Mindreader software, two different exercises were used to provoke nodding and head shaking - the most obvious way to trigger agreement and disagreement in the software. The exercise we used first was mental math, as described above. For those who did not understand the math well or had difficulty maintaining attention, a true/false exercise was substituted, with simple statements such as "this computer is white."

After several sessions we stopped using the exercises in favor of a completely unscripted conversation between the Groden Center Staff and the subject. The games elicited very specific head gestures and stereotypical acted expressions as each response by the subject had to be explicitly provoked. The unscripted conversation helped the subject avoid thinking about his facial expression so that these social cues came out more naturally. Meanwhile, the staff member would try to tell stories or ask questions that would lead to the expression of the desired social cues in the subject.

5 Data Collection

This section details the pre- and post-assessment for the pilot described above for recognition of social cues and details how videos were recorded and processed to seed the video test. The subject, Groden Center Staff, and stranger videos were recorded, segmented, and labeled during this preparation.

5.1 Recording Subject Videos

Videos are filmed twice a week for several weeks to obtain clips of natural facial expressions for use in the experiment. During each biweekly session, 4-6 videos of separate subjects are recorded and each video lasted between 7 and 20 minutes. The subject sits for the length of his video plus the time to put on and remove the Self-Cam.

The video recording takes place two days a week near where the subjects' class is in session. One day a week this takes place in a small, private room (Figure 16). The second day, the class is in another building and the recording takes place in a larger, private room (Figure 17). The subjects are taken, one at a time, into the room to record a video. A subject walks into the room and sits down across from the Groden Center Staff member. For most of the subjects, this is Meredith Phelps, a research assistant who the subjects see everyday. Occasionally, Matthew Goodwin, the research director at the Groden Center, steps in for Meredith Phelps, particularly to talk with subject 6. The staff member is wearing the Self-Cam when the subject arrives. With assistance, the subject puts on the Self-Cam. Then, the video recording is started for both the subject and the staff member. The videos are started at approximately the same time, but are not precisely coordinated. Once the recording begins, the focus is taken off of the equipment and centered on the interaction.

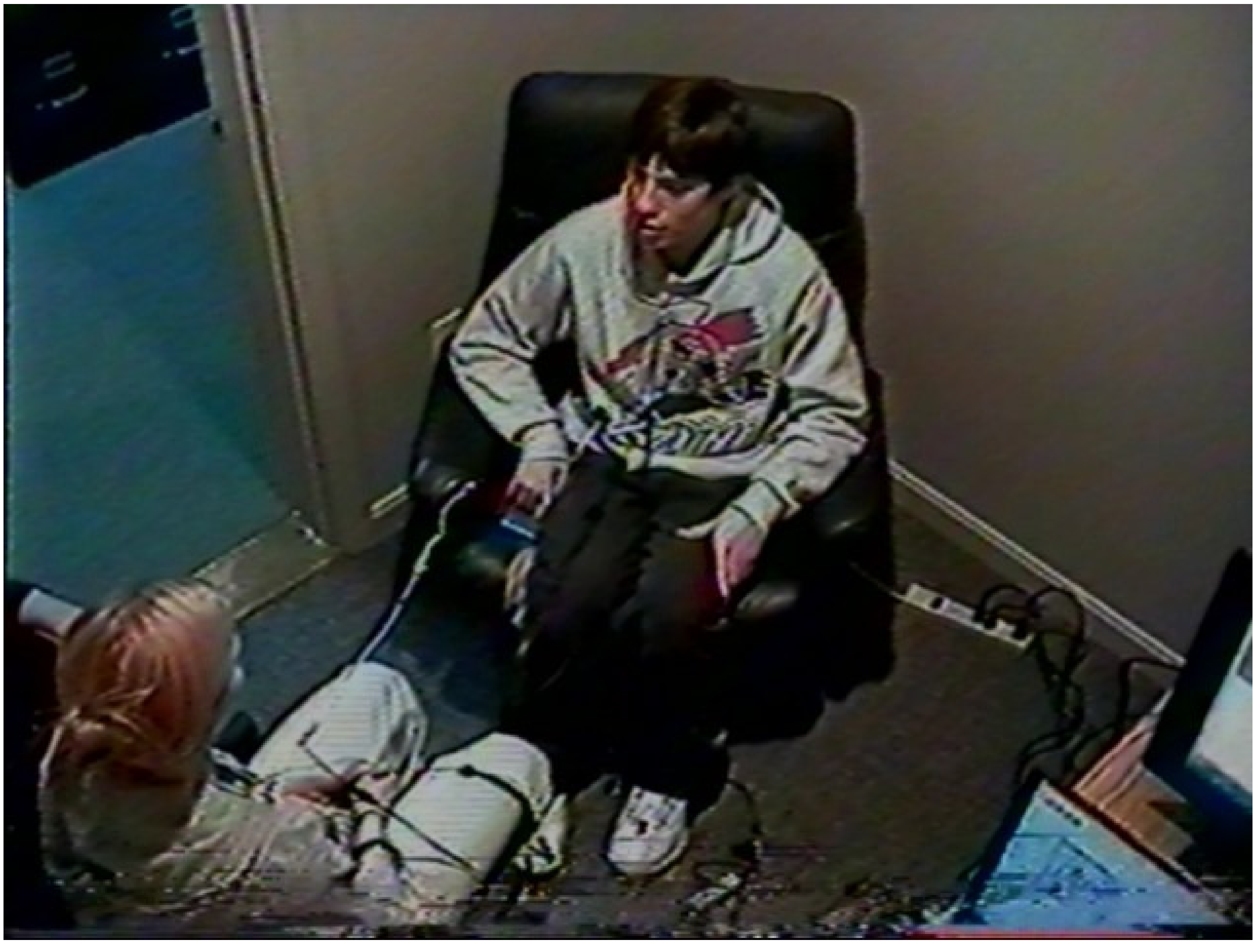


Figure 16: Recording Setup in Small Room at the Groden Center

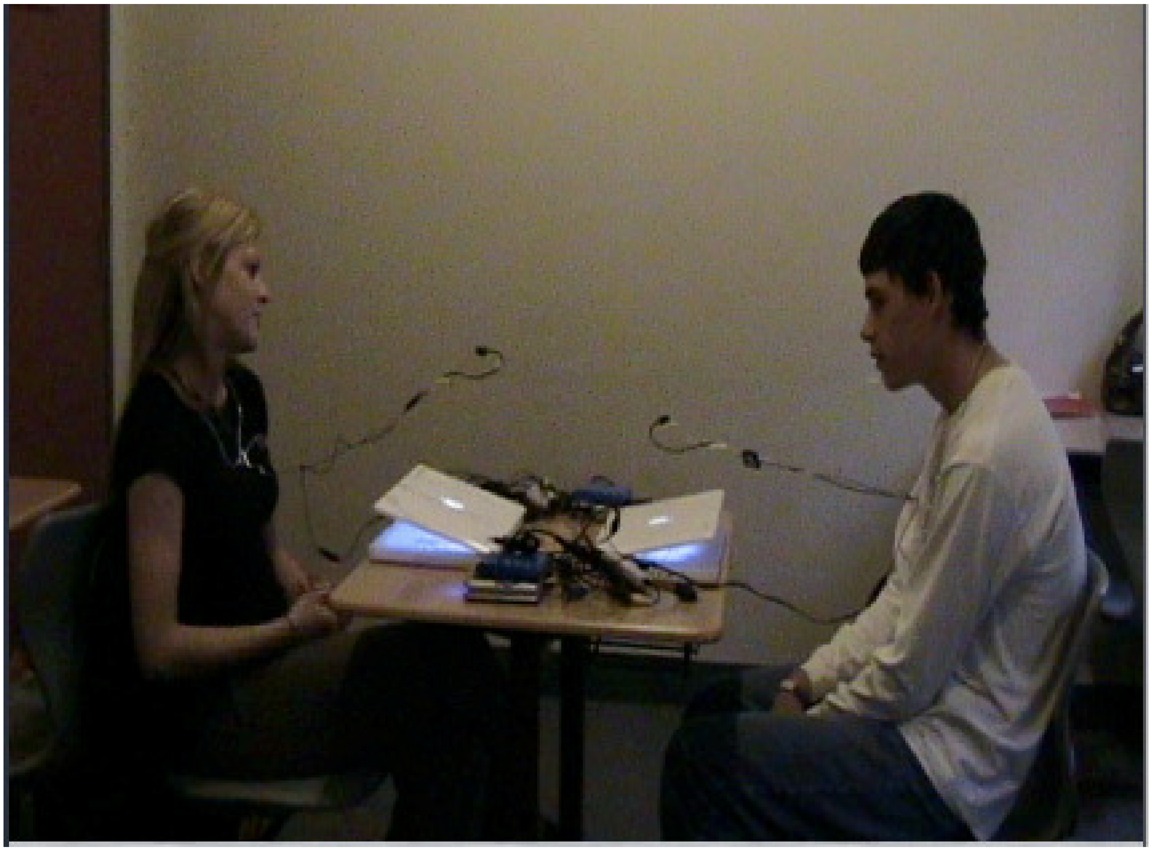


Figure 17: Recording Setup in Large Room at the Groden Center

During the interaction, the subject and staff member cover many topics such as sports, music, daily activities, class, friends, holidays, and weather. About half of the sessions start with improvisation-style games as described in section 4.5. They play these games before moving on to the unscripted conversation. The other half of the sessions consists of only unscripted conversation. During the conversation, the staff member often tells a story and spends much of the time asking questions to attempt to engage the subject and elicit a variety of emotions. In particular, the staff member tries to elicit six “emotions”: thinking, interested, confused, agreeing, disagreeing, and concentrating. The recording session ends either when the staff member has finished her story and gone through most of the topics listed above, when the subject becomes restless, or when time is limited and there are more subjects to see.

5.2 Recording Stranger Videos

Videos were taken of four people, ages 24-30, who are complete strangers to the subjects. The same methods were used as for the recording sessions at the Groden Center. However, the strangers did not do the theater exercises.

Stranger videos were taken for the video test in order to assess generalization of social cue recognition. The four strangers were volunteers from the MIT Media Lab who were available during the scheduled time slots. The apparatus used was identical to that used to film the subjects and Groden Center Staff members. The four strangers were split into two pairs. The first pair, both female, happened to be good friends and the other pair, both male, was a teaching assistant and his student. Each pair was asked to carry on a conversation of their choosing for a minimum of ten minutes and end at a natural ending point.

5.3 Video Segmentation

There is an inherent difficulty in collecting video expressions from those who have difficulty dealing with typical social conversations. Though the expressions may be more helpful and more significant to those same people, they happen less often than with a more socially functional person, if at all.

After working with theater exercises, math exercises, and other variations of expression elicitation activities, it was noticed that most of the nodding and what is recognized as "agreement" by the software is often something more like "acknowledgement." A nodding head is a social indicator that the conversational party is listening and is following the line of conversation. While culling the videos manually for expressions of "agreement," very few expressions were found while "acknowledgement" was abundant. This was also true for other social cues. Very little of the extracted cues and expressions represented emotions or internal mental states. Most of the facial movements corresponded to intentional communications and indications of states of attention and turn-taking.

The videos collected in the recording sessions were divided into short segments to isolate social cues. Video segmentation was performed in Final Cut Express and Adobe Premiere. Each segment was to be as long as possible while containing a single facial communication. The two people segmenting the videos earned scores of 33/36 on the eyes test (2001), a test of the ability to discern a person's state of mind from a picture showing a small rectangle around the eyes. While segmenting the videos, we discovered other important constraints. First, the playback software we were creating had a minimum practical video length of about three seconds. Second, the natural conversations were filled with overlapping facial expressions which made it very difficult to isolate a single emotional expression, social cue, or facial communication within the time constraints. In the end, we decided to relax the constraint on clarity in favor of a longer video (more facial context clues) and in favor of a more realistic set of natural facial expression clips. Each video then resulted in the extraction of ten to thirty-five clips of 3 to 15 seconds each. During the segmentation, any seemingly significant facial communications were noted by a starting and ending time in the video (Figure 18), but special attention was given to a set inspired by the Mindreader software and by prevalence in the collected videos: agreeing, disagreeing, thinking, unsure, interested, uninterested. The facial cues from the Mindreader software bias the social cues in both collection and segmentation. However, we included these six "mental states" because we intend to use the Mindreader software to automate the segmentation in the near future. Ekman's basic six emotions were also kept in mind (happy, sad, angry, frightened, surprised, and disgusted). The 3-20 second videos clips were inserted into the rating interface by their cue points to be verified by independent raters.

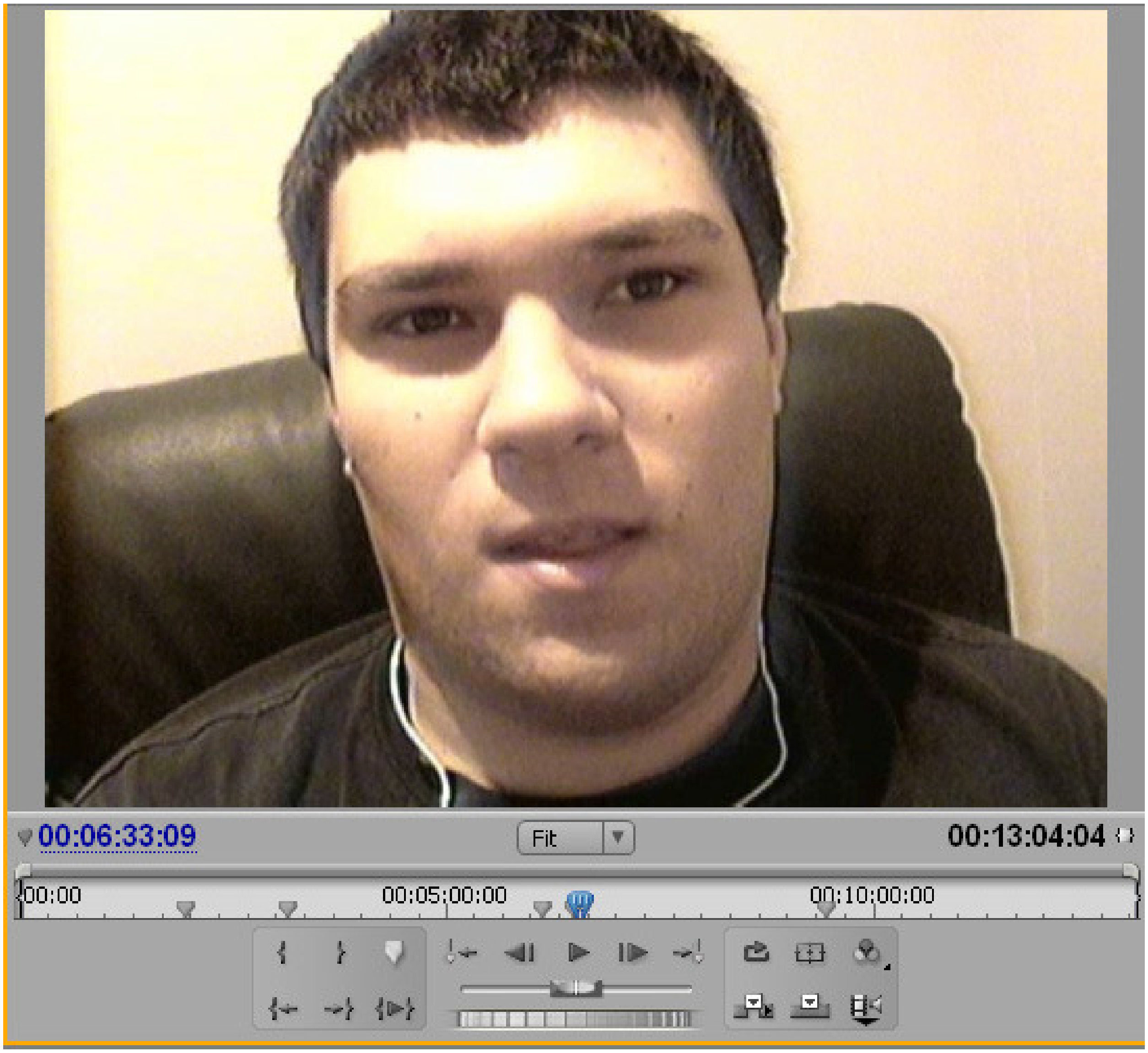


Figure 18: Snapshot of Segmented Video Clip in Adobe Premiere

5.4 Rating the Videos

The raters are a group of 10 people recruited by an ad on Craigslist (http://boston.craigslist.org). Raters were selected based on a good score on the Eyes Test (Baron-Cohen, Wheelwright, Hill et al. 2001). The top scorers were selected, with 28 points or higher out of 36, and are above average by gender. The top rater scored 35, and the average rater score was 32. Raters were recruited in a four day span and began immediately.

The rater experience takes place entirely remotely. The raters never need to come into the lab, and none of the raters are a part of the MIT or the Groden Center community. A rater experiences the study online, and most direct communication takes place through e-mail. Once chosen, the raters are asked to visit the experiment web site where they sign up for a username and password and then enter the rating interface. The rating takes place in sets. The first set consists of two fully segmented videos, a subject and the Groden Center staff member, for practice with the procedure of rating. The rater enters the site, watches each video as many times as she wants, then chooses all the labels that apply to that video clip, and finishes by loading the next video. The rater sees each video clip on the left half of the screen appearing with check boxes for the sixteen label choices on the right (see section 6). Below the video are indicators for the state of the video player and the location of the current video in the entire set of videos to be rated.

The first set of ratings not only give the raters a chance to become familiar with the rating process, but also allow us to test the content of the video clips compared to the choices given to the raters. Based on rater feedback, and the number of times a social cue was checked overall, the list of choices was modified and finalized, as described in section 4.4.

As each video rating is completed, it is stored to the database. Along with the clip number and each social cue checked, we store the date and time that the rating was submitted and the user number of the rater. These are used in the analysis of the ratings.

During the labeling of the video clips, the raters go through a small set of the videos twice. The first time is a practice run and the second time is part of the full set of video labeling. In this process, some of the practice labels were mixed up with the full set of labels. There were also some bugs in the rating interface that led to repeated ratings. This presented a small problem. For example, one video clip in the final smiling database is rated at 111% agreement, which reflects one more rating than unexpected (11 ratings divided by 10 expected ratings equals 111%). For video clips rated below 100% agreement, there was no reasonable way for us to find and remove extra ratings because of our study's time constraint. However, these initial ratings were only used to seed the video test and therefore slight errors in the statistics do not affect the final product. The video test questions were subsequently reviewed and changed if needed (see section 6).

6 Developing the Video Test for Recognition of Social Cues

In developing the video test, the goal was to create a concise evaluation tool that would provide a baseline for measuring improvement in processing of social-emotional information from face-head movements. This would give us an idea of the test takers' face-reading abilities based on their familiarity with the person they are watching. This section describes how the video test was seeded and how changes were made to create the final product. Video choice was a semi-subjective process, but did not interfere with the raters' independence, The test that was taken by subjects, raters, and "controls" (see section 7).

The video test utilizes a subset of all video clips, initially based on the ratings given by the independent raters. Each video test question contains two video clips showing different levels of expression of a social cue (Figure 15). While multiple social cues may be present in a single clip, the subject is asked to choose the clip with the strongest expression of one of five labels: thinking, interested, unsure, happy, and smiling. This is different than the rating interface. Instead of choosing labels for a single video, the subjects are asked to choose the video that is the "best example" of a single label. This was done for three reasons. First reason was to simplify the interface because the subjects have a hard time reading and choosing among multiple labels. Second reason was because raters seemed to miss some labels that were not immediately obvious (described below). Third, this interface allows us to differentiate between social cues. By asking, "which clip is the strongest cue?" we can declare a "correct" answer based on the ratings. When we ask, "What social cues are present?" even the smallest facial movements apply, so there are multiple correct answers. It is difficult to declare an answer "wrong."

A total of 90 questions were created including 10 questions from each of nine categories (referred to as subject types): each of the six individual subjects, the Groden Center Staff member, people unfamiliar to the subjects (strangers), and the Mind Reading DVD. Each subject sees all 30 questions depicting the staff member, the strangers, and the DVD actors. From the remaining 60 questions, each subject sees 10 of himself and 50 of his peers. The questions and the placement of the videos in each question are re-ordered randomly each time the test is taken.

The five labels chosen for the video test (interested, thinking, unsure, happy, and smiling) are a subset of the 16 choices given to the independent raters. These were the five labels that were most chosen and most agreed upon among the raters.

The answers that would be redefined until they eventually became the “right” answers on the 90 video test questions are called “correct” clips. Two “correct” clips were chosen for each social cue and subject type. These “correct” video clips were initially chosen by taking the top two most-agreed-upon clips as labeled by the raters (agreement refers to the percentage of raters who chose the social cue in question for any clip). When agreement was below 80%, the clip was checked for obvious presence of the labeled expression. More care was taken in choosing the distracters for these “correct” clips, described below.

We decided that each question should show the same person in all four video clips because facial expressions can vary greatly from person to person. For example, subject 6 might look happy when compared to another video clip of himself, but may look neutral or even negative when compared to someone with a big smile, as his expressions are very subtle. Upon later reflection, after the video test was administered, we thought that the inclusion of multiple people in a question might actually be helpful. Using two different faces would give an advantage to people familiar with those faces and this might allow us to better test the abilities of autistic subjects on faces that had different levels of familiarity. We did not have time to test this hypothesis.

After watching the “correct” clips below 80% agreement, we designed what we thought was a reasonable method for choosing distracters. Initially, we designed each question to have four (later two) video clips – a “correct” clip and three distracters (later just one). We would design one hard question and one easy question for each subject type and label. If the main clip is 80% agreement or above, a hard question would have two distracters around 40% agreement and one distracter below 30% agreement. An easy question would include two distracters with less than 18% agreement and one distracter around 20% agreement. Using this method, we designed one question for each of the stranger, DVD, staff member, and subject videos. We then went through each question looking at the video clips and revised our method because the questions seemed difficult.

We decided to base our methods on the agreement scheme of the DVD clips because there was a limited number of DVD clips in which all four videos could be of the same person. Going through the DVD clips, we found a workable agreement scheme for "correct" clips over 80% agreement in which a hard question's distracters were of 30%, 20%, less than 18% agreement and an easy question's distracters were of 30% and 2 less than 18% agreement. For the "correct" clips below 80% agreement, all distracters were less than 18% agreement. The DVD clips' agreement scheme served as the model for choosing distracters for all of the test questions. At this point, we showed the test to a couple of colleagues who watched and critiqued the four clips for a question side-by-side. Based on their input, we decided the test was still too difficult. Except for the DVD questions, we replaced all of the distracters above 20% agreement with distracters below 18% agreement. For the DVD questions, we went through manually and changed any we thought were still too difficult. Most distracters for these questions have agreements of 30% and 2 less than 18% (Table 1).

Reviewing the video test with colleagues, we decided to simplify further to make the test easier to complete and faster to take. We eliminated the distracters with the higher percentages, leaving only two video clips per question: the "correct" clip and a single distracter. Still, there are several clips that seem to have inconsistent ratings. Colleagues consistently thought the test was too difficult, so we continued to replace distracters and "correct" clips, often with clips that seemed inconsistent with their ratings (see discussion below). These video clips were manually viewed and chosen using personal interpretation by the same two people who segmented the videos (and scored well on the Eyes Test). The video clips with no ratings were segmented from previously unsegmented videos during the development of the video test when there were not enough examples of a particular social cue for a subject and after the ratings were otherwise completed. In particular, subject 6 had no happy and few smiling clips. These are the clips with seemingly inconsistent ratings:

Questions 11, 29, 36, and 45 with a "correct" clip at 33%, 33%, 43%, and 22% agreement respectively and question 36 with a distracter at 33% agreement (10 points below the correct clip), were chosen manually and were answered correctly by 12 of the 13 people taking the video test (Table 1). Questions 54, 64, and 72 had "correct" clips with no rating (segmented after the rating process). These were answered correctly by 10 of 13, 13 of 13, and 11 of 13 people, respectively. Questions 65 and 85 had "correct" clips at 33% and 43% agreement, were chosen manually, and were answered correctly by 13 of 13 and 8 of 13 people respectively. Only question 89, manually chosen with a "correct" clip at 33% agreement and a distracter at 29% seems questionable as only 6 of 13 people answered correctly, including 3 of 5 subjects, 1 of 4 controls, and 2 of 4 raters.

Question	Social Cue	"Correct" Clip	% Agreement	Distracter Clip	% Agreement
1	Interested	678	87	706	13
2	Interested	1666	100	1297	0
3	Interested	947	100	857	14
4	Interested	583	87	1772	17
5	Interested	281	89	603	14
6	Interested	426	89	1789	17
7	Interested	1914	100	619	0
8	Interested	653	85	1000	14
9	Interested	676	71	655	14
10	Interested	742	87	761	12
11	Interested	1274	33	1293	0
12	Interested	387	89	819	14
13	Interested	1781	83	1759	17
14	Interested	297	89	608	14
15	Interested	1019	71	1796	17
16	Interested	1903	100	1886	0
17	Interested	324	89	1930	20
18	Interested	666	71	1829	17
19	Thinking	723	100	728	13
20	Thinking	1688	67	1258	17
21	Thinking	389	100	849	14
22	Thinking	593	86	1776	17
23	Thinking	299	100	605	14
24	Thinking	1794	100	1799	17
25	Thinking	1886	100	1014	14
26	Thinking	331	100	1004	14
27	Thinking	672	86	1842	17
28	Thinking	778	100	771	13
29	Thinking	1300	33	1646	0
30	Thinking	970	100	804	14
31	Thinking	1786	83	592	14
32	Thinking	960	86	604	14
33	Thinking	1814	83	1810	17
34	Thinking	619	86	640	14
35	Thinking	648	71	646	14
36	Thinking	669	43	334	33
37	Smiling	686	100	710	13
38	Smiling	1652	100	1692	0
39	Smiling	804	100	351	11
40	Smiling	1775	100	584	14
41	Smiling	297	100	616	0
42	Smiling	1812	100	1796	17
43	Smiling	312	111	630	14
44	Smiling	321	100	646	14
45	Smiling	334	22	339	0
46	Smiling	712	100	721	13
47	Smiling	1637	50	1316	0
48	Smiling	370	100	814	14
49	Smiling	1776	100	580	14
50	Smiling	278	90	288	11
51	Smiling	1804	100	1803	17
52	Smiling	622	86	1007	14
53	Smiling	1001	100	994	14
54	Smiling	2092		1835	0
55	Happy	746	100	761	13

56	Happy	1653	33	1608	0
57	Happy	390	100	811	14
58	Happy	913	100	591	14
59	Happy	295	100	290	11
60	Happy	1019	86	1016	14
61	Happy	1862	100	308	11
62	Happy	644	71	645	14
63	Happy	2090		345	0
64	Happy	753	100	763	0
65	Happy	1650	33	1660	0
66	Happy	386	100	931	14
67	Happy	590	86	1782	17
68	Happy	292	100	283	11
69	Happy	1802	83	1016	14
70	Happy	1014	100	303	0
71	Happy	653	71	324	11
72	Happy	2091		668	0
73	Unsure	749	88	758	13
74	Unsure	1664	50	1321	17
75	Unsure	365	100	355	11
76	Unsure	1786	50	579	14
77	Unsure	616	57	296	11
78	Unsure	1794	84	1811	0
79	Unsure	1920	60	314	11
80	Unsure	333	44	327	11
81	Unsure	663	57	345	11
82	Unsure	717	87	719	13
83	Unsure	1663	50	1259	0
84	Unsure	358	89	398	14
85	Unsure	584	43	1765	17
86	Unsure	596	57	617	14
87	Unsure	430	78	1798	17
88	Unsure	311	56	637	14
89	Unsure	330	33	648	29
90	Unsure	340	57	1836	17

Table 1: Percentage Agreements for Pre-test Questions

One crucial difference between the Mind Reader DVD videos and the recorded videos is that the DVD videos utilize actors who are focused on a single emotion while the recorded videos include overlapping emotions and social cues (the harder challenge that people face in everyday life). The latter is a result of filming unconstrained conversation in a natural context. Within our rating interface, the raters tended to choose 1 – 3 labels that were most clear in the video clip. Therefore, the ratings of DVD clips were usually complete while the ratings of the recorded videos often left out some of the more subtle and complex overlapping social cues. An explanation for this may be that the raters had two different methods of rating. While some raters scanned through all social cues and asked “Is this present in the clip?” (the “Exhaustive Method”) most raters seemed to look for a right answer by viewing the clip, making a quick judgment of what social cue was strongest, and then only marking those cues that best matched that judgment (the “Intuitive Method”) Consequently, weaker cues in the video clip were not marked. For example, in clip 1810, the person is smiling and seems happy, but is clearly interested as well. The raters immediately registered happy and smiling and did not look at the rest of the social cues to see if they were present. Therefore, interested had low agreement for this particular clip. Additionally, the “correct” clip agreements for unsure are very low, but the video clips when viewed seem as if they should be highly agreed upon. It is possible that unsure was a secondary social cue in our video set, so the raters chose other social cues before considering unsure and only considered unsure if they followed the “Exhaustive Method.”

The video test contains 20 video clips that are used twice. Five of these video clips are used as “correct” video clips twice. None are used for the same emotion more than once, nor are they used as both smiling and happy. We decided that a video clip should not be use for both smiling and happy because smiling is a facial gesture that often implies happy. One video clip is used for both interested and smiling, two are used for both happy and interested, and two are used for both thinking an unsure. Ten repeated video clips are “correct” clips used as distracters for other questions. The remaining five video clips are repeated as distracter in two different questions.

7 Video Test Results and Discussion

The video test was taken by three groups of people: four raters who had previously labeled the larger set of video clips, four people recruited at MIT who had not seen the videos, and five of the six study participants (subject 2 was no longer available). We will call the group from MIT "controls," but note that this is not the same as the control group who will participate in the full study.

Overall, the average score of both the raters and the controls is 91% while the subjects scored an average of 71% (Figure 19). Looking at each individual (Figure 20) we find that the subjects all scored lower than any one rater or control - ranging from 57% to 82% - and the raters sandwiched the control scores - 90% to 92% - with scores of 89, 90, 93, and 94%.

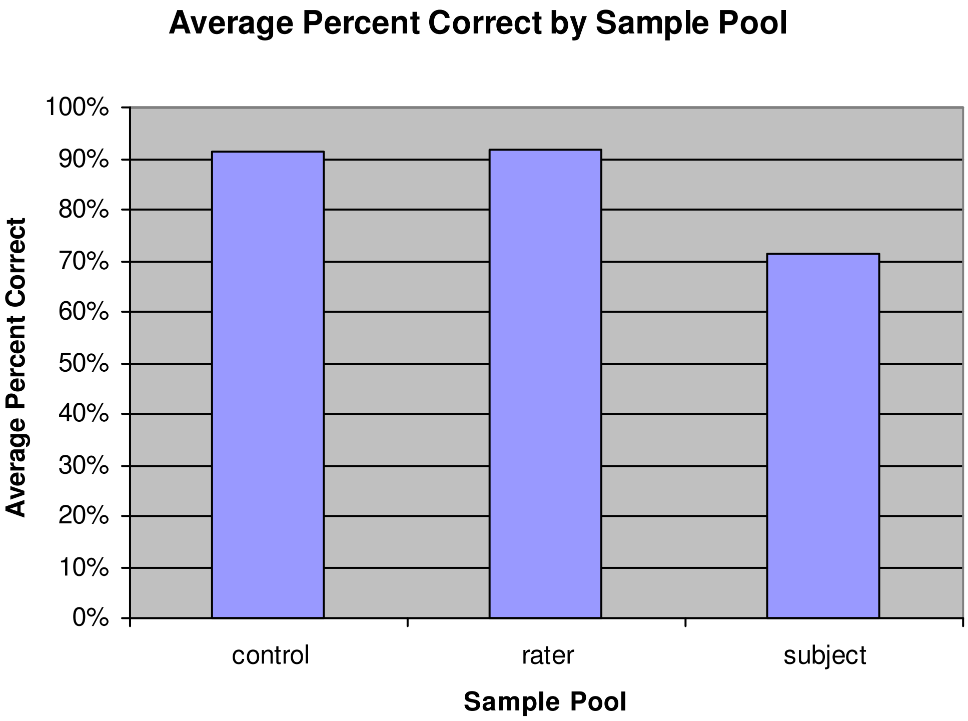


Figure 19: Average Percent Correct by Group

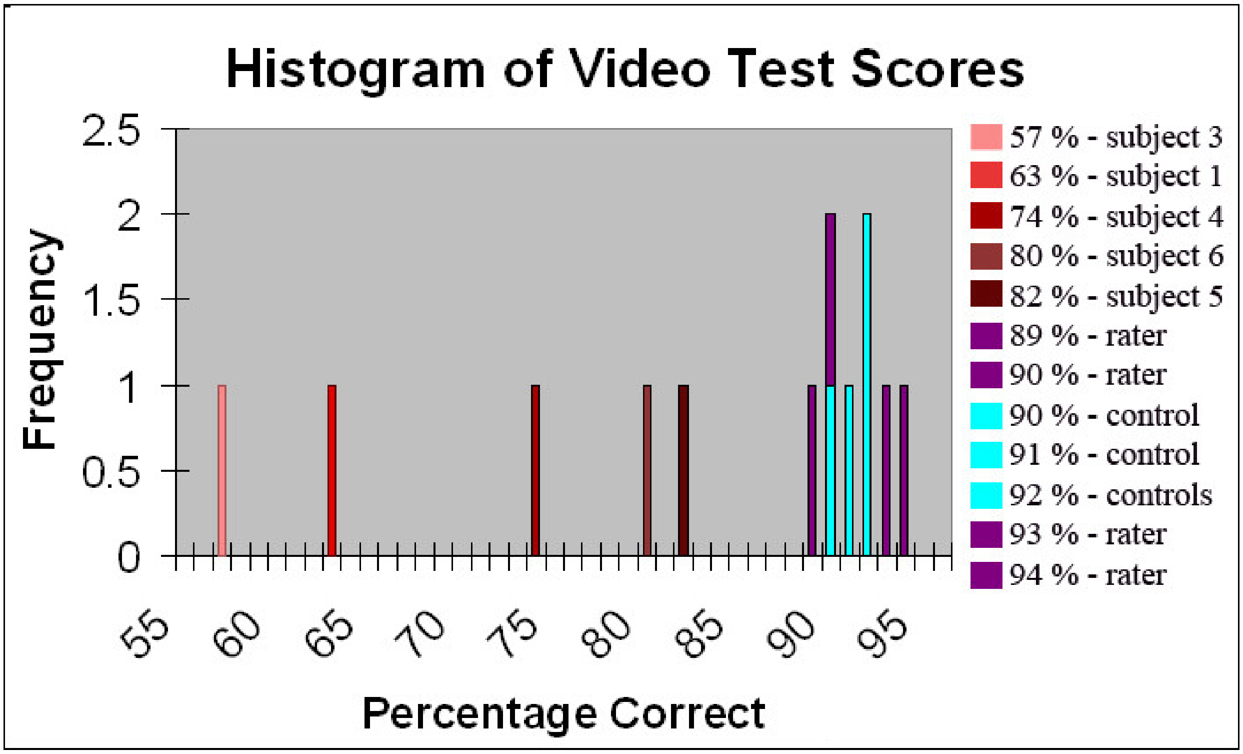


Figure 20: Histogram of Video Test Scores

Breaking down the scores by question (Figure 21) and social cue (Figure 22, Figure 23, Figure 24, Figure 25, Figure 26), we can see clusters of incorrect answers that may point out flaws in the test: Where raters and controls both scored near 50%, the expected average for random guessing, the question may have been overly ambiguous or represented two interpretations of a social cue. (note: the missing answer that leaves a shorter column in Figure 21 and Figure 24 is due to a software glitch that prevented one rater from answering that video test question). For example, question 20 has one example that shows classical, stereotypical thinking (the person tilts his head and looks up while pursing his lips) followed by a headshake (the Mind Reader DVD classifies this as disinclined) while the second example (classified in Mind Reading as fascinated) shows a natural style of thinking that is more interactive. While the classical thinking is the "correct" answer, the other example better represents our goal of recognizing conversational social cues.

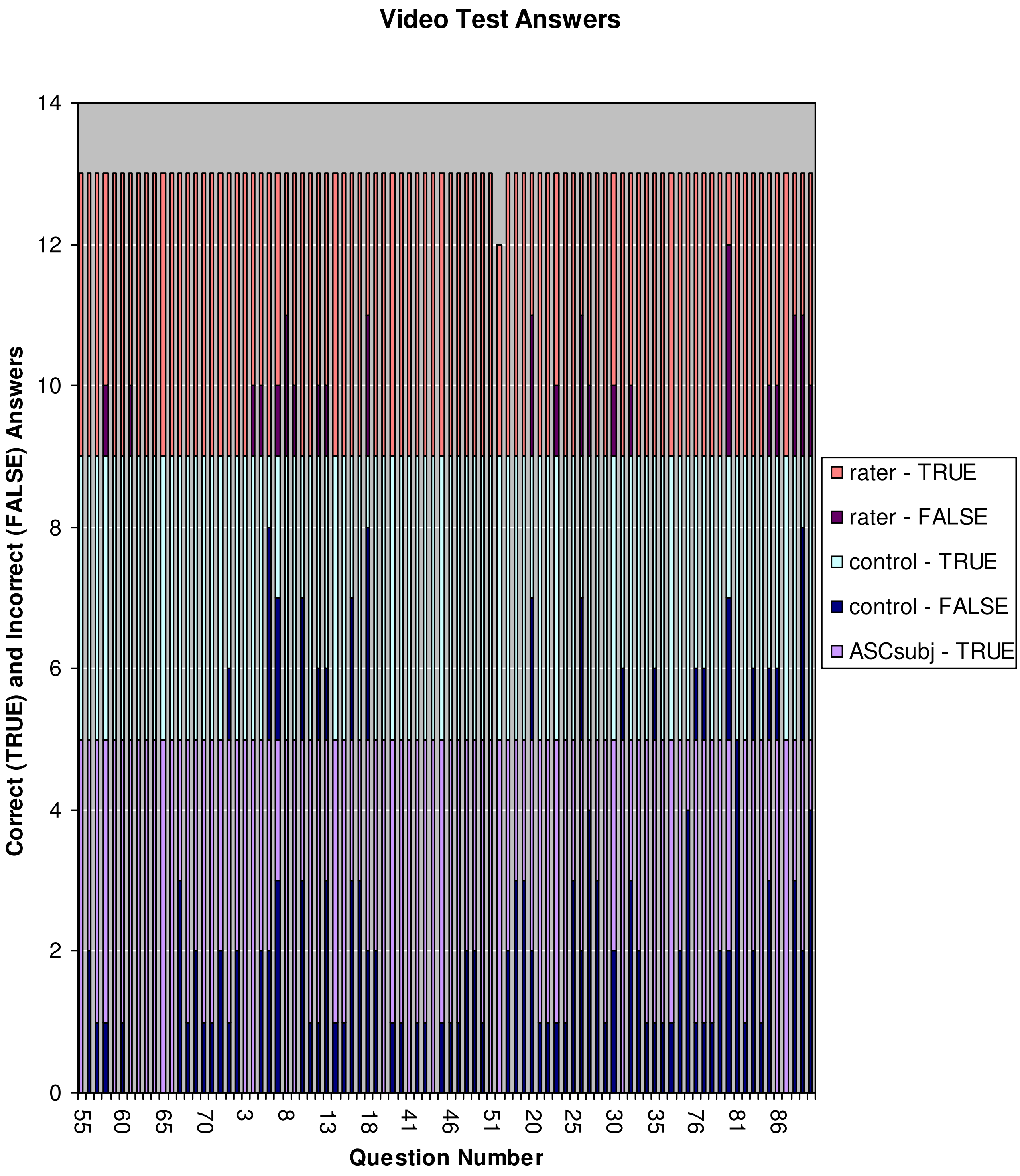


Figure 21: Individual Video Test Answers Compared by Test-taking Group

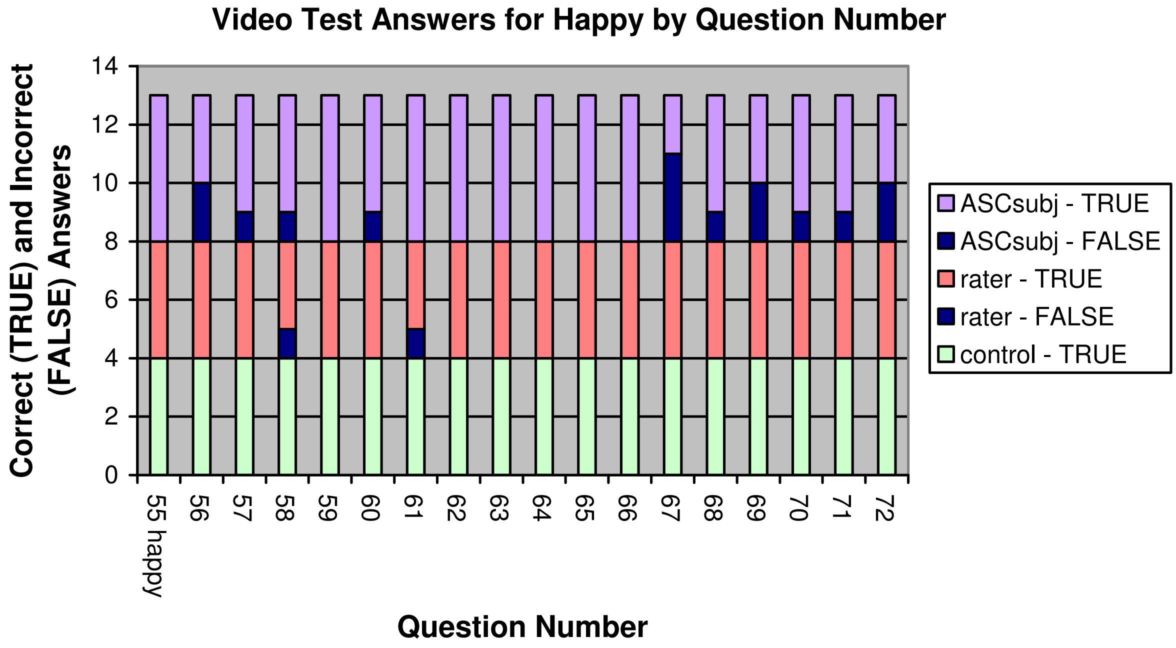


Figure 22: Video Test Answers for Happy Compared by Test-taking Group

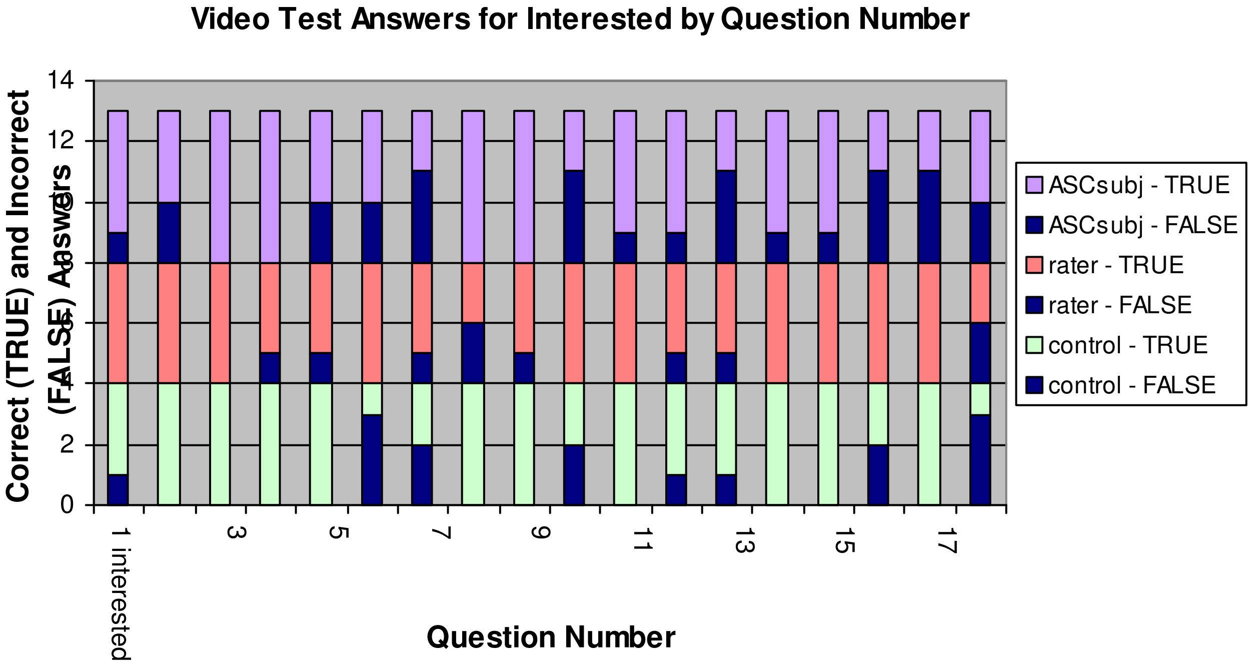


Figure 23: Video Test Answers for Interested Compared by Test-taking Group

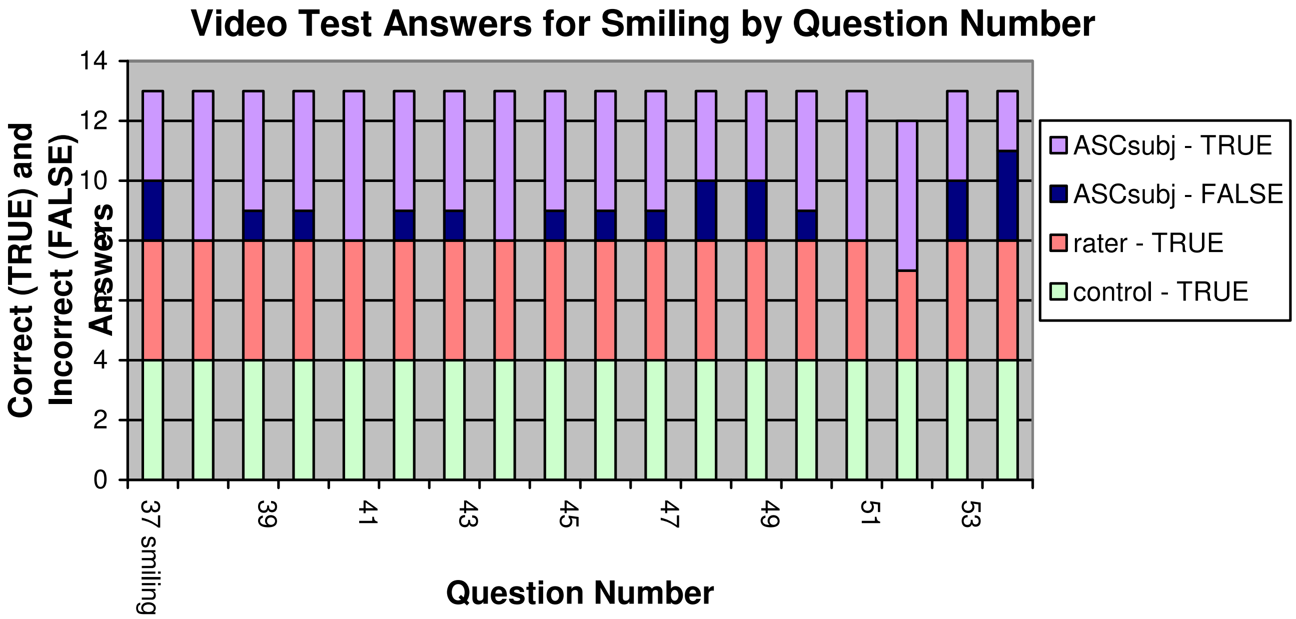


Figure 24: Video Test Answers for Smiling Compared by Test-taking Group

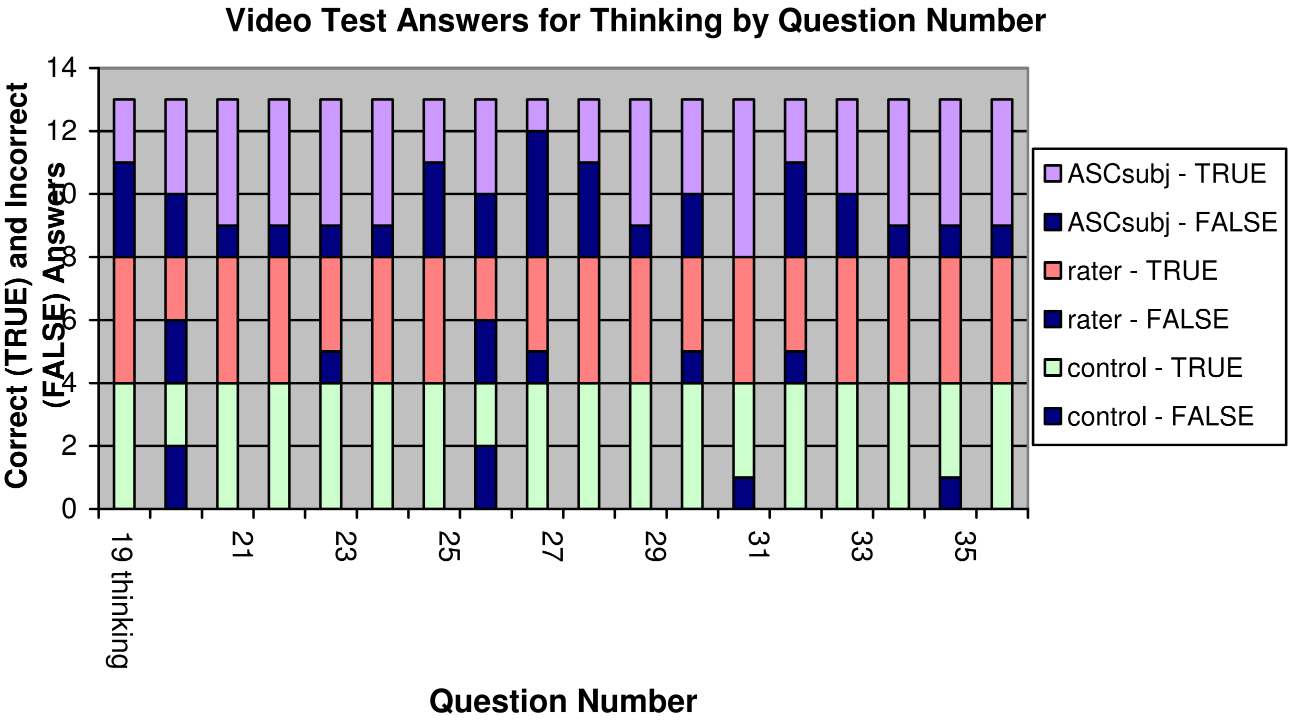


Figure 25: Video Test Answers for Thinking Compared by Test-taking Group

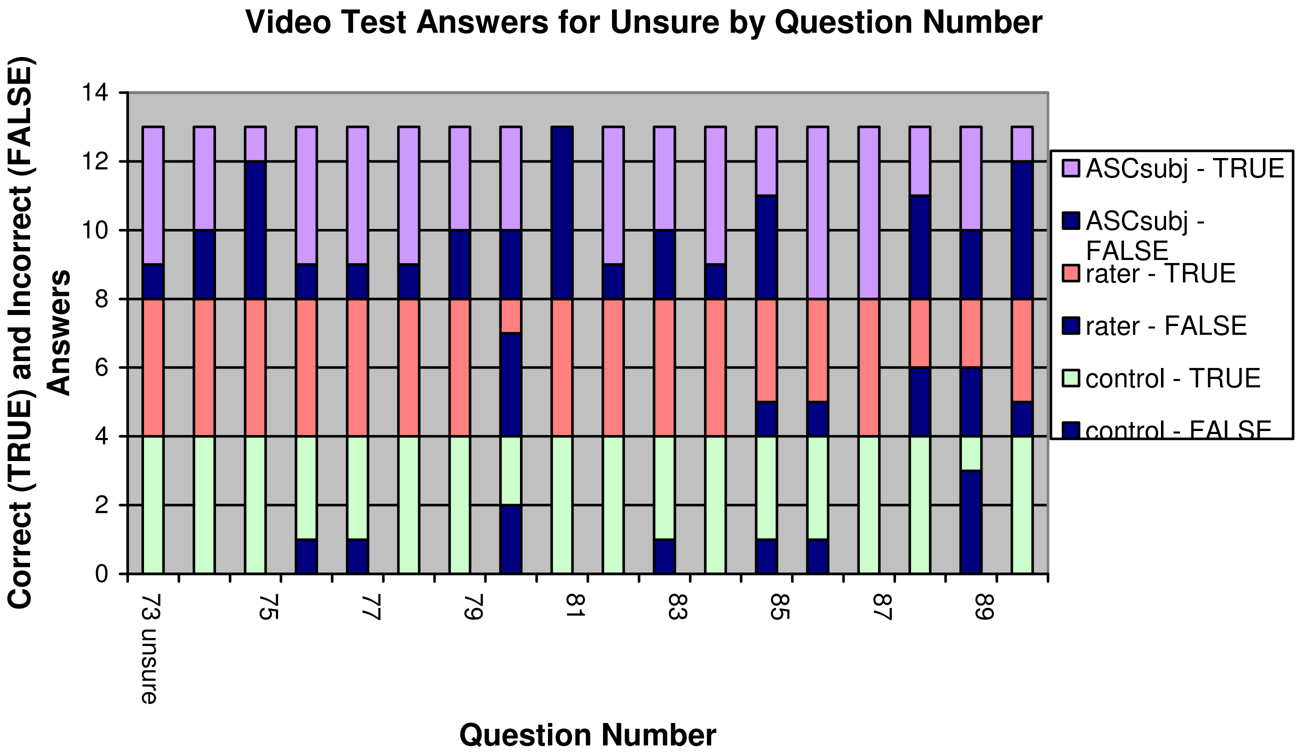


Figure 26: Video Test Answers for Unsure Compared by Test-taking Group

Looking at how the groups and each individual subject scored by social cue (Figure 27), we see that again raters and controls scored similarly while subjects scored in a different order. Smiling and happy had the highest scores for raters and controls (smiling 100% and 100 % , happy: 97% and 100% respectively). This is followed by thinking (90%, 92%), unsure (86% and 86%), then interested (86% and 79%). For subjects, happy was the highest scored social cue (83%) followed by smiling (79%). Interested (69%) had higher scores than thinking (64%), and unsure had the lowest score (61%). It appears that the scores of the subjects are positively correlated with the valence of the social cue (more positive social cues got better scores).

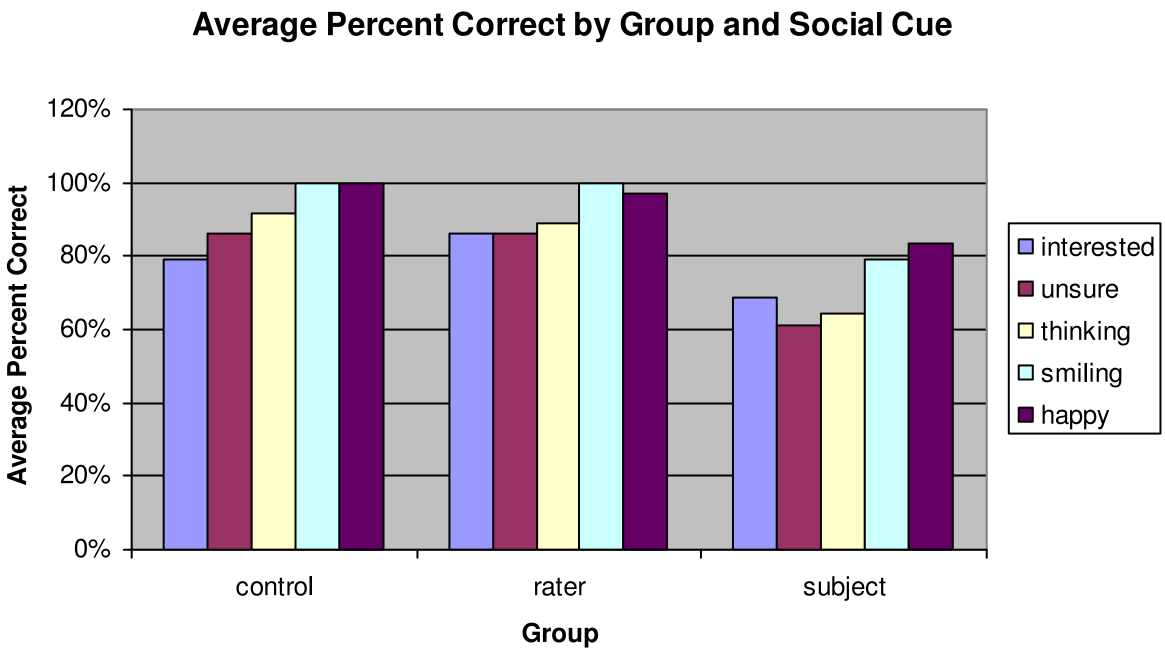


Figure 27: Average Correct by Group and Social Cue

Sorting the scores by the type of videos watched (Figure 28), we can compare how subjects, raters, and controls did on faces at different levels of familiarity to the subjects. While the subjects rated autistic peers, familiar staff, strangers, and actors, the raters and controls, because of different relationships to the people in the videos, saw autistic strangers, neurotypical strangers, and actors. However, there seems to be little variation across subject type in the averages. The biggest difference between how the subjects scored and how the raters and controls scored is a reversal when rating subjects 4, 5, and 6. Where subjects scored better on subject 5 (74%) than subjects 4 and 6 (66%, 56 %), raters and controls scored better on subject 4 (90%, 90% respectively) and subject 6 (88%, 93% respectively) than on subject 5 (78%, 80% respectively). We did not do an analysis for statistical significance because the number of subjects was small, but this very interesting difference may warrant further investigation. Subject 5 is very quiet in front of the camera, and his face may be atypically, but strongly expressive. This could put his peers at an advantage if they have already learned to read his social cues.

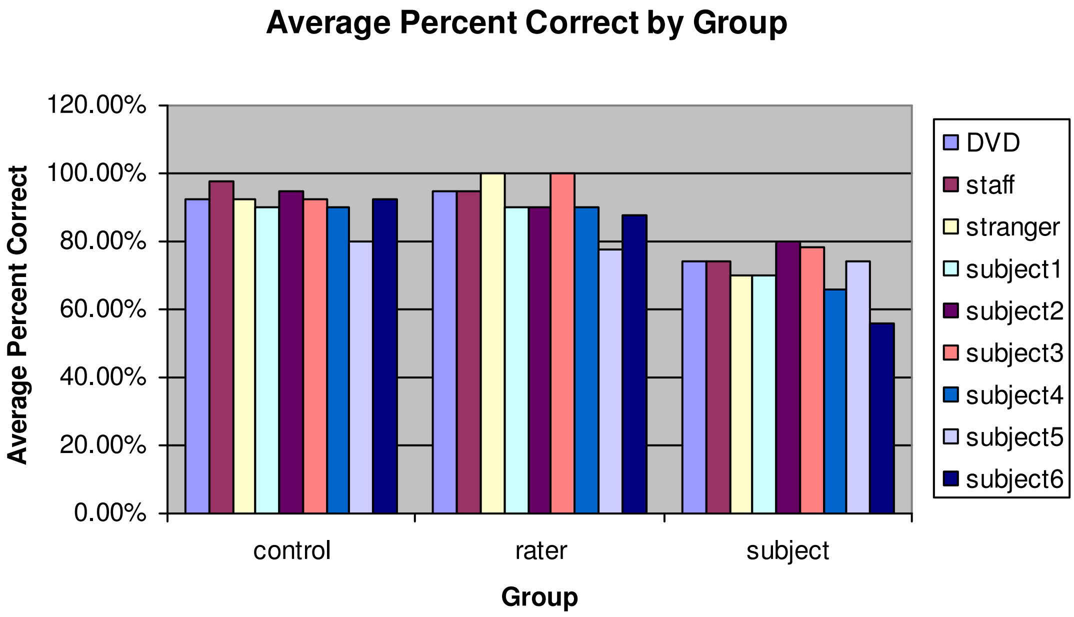


Figure 28: Average Percent Correct by Group

On average (Figure 29), the subjects did the best on three of their peers, subjects 2, 3, and 5 (80%, 78%, 74%) and equally but not quite as good (74%, 74%) on actors and familiar staff. Strangers were on par with subject 1 (70%, 70%), and this is followed by subject 4 (66%) and subject 6 (56%). In general, the higher scores reflect the general subjective expressiveness of the person in the videos. Except for subject 5, it is interesting to note that the order of the overall average scores for the individual subject on their peers are inverse to the order of how well peers scored on each of the subjects. Note especially that subject 3 scored lowest (Figure 20: 57%) and was the best understood (Figure 29: 78%). While subject 6 scored second-highest (80%) and scored the most 100% s (staff and subject 5), he was the least understood (average score 56%). If being well understood reflects general expressiveness, then perhaps the less expressive subjects are instead paying more attention to the social cues of others. It would be interesting to see how this correlates to introversion and extraversion as well.

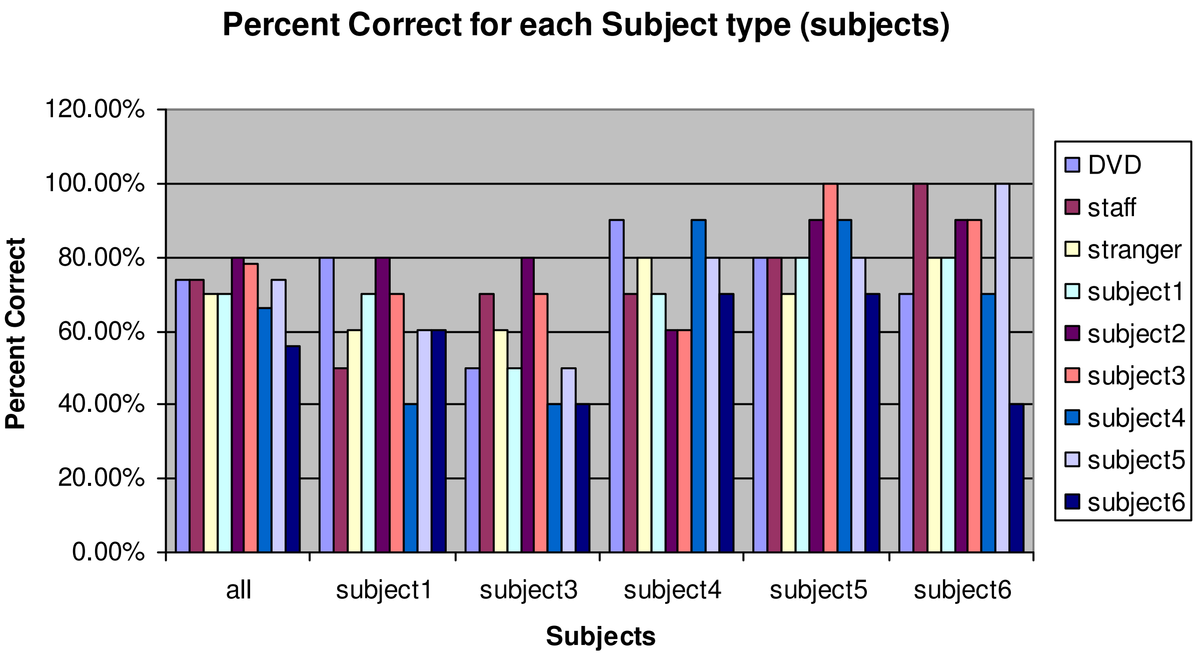


Figure 29: Percent Correct for each Subject Type

Observing the subjects as they took the video pre-test, I saw some possible complications to the results. Each subject displayed personality characteristics that make it difficult to standardize scores. Some of these are related to an ASC diagnosis and others are not. To help contextualize these comments, the times each subject took to complete the video test are given in
Figure 30.

Subject Number	Time to Complete First 50 Questions (min)	Time to Complete Remaining 40 Questions (min)
1	35	23
3	41	30
4	21	22
5	30	22
6	25	18

Figure 30: Times for Subjects to Complete Video Test

Subject 1 is easily distracted and his attention wandered from the video pre-test several times. This was more frequent as the test progressed. At first, the Groden Center Staff member walked him through each step, working the touch pad and clicking an answer according to the video that subject 1 pointed out on the screen. The social cue was read to him before and after he watched the videos for each question. Halfway through the first 40 questions and the next day when he took the remaining 50 questions, he took over the mouse work and I switched in for the Groden Center Staff member. Frequently, it was difficult to tell if he was watching and paying attention to the videos as he played them. He would sometimes appear to look away, but then go directly to what looked like the right answer. Other times, he would be looking, but would appear to choose randomly as if he was trying to get through quickly. There was no way for me to tell how well I was reading his actions and motivations.

Subject 3 also seemed to tire near the end of the second day's 50 questions. He was slightly distracted and played with the mouse on the screen. When reminded, he came back to the task. He did not appear to examine the faces closely. At first, he was delighted to see his own face in the videos. By the end, he became accustomed to it and did not react.

Subject 4 had his birthday on the second day and expressed that he liked the happy questions because he was happy. He also said he liked the smiling questions, but not the thinking questions.

Subject 5 was gone the first day, so he completed all 90 video pre-test questions on the second day. He did not appear to become distracted. For subject 5, I was reminded of Clever Hans, the math horse, who appeared to be able to do math as long as his owner was in the room. Subject 5 appeared to be doing very well on the video test and I did not know if he could be reading answers from me in his quiet concentration. In case it was possible I was sending out signals without my knowing, and he was picking up on them, I tried to not choose an answer or to find a reason to choose the answer that was not immediately obvious. However, none of this seemed to make a difference as he chose his answers without any obvious correlation to my choices.

Subject 6 had two notable dislikes: he did not like Macintosh computers, and he did not like to look at videos of his own face. He verbally expressed his preference for Windows and so the Groden Center staff moved the cursor and clicked for him on the first day. Subject 6's dislike of seeing himself on screen was more profound. He verbally stated he did not like to watch those videos and hid his face in his arm the second time his videos came up. The Groden Center staff member convinced him to continue by reassuring him that there were not many more of his own face. However, it seemed that he completed those questions quickly to get through them and I do not think he watched them more than fleetingly. The second day, there were no more videos of his own face, so he completed those 50 questions more easily.

One other thing I noticed is that many of the labels given to the Mind Reader DVD video clips by our raters did not seem to match up with the expression originally defined (and acted) for that clip. For example, a “fascinated” clip (according to the DVD labels) had labels of thinking from our coders – which is not a directly related expression in Simon Baron-Cohen’s taxonomy. Though I have not yet done a thorough comparison of the DVD defined labels to all the labels collected from our raters, it is interesting to note that other labels can easily be read into the acted expressions when raters are asked to label expressions in the context of natural interactions.

8 Conclusions

This thesis develops and implements a video test of natural social cues as a step toward creating automated tools to assist people with communication. We do this by defining a process to collect, segment, label, and use video clips from everyday conversations. We begin to map out the current skills of recognition in people on the autism spectrum and the development of those skills over time. The motivation to use technology comes from the ease with which technology completes repetitive tasks and the appeal of consistent, systematic, logical, and predictable learning mechanisms to people on the autism spectrum.

The implemented the pre-assessment for the pilot of a longer study set a baseline for the subjects involved in that pilot. The video test was developed to measure the progress of social cue understanding over time and its use established the ability of autistics to rate videos of social cues. The pool from which we drew video clips for the 90 questions on the video test included 1128 video clips taken from 51 videos an took countless hours to process by hand. The test demonstrated that, in autistic subjects, there are differences when rating different social cues and when looking at videos of self as opposed to videos of others. There may also be people and faces for which autistics may understand comparatively better because of familiarity.

The video test contains clips to gauge a person's generalization of reading face and head gestures and specific abilities with faces of self, peers, familiar adults, strangers, and unknown actors. Except for subject 6, who did not like to watch videos of his own face and who did very poorly on those ten self-videos (perhaps because of this dislike), there was not a clear difference in the scores on the types of videos before the intervention. However, we hypothesize that the difference in generalization will show up after the intervention takes place, where we predict better generalization when non-strangers are used in the intervention.

The capture of videos of natural interactions succeeded with the use of Self-Cam, but the social cues captured were somewhat limited in scope because of the limited situations captured. In the future, it may be possible to increase the number and diversify the types of social cues captured by improving Self-Cam to make it more subtle and by sending people out to capture video at different times of day and in different environments. At times, the equipment interfered with the quality of the video. Some videos bobbed up and down due to bouncing of the Self-Cam wire. Other videos seemed to be impaired by a subject's nervousness while being recorded (particularly with subject 5, whose nervousness did not seem to abate over time). For the most part, the captured video and resulting social cue examples appear very natural.

This thesis encompasses steps toward easier analysis of our facial expressions and those of the people around us so that we may better understand ourselves and others and use this knowledge to improve communication. Interesting results include the comparison of individual video test questions where raters/controls and autistic subjects had nearly opposite answers, a peak in the subjects' ability to identify social cues in subject 5 (the opposite of raters/controls), and an impression of some of the reasons behind the abilities of some very different people who are a small part of the autism spectrum. I learned that, as neurotypicals, we assume a lot about other people based on how we ourselves would feel in their situations. With everyone, but in particular with autistics, we are probably more wrong than we could possibly think. Through this research, I hope we've come a little bit closer to understanding those differences.

9 Future work

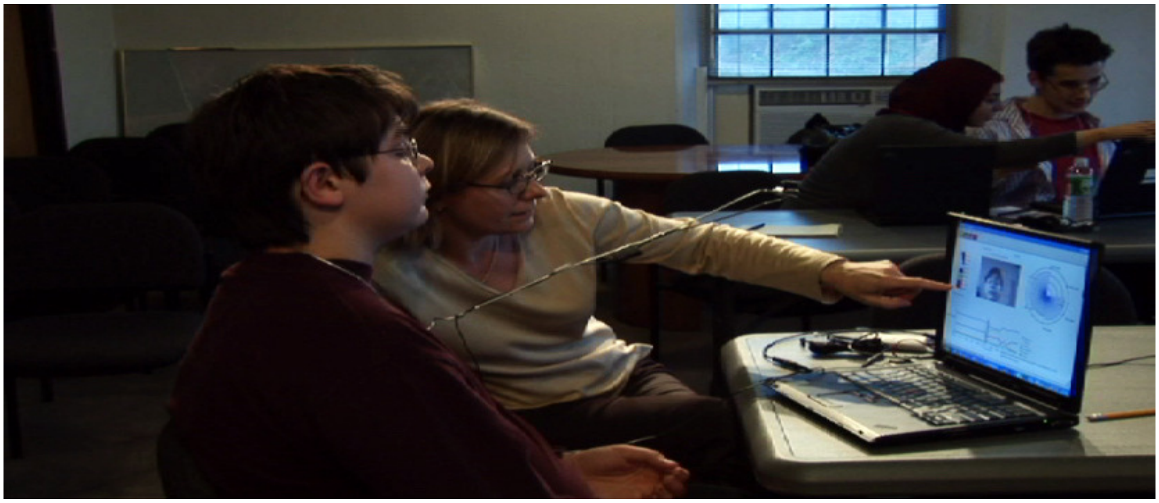


Figure 31: Reviewing Facial Expressions with the Mindreader Software

There are several observations we have made over the course of the work which could lead to better data collection and possibly lead us in new directions as we continue to implement this study.

During data collection, subject 5 was very shy on camera, though he did not seem to mind putting on the equipment. I wonder if standing and having more free and active movement during the conversation might keep him from closing up. This may counteract the oppressiveness of an experimental structure.

During completion of the video test, the subjects required assistance, and this was performed by the experimenters. To prevent possible biasing of the results, an independent party could perform this task, but would still need careful instructions to try to prevent them from choosing what they think is the best answer and possibly influencing the results. A better solution would be to continue work on making the video test interface self-administered. This could include automatic reading of instructions and of each social cue as it comes up (an audio file could be played by the computer at the start of each question and after each video is played). However, someone may still needto be present in some cases, to keep the student on task.

Once the study is complete, there will still be many directions to explore. One of the near-term goals of this line of research is to develop a real-time, wearable tool that can be used to develop new, face-related social skills over time by providing live and post-recording feedback related to expressions of emotion and social cues. To this end, there are three main pieces of work to be explored. These are: improving the wearable device including the Self-Cam and processor, integrating the Mindreader system to provide live feedback, and incorporating additional sensors and feedback devices to improve situation recognition and to accommodate additional sensory modalities and sensitivities. The current study will also be improved and run on a larger scale.

The wearable device may be improved by making it less obvious to observers or by designing a look that is more integrated into technology fashion. For example, a wide-angle lens could allow the placement of the camera to be moved closer to the face. Though this would require a more powerful computer to do the real time processing or retraining of the Mindreader software, it would allow the small camera to be placed subtly on the brim of a hat or on a goose-neck support extending only slightly from the strap of a backpack or earpiece. Another possibility is to use a small mirror, similar to a rear view mirror mounted on a bicycle helmet. The mirror would allow a larger camera to be used, as it could be mounted close to the body.
Creating a technology that appeals to a wide population may also make the device less cumbersome or embarrassing to wear in a wider context than labs and homes. The use of the Self-Cam for entertainment or general communication may make it as natural to wear as the cellular phone or bluetooth headset is today. In addition to the physical appearance, the processing power of small devices will improve over time. Making a dedicated processing system would speed up the real-time data analysis and could be designed to be self contained and easier for the user to operate.

In future versions of the video capture sessions, we hope that Mindreader will replace the manual division and hand coding of the videos and video clips. When faces are successfully tracked by separate software, Mindreader (Figure 31) can be trained to recognize social cues in addition to the six it now evaluates depending on the application. Mindreader can also be trained on the corpus of natural conversation videos developed in this thesis for better recognition during everyday use of Self-Cam. If Mindreader is able to successfully recognize, mark, and save video clips in real time (also dependent on working with a good tracker), there are endless applications that can be designed for reviewing the videos, and giving live feedback. Whether these applications help autistics learn neurotypical social cues, help teachers and students develop better public speaking skills, or help machines interact more fluidly with humans, the ability to create an objective database of recurring facial movements and patterns has potential to give us valuable insights into the way we function as people.

Also useful in future versions of Self-Cam would be additional sensors and actuators to assist in understanding the scenario for appropriate application of teaching techniques. Skin conductance and audio recognition of things like social turn taking and vocal affect could provide further clues to underlying mental states as well as displayed social cues. These sensors have the potential to better discern social context and provide better evidence for giving appropriate feedback to users. Such feedback may consist of audio marking of a chosen social cue, such as confusion, spoken feedback as reminders of congruent behavior to detected social cues (possibly by bluetooth headset), or an LED or tactile buzzer (possibly through a cell phone or pager) as a set of markers or reminders. These different actuated modalities would allow a user to choose a preferred mode of feedback. This choice is especially important for the autistic population as many autistics are hypersensitive in certain sensory modes and/or have trouble with sensory integration.

And finally, the most pressing future work is the completion of this pilot study and the implementation of a full study using the preliminary results to guide the design.

10 Appendices

10.1 Self-Cam Trials

This section describes the first uses of Self-Cam as it was initially developed. Self-Cam was used with a variety of cameras and recording systems and was tested in real-life situations to get a feel for how the interactions might proceed and for what type video scenarios might be interesting to capture and process. Based on these trials, the current version of Self-Cam was formed.

10.1.1 Self-Cam to CBC

The first trial run of Self-Cam was a lunch trip to a local restaurant using the Dejaview camera system. Dejaview is a commercial system that allows the user to push a button to save the last several seconds of video, which is continuously buffered. The small camera was placed at the end of the Self-Cam wire and triggered manually. The wearer triggered the Dejaview frequently, attempting to capture the most dynamic facial displays, and observed that it was difficult to act naturally while constantly thinking about her own face and appropriate times to push the record button.

Apart from the wearer, people walking down the street seemed to ignore the camera, as did the waitress at the restaurant. The question came up whether this was because people in the vicinity of MIT were used to technological curiosities or if it was a natural tendency. At the restaurant, eating was made more difficult by the wire directly in front of the wearer, but possible. The wire was also worn under a winter coat with no difficulties.

10.1.2 Roz/Barry Conversation

On the second trial of Self-Cam, the Dejaview was made capable of continuous recording for up to 5 minutes. Rosalind Picard and Barry Kort sat across from each other in comfortable chairs, each wearing a Self-Cam system. The object of this trial was to attempt to simulate a monologue situation, as monologuing is a common problem cited by parents of kids with Asperger's syndrome. Therefore, Barry chose a subject to talk about and attempted to talked continuously while Roz listened.

The conversation was recorded by a standard video camera as an overview of the entire experimental process. The process was interrupted every 5 minutes to download the data from the Dejaview boxes onto a computer, thus freeing them for the next 5 minutes of data. During each transition, the situation was discussed and small changes were made to better simulate a natural one-sided conversation.

The frequent interruptions were "difficult to get used to," as people are not accustomed to stopping their conversations abruptly, by a strict time period. There was no indication of the system being near full, but only an indicator when the memory card was completely full. Barry observed that wearing the Self-Cam was strange at first, but that sometime between 15 and 30 minutes of wearing it, the awkwardness was no longer an issue. During the conversation, Roz observed that she looked up and left when she was thinking. She surmised that this natural signal indicated a need for time to think or to ask a question. Roz also observed that the Self-Cam made her reflect more on her own social signals to others and those of the people around her than any other recording experience. Technically, it was observed that standing, sitting, and slumping made a difference in the view of the camera. The slope of the chest at any given rest position dictated the view of the camera. Though this view change was not dramatic, it was significant if a computer was to do any automatic post processing. Therefore, it was discovered that it was important to conduct a camera view check at the start of each capture session.

During review and processing of the video data, Mindreader caught a clear transition that looked like a move from interest toward boredom or exasperation. Mostly a change of head position, the movement was read by the software as a dramatic change of mental state. On later review, Roz indicated that the continuous information received from Barry was interesting, but that she had needed a small break to process what was being said. The fact that there was no such break in the one-sided conversation caused some frustration and caused her mind to eventually lose track of what was being said. This particular situation, if found to be a common indicator, could be one key to creating a function live feedback system for people who have difficulty reading social cues in natural situations.

10.1.3 BSN Data – break, lunch, dinner cruise

During the Body Sensor Networks conference, the Self-Cam system was worn by two people while socializing with other conference attendees. During breaks, lunch, and during a dinner cruise, Self-Cam was worn by two group members. During the cruise, a Hat-Cam was also worn. The wire arm was redesigned into a single wire extension resting on the chest for a cleaner look and extended for a wider view of the wearer's face.

During the conference, it was found that the camera systems were a "magnet for curiosity." The system gave attendees an excuse to approach the wearer and to ask questions about the sensor. Many people guessed that the camera, because of its tiny size, was actually a microphone. It was also observed that people tended to stand side by side while one was wearing a Self-Cam during conversation. This was observed in one-on-one conversation, as this position is quite common in groups. Standing side-by-side allowed conversers to be closer to each other while avoiding the introduction of an object directly in the middle of the conversation.

10.1.4 Group Lunch

The Archos digital video recorder was used in conjunction with the Self-Cam capture video during a group meeting over lunch at MIT. The Archos device extend the time limit of the recording period from 5 minutes to at least 15 minutes of interaction. Five people were recorded while sitting and conversing at a round table. Unfortunately, there were problems with the settings on the Archos devices and 2 sets of data were lost.

10.1.5 Open Source Magic Show

Performed by Seth Raphael, the open source magic show was the first test of recordings taken of a performer and audience in coordination. Several audience members agreed to wear the Self-Cam throughout the performance and the entire audience was recorded using a standard video camera. Seth was recorded by Self-Cam and also by a camera with a standard view of the performance. Various devices were used for capturing video in order to increase the number of audience members who could participate. Devices included laptops, Archos DVR's, and small handheld video cameras, all placed on a wire at arm's length from the body.

The videos were coordinated by turning the auditorium lights off, flashing a camera flash, then turning the lights on again. It was later discovered that this did not give a precise timing as the lights did not change instantly and the camera flash was not visible to all video devices. Despite this, the faces were coordinated in time as best as possible and collaged into a grid of videos showing the performance, the performers face, and the audience faces as they reacted to the show. No digital processing of the faces in the video was done, but the display of the faces together provided an interesting view of the various reactions and lack of reactions to the performance of magic.

10.1.6 China

As a continuation of the capture of reactions to magic, Seth Raphael took the Self-Cam to China to test on the streets and begin his research into the exploration of wonder. Performing for various people he met, Seth had them wear the camera and qualitatively explored the types of reactions displayed.

10.1.7 Mindreader and SIGGRAPH

Self-Cam was taken to the SIGGRAPH 2006 conference as a poster session and live demo for attendees to try out themselves. In this location, the users wore Self-Cam and were placed directly in front of the screen where they had direct visual feedback of their faces and Mindreader's tracking points and facial features, optional audio feedback of the most likely mental state expressed (recognized by Mindreader), and indirect visual summaries that included a line graph of all six recognizable mental states and a pie chart record of the most likely mental states, accumulated over time. The feedback was explained to users and they were asked to try out the different mental states to see if they could trigger recognition by Mindreader. Each person was then given a screenshot printout of Mindreader's visual output with his or her face.

10.2 Smile Detector

A small addition was made to the Mindreader software for demonstrations at the Media Lab. When activated, the live feed would freeze upon detection of a lip pull (smile). This is an example of possible future applications where social cues could be detected automatically and signaled in real time to a person wearing the Self-Cam apparatus.

10.3 The Groden Center's Asperger's Group

After SIGGRAPH, the Self-Cam system was stable enough to test with the target test population in order to get their feedback on the design and to brainstorm ideas for its use as a tool for Autism and Asperger's syndrome. We took several systems to a meeting of a group of teenagers with Asperger Syndrome at the Groden Center in Providence, RI. This group tried out the Self-Cam while looking at the Mindreader feedback as described for SIGGRAPH, and also used the Hat-Cam with the Archos DVR as a simple recording device. Most of the teenagers used Self-Cam with some trepidation at first, but were very interested in exploring the reactions of the output on the computer screen. A couple of teenagers also turned the camera on Self-Cam to face the person explaining Mindreader and asked the person to make faces to trigger the mental states recognized by Mindreader. One turned it on his pet lizard in hopes that Mindreader would recognize the lizard's face (it did not).

While using the Hat-Cam, the teenagers stared into the Archos while walking around the room. Some approached other people and recorded them briefly while still staring into the screen. Afterwards, the teenagers and their parents sat down to discuss the technology and ask and answer questions. As a whole, they were very excited about finding new technology that might help them in their daily lives.

10.4 AANE

At the Asperger's Association of New England meeting in Waltham, we explained Mindreader and Self-Cam to a group of mostly adults and a couple of teenagers with Asperger's syndrome and asked them for their comments. They then were able to try wearing the device while viewing the feedback on the computer screen. Some were very excited at the prospect of being able to read another person's face. One man was particularly interested in using such a system for dating, to recognize a signal to "back off" or "don't go there." One person thought he would be anxious with audio feedback, always wondering "is it going to tell me something now?"

REFERENCES

American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders. Washington DC, American Psychiatric Association.

Baron-Cohen, S. and O. Golan (2004). Mind Reading: The Interactive Guide to Emotions. London, Jessica Kingsley Publishers.

Baron-Cohen, S., S. Wheelwright, J. Hill, et al. (2001). "The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism." The Journal of Child Psychology and Psychiatry and Allied Disciplines **42**(02): 241-251.

Centre, A. R. (2001). "Eyes Test (Adult - Part 1 and 2)." Retrieved May 2007, from http://www.autismresearchcentre.com/tests/eyes_test_adult.asp.

Ekman, P. and W. V. Friesen (1971). "Constants Across Cultures in the Face and Emotion." Journal of Personality and Social Psychology **17**: 124-129.

Ekman, P. and W. V. Friesen (1978). Facial Action Coding System, Consulting Psychologists Press.

el Kaliouby, R. (2005). Mind-Reading Machines: Automated Inference of Complex Mental States, Phd thesis, University of Cambridge, Computer Laboratory.

el Kaliouby, R. and P. Robinson (2005). "The Emotional Hearing Aid: An Assistive Tool for Children with Asperger Syndrome." Universal Access in the Information Society **4**(2).

el Kaliouby, R. and P. Robinson (2005). Generalization of a vision-based computational model of mind-reading. First International Conference on Affective Computing and Intelligent Interaction
Beijing.

el Kaliouby, R. and P. Robinson (2005). Real-Time Vision for Human-Computer Interaction, Springer-Verlag 181-200.

Gemmell, J., G. Bell, R. Lueder, et al. (2002). "MyLifeBits: fulfilling the Memex vision." Proceedings of the tenth ACM international conference on Multimedia: 235-238.

Golan, O. and S. Baron-Cohen (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.

Healey, J. and R. W. Picard (1998). StartleCam: A Cybernetic Wearable Camera. International Symposium on Wearable Computers, Pittsburgh, PA.

Hodges, S., L. Williams, E. Berry, et al. (2006). "SenseCam: A retrospective memory aid." Proc. 8th International Conference on Ubicomp.

Lockerd, A. and F. M. Mueller (2002). "LAFCam: Leveraging affective feedback camcorder." Conference on Human Factors in Computing Systems: 574-575.

Mann, S., M. I. T. (1997). "Wearable Computing: A First Step Toward Personal Imaging." Contact **30**: 25-32.

Moore, D. (2000). "Computer-Aided Learning for People with Autism–a Framework for Research and Development." Innovations in Education and Teaching International **37**(3): 218-228.

Sinclair, J. (1993). "Don't mourn for us." Our Voice **1**(3).

Starner, T., S. Mann, B. Rhodes, et al. (1997). "Augmented reality through wearable computing." Presence: Teleoperators and Virtual Environments **6**(4): 386-398.

Teeters, A., R. el Kaliouby and R. Picard (2006). "Self-Cam: feedback from what would be your social partner." International Conference on Computer Graphics and Interactive Techniques.