Understanding Children’s Gaze Behaviors

An Undergraduate Thesis
Presented to
The Academic Faculty

By
Yongxin Wang

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science in Computer Science in the
School of Computer Science

Georgia Institute of Technology
April 18, 2018

Understanding Children's Gaze Behaviors

Approved By:

Dr. Jim Rehg, Advisor
School of Interactive Computing
Georgia Institute of Technology

Date: 04/26/ 18

Dr. Agata Rozga
School of Interactive Computing
Georgia Institute of Technology

Agata Rozga

Date: 04/24/2018

Table of Contents

Abstract

Identifying early signs of autism has been a challenging problem in the medical field. Many research studies aim to detect behavioral patterns of children with autism in the first three years of life. Early detection of autism allows early intervention to be initiated and thus is essential to achieving the best long-term outcomes for children with autism.

Under the guidance of Drs. Jim Rehg and Agata Rozga, and mentored by PhD student Eunji Chong, I explored the challenge of developing automated measures to analyze children's gaze behaviors during social interactions. We focused on using computer vision and deep learning techniques to analyze children's attentions to objects and their social partner during play. In this thesis, I will first talk about related research works in the area of detecting autistic behaviors in children. I then introduce the two major pathways that I have taken to explore this problem, which are children's head pose analysis and children's gaze analysis. Details of the algorithm and the results are also presented and discussed. Finally, I will also address some failure cases and propose potential future works.

Introduction

Research aimed at measuring early behavior in autism usually involves manual annotation of pre-recorded videos of children engaged in a variety of structured and semi-structured interactions (J. Hashemi et al., 2012). While this approach generates rich data, it is inefficient and tedious. To achieve more efficient and scalable means of analyzing children's behavior, researchers have turned to sensors, machine learning techniques, as well as computer vision.

In the study by Cheol-Hong Min (Min, C.-H. 2017), child behavior data are recorded by several wearable accelerometers, based on which the behaviors of children with autism are analyzed using a Hidden Markov Model (HMM). J. Hashemi et al. use a computer vision approach to analyze behaviors of children with autism (J. Hashemi et al., 2012). In another research by Abbas, H., et al., the behaviors of children with autism were measured by analyzing parent-questionnaires and home recorded videos (Abbas, H., et al. 2017). The questionnaire data and video data are further fed into a Random Forest algorithm for the analysis of child behaviors. Rajagopalan, S. S. (n.d.) presents a method that utilizes computer vision to detect children's repetitive motion patterns, which Rajagopalan believes are atypical child behaviors indicating a risk of autism (Rajagopalan, S. S. (n.d.) 2017).

These aforementioned research projects provide great methods in detecting the behaviors of children with autism, but they all have some weaknesses. Cheol-Hong Min utilizes wearable sensors (Min, C.-H. 2017), which can be intrusive to the child. J. Hashemi et al.'s did not take advantage of modern deep learning (e.g. J. Hashemi et al. 2012). Traditional clinical methods, like filling questionnaires and analyzing videos (Abbas, H., et al. 2017), require time-intensive manual annotation, which is very inefficient and prone to human errors.

Our research focuses on developing non-intrusive and efficient behavior measurement methods that leverage state-of-the-art computer vision and deep learning techniques. We have recorded over 100 videos sessions where children play with toys and interact with an experimenter during a semi-structured, standardized play protocol, the Early Social Communication Scales (ESCS; Mundy et. al 2003). The children are recorded with two cameras, a camera embedded in a pair of glasses worn by the experimenter and a tripod-mounted camera that captures the child upper body, the tabletop where the toys are presented, and the experimenter in profile. We use a tool called openpose, developed at Carnegie Mellon University, to detect the facial landmarks of each child in both types of videos (Z. Cao et. al.) (S.-E. Wei et. al.) (C. Qian et. al.) and developed an algorithm to determine the head pose based on landmark detections. Details of the algorithm and our results are discussed in later sections. We also designed a deep convolutional neural network to predict children's gaze target by extracting their gaze angles and saliencies during these interactions. Finally, we show visual results of our gaze detection method, and discuss future work.

Literature Review

Past research on automated measurement of behaviors in autism can be categorized roughly into three types: methods with wearable devices, methods with interactive software, and methods with computer vision techniques.

i. Wearable Device Based Methods

The work of Cheol-Hong Min introduces a method that automatically detects and labels self-stimulatory behavior patterns, which are commonly observed in children with Autism Spectrum Disorder (ASD) (Min, C.-H. 2017). Min places wearable sensors on children's wrists, waist, torso, legs, and ankles to record movements produced when children engage in these behaviors. With a Hidden Markov Model, Min is able to classify and identify the children's self-stimulatory activity patterns from the recorded behavior data. This research introduces an HMM model to detect behaviors often observed in autism. However, the use of wearable sensors can be intrusive to the children, something our work is trying to eliminate.

ii. Interactive Software Based Methods

In the research by Mohamed, A. O., et al., a system consisting of an interactive computer game and several autism experts is introduced. (Mohamed, A. O., et al. 2006). Though the final goal of this system is to help children with autism in their rehab processes, this paper focuses a robust method of attention measurement. During experiment sessions, children play the computer games, while being observed by experts who assess the children's attention. The computer system allows experts to see when and for how long the children are looking at. Thus, the experts can measure the levels of attention by estimating the correlation between the children's gaze and the trajectories of objects presented to the children by the computer system. If the gaze and the trajectories are correlated, the child is considered attentive. Otherwise, the child is inattentive. Though this work provides a robust measurement of children's attention, computer technology only plays a small part in the measurement pipeline. Their approach heavily relies on human observation, and is thus error prone and time consuming. This is also something that my research aims to avoid.

The work by Abbas, H. et al (Abbas, H., et al. 2017) presented a machine learning approach that leverages parent-questionnaires and home recorded videos to measure the behavior of children with autism early in life. Abbas, H., et al develop software platform that guides parents to fill out a survey and record videos of their children. The survey data is then sent to the researchers and fed into a Random Forest Classifier. The video is viewed by expert analysts who then fill out another data survey, which is also fed into the Random Forest Classifier algorithm. Finally, Abbas, H., et al run the learning algorithm to estimate the likelihood of autism for the participating children. In the video analysis portion of their work, classifiers are trained based on the ADOS (Autism Diagnostic Observation Schedule) assessment, which is considered a gold standard of clinical methods to diagnose autism. This work suggests several great aspects of detecting autism, such as utilizing ADOS modules. However, significant human effort is still required.

iii. Computer Vision Based Methods

> In the paper “Computational Behavior Modeling for Autism Diagnosis” (Rajagopalan, S. S. 2013), Rajagopalan introduces a computer vision approach to detect repetitive motion patterns of children with autism. Rajagopalan uses a clustering algorithm to cluster motion trajectories in prerecorded videos based on the mean positions of the trajectories, and then trains a Support Vector Machine to determine if a motion is repetitive. The biggest strength of the method is that a detailed algorithm is provided for classifying repetitive behaviors, which can be useful in diagnostics of autism. However, identifying the relevant portions of children’s motions is still done by hand.
>
> J. Hashemi *et. al.* proposes another method that also uses computer vision to analyze children’s faces and body poses (J. Hashemi et al 2012). The paper also provides useful metrics for detecting relevant behaviors such as Shared Interest (“ability to use eyes to reference and share interest in an object or event with another person”), Visual Tracking (“ability to visually follow a moving object laterally across the midline”), Disengagement of Attention (“ability to disengage and move eyes/attention from one of two competing visual stimuli”), and Atypical Motor Behavior (“behaviors like atypical gait, locomotion, motor mannerisms/postures or repetitive motor behavior”). Though it is one of the most related work to our research, this work is not leveraging much of the modern deep learning techniques.

As described in the previous paragraphs, there are many related works in the field of detecting behaviors in children with autism. Some provide concrete descriptions of how to estimate repetitive motion patterns in a given video, and others give insights into how to integrate clinical gold standard instruments into machine learning.

Despite these strengths, the drawbacks of these methods lie in substantial manual effort, wearable devices that can be intrusive and not tolerated by many children, and inadequate integration with modern deep learning techniques. Our research aims to fill in the gap by using non-intrusive sensors and computational methods that leverage modern deep learning techniques to study a class of behavior (attention to eyes and toys during play interactions) that have been found to capture the earliest signs of autism.

Methods and Materials

i. Head Pose Analysis

a. Data

In collaboration with researchers at Weill Cornell, we invited children with Autism Spectrum Disorder (ASD) and Typically Developing (TD) to come to our labs and participate in Early Social Communication Scales assessment (ESCS; Mundy *et al.*, 2003). The ESCS is a semi-structured, play protocol that assesses children's nonverbal communication. An examiner presents different toys to the child that are intended to elicit nonverbal behaviors such as gaze shifts and gestures. Children's interaction with the examiner is recorded by two cameras, a tripod-mounted camera and a wearable camera that the examiner wears. Figure 1 shows our recording setup (E. Chong *et. al.*). Trained annotators use specialized software to review the videos and mark the onset and offset of each instance of the child making eye contact with the examiner or attending to the toy.

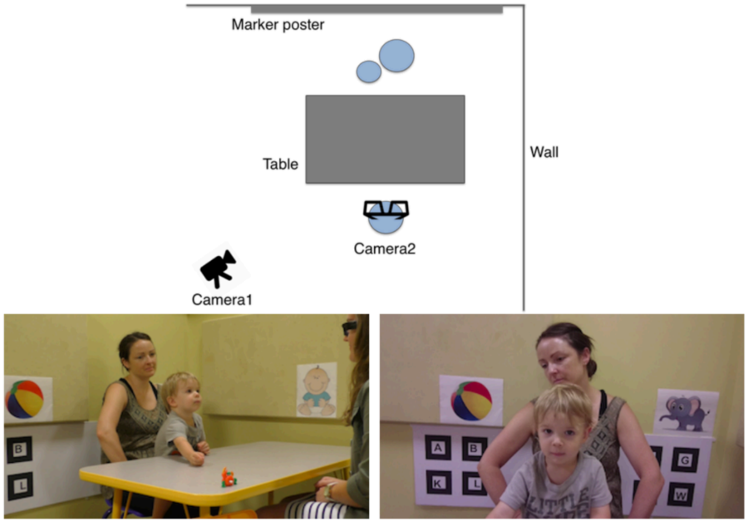

Figure 1. ESCS recording setup (E. Chong et. al.)

b. Methodology

We first focus on analyzing children's head movements. This is done by calculating children's head pose changes during the sessions. There are three steps to calculate the children's head pose. First, we detect facial landmarks using openpose, a tool created at Carnegie Mellon University that uses deep learning to extract 68 facial landmarks from images (Z. Cao *et. al.*) (S.-E. Wei *et. al.*) (C. Qian *et. al.*). Next, once the facial landmarks (u, v) are obtained, we solve a 3D to 2D translation matrix using a generic 3D child head model with 3D coordinates (x, y, z) of the corresponding 68 facial landmarks. Finally, we calculate the head pose based on the solved 3D to 2D translation matrix.

ii. Gaze Analysis

The interactions between children and an adult are important in the diagnosis of developmental disorders such as autism (Rehg et. al.). To understand these interactions, one important element is gaze analysis. We designed a deep convolutional neural network to predict the child's gaze (i.e. where a child is looking at) in an image. A non-intrusive, automatic gaze measurement, our method can predict not only the gaze of a child, but also that of a person in general.

This part of our work is submitted to European Conference of Computer Vision (ECCV). Because our paper is still in the process of review, I will discuss our work at a high level and show some visual results. For details such as network implementation, network architecture, training scheme, and experiment design, please refer to our paper: "Connecting Gaze, Scene and Attention."

a. Data

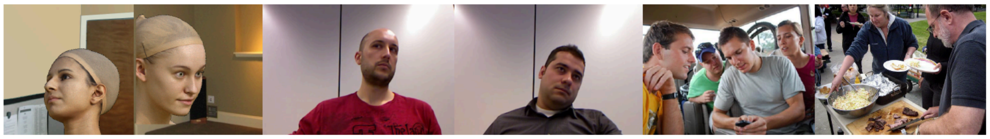

Figure 2. Example datasets used, NVIDIA SynHead (Left two) (Gu J. *et. al.*) *, EYEDIAP (Middle Two)* (Funes Mora *et. al*)*, MIT GazeFollow (Right Two)* (A. Recasens *et. al.*)

As shown in Figure 2, we use three main data sources: NVIDIA SynHead (Gu J. et. al.), EYEDIAP (Funes Mora et. al), and MIT GazeFollow (A. Recasens et. al.). The NVIDIA, SynHead, and EYEDIAP datasets contain images with ground truth gaze angle in yaw and pitch, and the MIT GazeFollow dataset contains images with gaze target annotations. The MIT GazeFollow dataset assumes that the people in the image are only looking inside the image frame. Even though a person is clearly looking at something that is outside the image frame, this dataset still labels a gaze target inside the frame. This annotation scheme could be potentially problematic because it does not distinguish between looking inside and looking outside cases, leading to wrong saliency predictions when a person is looking outside the image. We modify the MIT GazeFollow dataset by adding a binary label indicating whether a person is looking inside or outside the frame. Finally, our deep learning model uses the SynHead dataset, the EYEDIAP dataset, and the modified MIT GazeFollow dataset to learn gaze angle prediction, visual saliency prediction, and the final gaze target prediction.

When testing our model, we use the MIT GazeFollow as well as the Multimodal Dyadic Behavior (MMDB) dataset proposed by Rehg *et. al.* Similar to the ESCS dataset that we used to analyze head poses, the MMDB dataset contains footages of a child and a trained examiner participating in an interactive game session. Instead of using the ESCS, the MMDB dataset employs the Rapid-ABC play protocol (Ousley *et. al.* 2012), which contains *five* stages: Greeting: the examiner greets the child; Ball: the examiner plays a game where she rolls the ball back and forth; Book: the examiner presents a book and invites the child to read it with her; Hat: the examiner places the book over her head and pretends the book is a hat; Tickle: the examiner plays a tickling game with the child. In each play session, the child and the examiner sit at opposite sides of a tabletop where the ball and the book are presented. The child's upper body movements are recorded by a tripod-mounted Basler camera with 1920x1080 resolution and 60 FPS. We perform gaze analysis on the video data on a frame-by-frame basis.

b. Methodology

We designed and trained a deep convolutional neural network to perform the gaze analysis. Our network takes in three inputs, which are an image containing the person whose gaze we want to predict, the face location of that person, and a close-up face image of that person. Our network has three outputs, a saliency heat map similar to that by (A. Recasens et. al.), a gaze angle prediction, and a final gaze target prediction indicating where the person is looking at. We use the three main sources of data described previously to train our neural network. Since our work is currently under review and pending publication, details related to implementation, network architecture, training, and experiments design etc. will not be discussed here.

Results

i. Head Pose Analysis

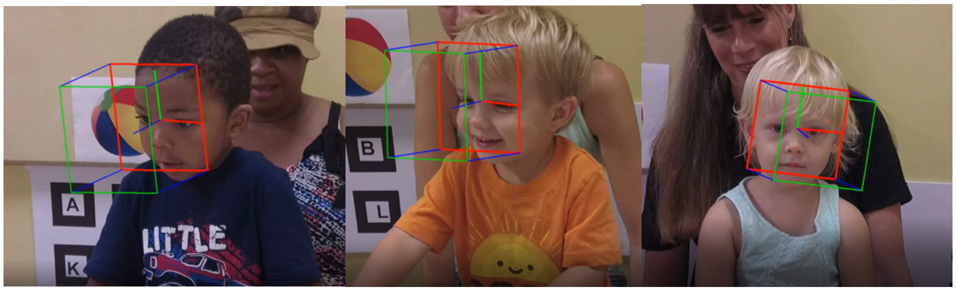

Figure 3. Example visualizations of head pose estimations

Shown in Figure 3 are some visualizations of children's head poses. The blue axis points in front of the children's head, the green axis points upwards, and the red axis points to the right. There is also a cube shown, where the red edges constitute the base of the cube, blue edges are the pillars, and the green edges make up the top of the cube. These are some of the successful cases where the head poses are visually correctly captured. We can observe that these visualizations give a general prediction of where the children's head is pointing to, though some errors still exist. For example, in the right most image, the child's head is almost pointing directly into the camera, but the pose estimation tells us that she's looking towards the bottom right. We notice this situation and believe that this is because there are few differences in facial landmark detections between looking directly into the camera and looking slightly towards the bottom right. In other words, it is inherently difficult to distinguish based on facial landmarks alone whether the child is looking straight into the camera or is looking only slightly away from the camera. Another limitation of our method is that it heavily depends on openpose facial landmark detections (Z. Cao *et. al.*) (S.-E. Wei *et. al.*) (C. Qian *et. al.*). Once openpose gives noisy detection results, our head pose estimation also suffers.

These issues and limitations lead to potential future works. We are interested to explore other methods to predict head poses more accurately, potentially leveraging the state-of-the-art deep learning techniques. We also want to find ways to improve facial landmarks detections, which our currently method relies on.

Overall, despite the issues, our method can capture the general direction along which a child's head is pointing to.

ii. Gaze Analysis

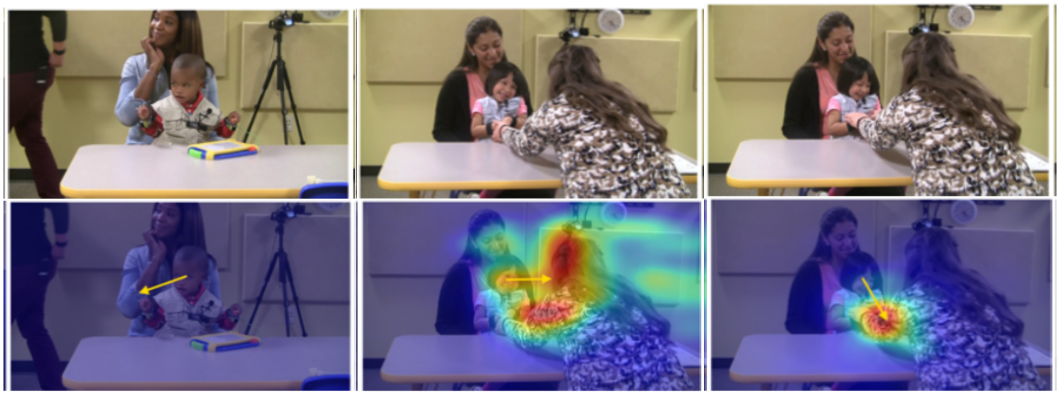

Figure 4. Gaze analysis results on MMDB dataset

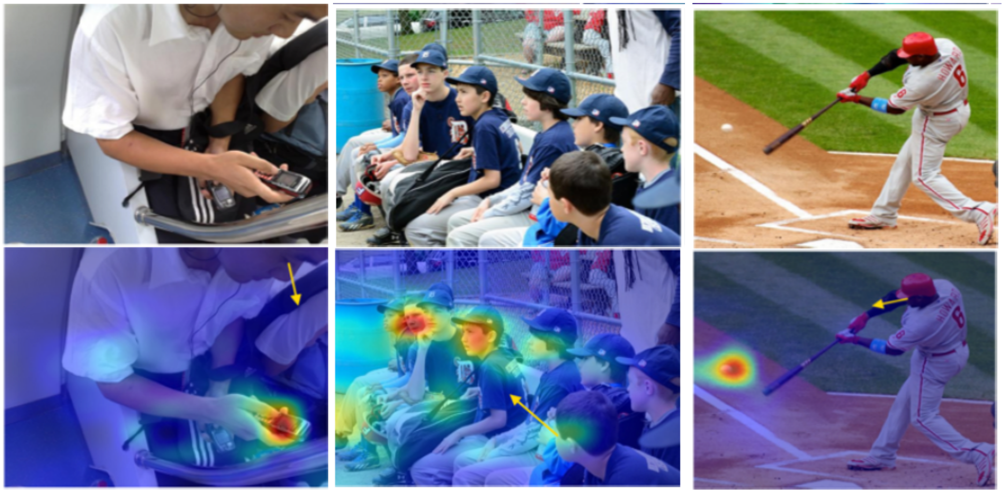

Figure 5. Gaze analysis results on MIT GazeFollow Dataset

Figure 4 and Figure 5 show some of the successful results of gaze detection on both our MMDB dataset and the MIT GazeFollow Dataset. Our model generates a heat map where the red areas are the areas that the person is most likely looking at. The blue areas are the least likely ones that the person is looking at. When potential salient objects are present inside the frame, our model correctly gives these objects higher probabilities of being observed. When the person is looking outside the frame (left column in Figure 4), our model generates zero probabilities for all pixel inside the image.

Moreover, Figure 4 shows three different cases: the child is looking outside; the child is looking at the examiner; the child is looking at the object presented. Under each of the cases, our model correctly predicts where the child is looking at. This shows that our method is robust in predicting children's gazes under different circumstances, and thus will potentially be very useful for understanding child-adult interactions.

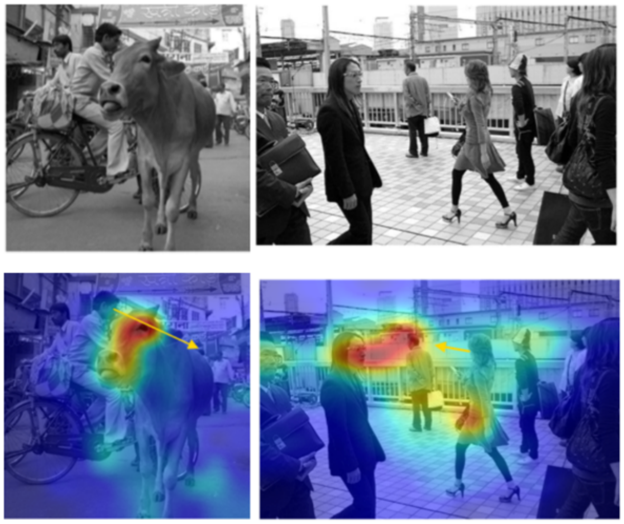

Figure 6. Failure cases

There are some challenging cases as well, as shown in Figure 6. We can see that the person in the left image is actually looking somewhere behind the cow, instead of at the cow's head. The lady in the right image is looking straight ahead instead of looking at the people in front of her. These cases are challenging because our model currently lacks the ability to understand depth, and there are some occlusions that can mislead our detector. However, this is a very exciting future work direction. Once we can incorporate depth understanding and deal with occlusions, it is very likely that the gaze detection accuracies will be greatly improved.

In addition, another potential direction of future work is to use our gaze prediction model and compare gaze shifts between children with and without autism. We are interested to see whether there exist any differences of gaze shift patterns between these children.

Conclusions

In this thesis, I focused on two areas of research, namely head pose detection and gaze detection. Under the guidance of Dr. Jim Rehg, Dr. Agata Rozga, and PhD student Eunji Chong, I took part in and assisted in developing novel methods to enable researchers to study and understand children's social attention, which is highly relevant when it comes to studying and treating autism. We designed and tested an algorithm to determine children's head pose using data generated from openpose. We also developed a deep learning model to predict where children are looking during a tabletop social interaction. I presented qualitative results of our methods and noted future directions for this work.

This research experience is very meaningful to me as an undergraduate research assistant. Not only did I figure my research interest in human behavior analysis, I also learned many valuable skills, ranging from learning what research is all about to reading and digesting other people's research work, from understanding basic computer vision topics such as stereo correspondences to using specific deep learning tools such as Caffe. This 1.5-year research experience sets the stage for my early research life, and will certainly be very helpful in my future research career.

Acknowledgments

The undergraduate research experience in Dr. Jim Rehg's lab has opened the world of behavioral imaging to me, and I have found my strong interest in this field. I want to specially thank Dr. Jim Rehg, Dr. Agata Rozga, Eunji Chong, and Nataniel Ruiz for your guidance and for always being willing to help. This thesis concludes my work as an undergraduate researcher, and I will take what I have learned with my passion to expand field of behavioral imaging in next stages of my life. It has been a fantastic journey at Georgia Tech.

Bibliography

J. Hashemi *et al*., A computer vision approach for the assessment of autism-related behavioral markers, 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), San Diego, CA, 2012, pp. 1-7. doi: 10.1109/DevLrn.2012.6400865

Min, C.-H. (2017). "Automatic detection and labeling of self-stimulatory behavioral patterns in children with Autism Spectrum Disorder." Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual InternationalConference of the IEEE, Jeju Island, South Korea, South Korea.

Abbas, H., et al. (2017). "Machine learning approach for early detection of autism by combining questionnaire and home video screening." CoRR **abs/1703.06076**.

Rajagopalan, S. S. (n.d.). “Computational Behaviour Modelling for Autism Diagnosis.” Proceeding of ICMI '13 Proceedings of the 15th ACM on International conference on multimodal interaction. Retrieved September 9, 2017.

Mohamed, A. O., et al. (2006). “Attention analysis in interactive software for children with autism.” Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility. Portland, Oregon, USA, ACM**:** 133-140.

E. Chong, et al. (2017, Sept). "Visual 3D Tracking of Child-Adult Social Interactions". IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). Lisbon, Portugal.

Z. Cao, T. Simon, S. Wei, and Y. Sheikh. “Realtime multi-person 2d pose estimation using part affinity fields.” arXiv:1611.08050, 2016.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. “Convolutional pose machines.” In IEEE Conference on Computer Vision and Pattern Recognition, 2016

C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. “Realtime and Robust Hand Tracking from Depth.” In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1106–1113, 2014.

A. Recasens*, A. Khosla*, C. Vondrick and A. Torralba. “Where are they looking?” *Advances in Neural Information Processing Systems* (NIPS). 2015

Funes Mora, K.A., Monay, F., Odobez, J.M. “Eyediap: A database for the develop-ment and evaluation of gaze estimation algorithms from rgb and rgb-d cameras.” Proceedings of the ACM Symposium on Eye Tracking Research and Applications, ACM (March 2014)

Gu, J., Yang, X., De Mello, S., Kautz, J. “Dynamic facial analysis: From bayesian _ltering to recurrent neural network.” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)

Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaro_, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al. “Decoding children's social behavior.” Proceedings of the IEEE conference on computer vision and pattern recognition. (2013) 3414{3421}

P. Mundy, C. Delgado, J. Block, M. Venezia, A. Hogan, and J. Seibert, “Early social communication scales (escs),” Coral Gables, FL: University of Miami, 2003.

O. Y. Ousley, R. Arriaga, G. D. Abowd, and M. Morrier. “Rapid assessment of social-communicative abilities in infants at risk for autism.” Technical Report CBI-100, Center for Behavior Imaging, Georgia Tech, Jan 2012. Available at www.cbi.gatech.edu/techreports.