Visual 3D Tracking of Child-Adult Social Interactions

Eunji Chong*, Audrey Southerland*, Abhijit Kundu*, Rebecca M. Jones†, Agata Rozga*[1] and James M. Rehg*[1]
*Center for Behavioral Imaging and College of Computing, Georgia Institute of Technology, Atlanta, GA.
†Weill Cornell Medicine, Center for Autism and the Developing Brain, White Plains, NY.

***Abstract*—We describe an approach to continuously capture children's 3D head pose and location during a tabletop social interaction with an adult examiner. Our approach, called face plus context, utilizes a fixed room camera in conjunction with a head-worn camera on the examiner to simultaneously capture the child's face along with the toys and social partners that provide context. Our system performs head tracking and pose estimation along with multi-target tracking to provide 3D localization and disambiguate identity. We evaluated our method on a dataset of 16 children, including both typically developing and autistic children. We present encouraging results for measuring children's social behaviors, along with validation results using an IMU.**

I. Introduction

The problem of capturing and analyzing social behaviors during naturalistic interactions is an important and challenging task with a broad range of applications in automated behavior analysis and social robotics. For example, the use of sensors and machine learning methods to analyze social behaviors has emerged recently as a promising technology for understanding and treating developmental conditions such as autism [1], [2]. Moreover, in the area of human-robot interaction, there is a long-standing interest in creating social robots with nonverbal communication capabilities [3], [4]. While motion capture technology can be used to record social behavior, it requires the use of professional actors as marker-based methods are too invasive to capture spontaneous naturalistic interactions between multiple people. This is particularly true in the case of *children's* social behaviors. There has been a limited amount of prior work on the analysis of children's behavior from video [5], [6], [7]. These works have tended to focus on facial expression analysis or the detection of specific behaviors such as eye contact. While several software packages exist for tracking facial landmarks [8], [9], facial expressions are only one element of social behavior. In particular, facial expressions are coordinated with shifts of attention, and attention in turn requires the coordination of *head movement* with the eyes. Head movement provides additional nonverbal communication cues, such head nods and shakes for "yes" and "no." In addition, the 3D location and pose of the head identifies the portion of the scene that the person is facing and is likely to be attending to. It follows that the ability to track head pose and localize heads in 3D is a key capability for social behavior capture and analysis.

[1]Rozga and Rehg share senior authorship of this work.

While there have been a variety of prior works on head tracking from video [10], [11], [12], few of these methods are designed to work with multiple video cameras, and as a consequence they are limited to relative head pose and cannot localize the head in a 3D room coordinate system. There have been a few works on multicamera head tracking [13], [14] along with works that focus on more general multicamera reconstruction which can recover 3D head location [15], [16]. Unfortunately, these methods are not suitable for the large-scale capture of children's social interactions due to the expense and complexity of their multi-camera setup. It is commonplace to record assessment and therapy sessions with children using a single room camera, but is not practical to capture with the large number of cameras needed for dense reconstruction.

As an alternative, we have developed a practical and effective approach to capturing children's social behaviors which combines a single room camera with a wearable camera worn on an adult social partner. This setup reflects the fact that measurement of a child's behavior frequently occurs via interactions with an adult, such as a clinician, therapist, teacher, or caregiver. We call this setup *face plus context* because the head-worn camera of the adult examiner provides almost continuous capture of the child's face and head, while the room camera provides access to the social context, and supports localization in a 3D room coordinate system.

We present a novel multi-camera system for 3D head tracking and localization which is suited to the face plus context scenario. We combine continuous tracking and calibration of the head-worn camera with 3D localization and pose estimation of all heads in the social scene. Our system uses state-of-the-art methods for face tracking and head pose estimation combined with multi-target tracking to provide 3D localization and disambiguate identity in the case where there are multiple people present.[2] Our system automatically tracks all heads in the scene and reconstructs the pattern of social interaction between the participants based on head movement. This is a first step towards a more comprehensive 3D social capture system which will incorporate gestures and gaze shifts in addition. This work makes the following contributions:

[2]For example, it is common for very young children to sit on a parent's lab during an assessment, with the result that the parent's head becomes a distractor for the task of tracking the child.

978-1-5386-3715-9/17/$31.00 ©2017 IEEE

- We present the first 3D multi-head tracking and capture system for the *face plus context* measurement scenario, which is designed to be applicable to a wide range of child assessment scenarios including behavioral screening, therapy, evaluation, and skill training contexts.
- We provide the first experimental results for 3D head tracking of children and their adult social partners during naturalistic face-to-face social interactions.
- We demonstrate that head shifts detected using our 3D head tracking approach are a useful step in detecting gaze shifts, based on experiments including both typically developing children and children with autism. Our results are promising in light of the difficulty of deploying standard gaze tracking technology in naturalistic social scenarios.

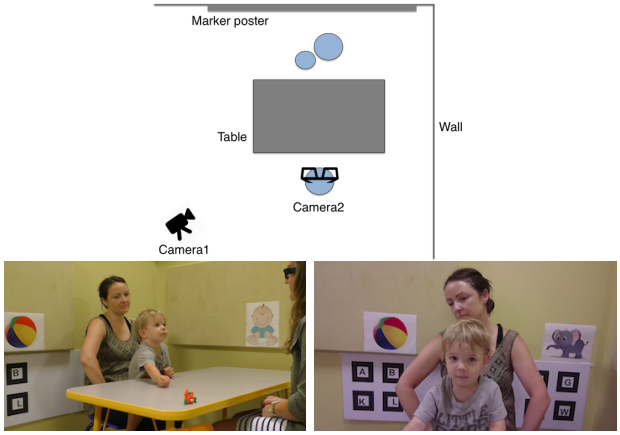

Fig. 1. Our setup "face plus context" and sample images from each camera.

II. Related Work

There are three main areas of prior work which are relevant to this paper: 3D head pose estimation, head pose-based social sensing, and video-based measurement of children's behaviors.

3D Head Tracking: There is a large body of work on face localization and tracking from video [12]. Single camera tracking systems can estimate the head pose relative to the camera and track the scale of the face, but are unable to localize heads in 3D reliably. This includes methods based on facial landmarks [8], [17] and other alignment methods [18], [19], as well as full 3D head models [20], [21]. If two or more cameras are used, then it is possible to recover the complete 3D location and pose of the head. Since children frequently lean towards the objects and people that they are interacting with, the 3D head location is a valuable cue for social understanding. There have been several prior works on multi-camera head tracking [22], [13], [14], [23], [24]. One area of prior work, of which [14] is representative, use multiple cameras to estimate where a user is looking in a smart environment application. The goal is to determine which of a discrete set of gaze targets is being attended to. Other works in meeting room understanding, described below in more detail, use head pose (along with audio cues in some cases) to understand patterns of conversation, floor holding, and other communicative acts. None of these works produce continuous measurements of 3D head location and pose, a capability provided by our approach which creates the possibility for more fine-grained measures of social behavior. Our use of multiple hypothesis tracking (MHT) is related to the work of Benfold and Reid [25], which demonstrated the ability to track multiple pedestrians using a single camera in an outdoor surveillance application and determine where they were looking. In contrast, our MHT approach can produce full 3D estimates of head locations and pose over time from multiple cameras.

Head Pose-Based Social Behavior Sensing: Prior works on social sensing have utilized estimates of head pose as a cue for social attention, although none of these works have addressed children's behavior. One representative problem is understanding patterns of conversation and attention between adults during business meetings [23], [24]. In comparison, our task requires the consideration of a broader range of gaze targets, as children will interact with toys as well as their social partners. Moreover, these works assume that participants don't approach each other closely and can therefore be tracked without a multiple hypothesis tracking framework. In our case, when children sit on their parent's lap or approach the examiner it creates a more challenging tracking problem. Another line of work [26], [27] uses head orientation to understand social group formation and detect specific types of social interactions, such as conversational group detection based on f-formation theory from social psychology. In contrast, we are interested in detecting specific behaviors of interest, such as a shift in gaze based on head movement, not in classifying the type of interaction.

Our use of a wearable camera connects us to other works that study social attention from a first person vision perspective. The closest work is [28], which shows that patterns of social interaction within a group can be classified based on the change in the head poses of the group over time, as captured by a wearable camera. This work assumes that there are many people looking at the same gaze targets, and requires that the targets be visible to the wearable camera. In contrast, our case is primarily a dyadic interaction, and often the child's gaze targets are not visible in the examiner's camera (requiring the use of a room camera). Another representative work identifies social "hotspots" that arise when many people, all wearing head-mounted cameras, look at the same location [29]. This approach does not handle behaviors like eye contact, and it would require instrumenting the child, which would limit the applicability of the method considerably. One point that we have in common with all of these prior works is an assumption, often called "center bias," that head orientation is a powerful indicator of one's direction of attention [30]. Our experimental results confirm that this bias is a useful cue in analyzing children's attention as well.

Vision-based Measures of Children's Social Behavior: Other works have demonstrated the ability to measure aspects of a child's social behavior using vision. In dyadic face-to-face interactions, it has been shown that facial expression analysis

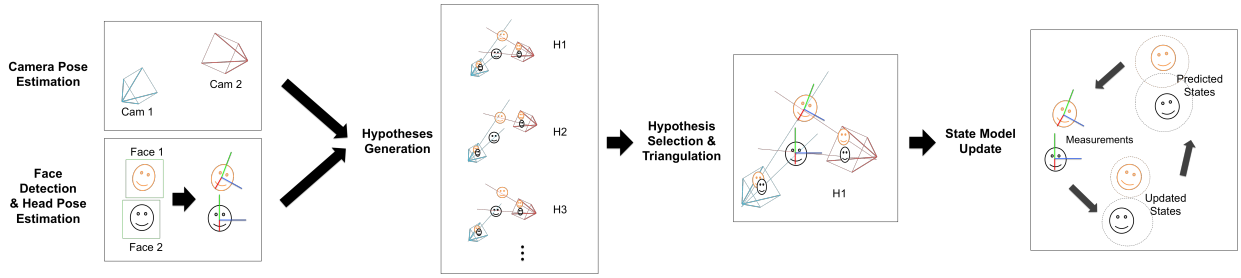

Fig. 2. **System overview**: First, each camera pose is estimated based on the patterns placed on the wall (Sec. III-B). Also, face is detected and head pose w.r.t. each camera is estimated using facial landmark alignments (Sec. III-C). Finally, the most likely combination is used to update head state models (Sec. III-D).

can be used to discover synchrony between an infant and their caregiver [7]. In earlier work, we demonstrated the ability to detect moments of eye contact using a wearable camera [5]. We go beyond these works by addressing children's attention to objects in addition to faces. A small number of works address activity recognition in children from video [31], [32], and while this has no direct bearing our work it is part of the broader story for behavior capture using sensors.

III. Approach

Our sensing approach for the *face plus context* capture scenario utilizes two cameras to maximize coverage of the child's behavior while facilitating automated measurement. The setup is illustrated in Fig. 1. The first camera is statically mounted and records the scene context. The second camera is concealed in a pair of glasses worn by the examiner and captures the child's face (see Sec. III-A for details). Our analysis pipeline is illustrated in Fig. 2. The first step is to calibrate the two cameras so they can be combined to localize the heads. Since the wearable camera is in constant motion, it is continuously-calibrated using AR Tags (see Sec. III-B). Separately, faces are detected and relative head pose is estimated from each camera (see Sec. III-C). The set of detections is processed by a multiple hypothesis tracking algorithm (see Sec. III-D) which maintains the identity of each subject over time and fuses both camera views to produce 6 DOF location and pose for each head. Head movements are then analyzed to predict attention to elements of the scene (see Sec. III-E). We now describe these elements in more detail.

A. Face Plus Context Setup

Our goal is to support the capture of a child's behaviors in the context of interaction across a tabletop, which is a standard assessment paradigm in psychology. The setup is illustrated in Fig. 1. Young children may sit on a parent's lap, with the result that the parent is frequently visible in both cameras. The tabletop serves as a convenient surface for toy manipulation and also helps to constrain the child's movement. In this setting, the examiner administers a set of play protocols by interacting with the child using a set of toys. In the case of assessing children with autism, the protocols are designed to elicit social behaviors such as eye contact or pointing, which are key elements of joint attention. This procedure makes it possible to tap a variety of social behaviors in a well-defined context.

We utilize two cameras to capture the child's social behaviors. The first camera is mounted on a tripod and is placed at an angle that covers the table and the people in the scene. It provides capture of the entire social scene from a fixed vantage point, ensuring that the child and any toys they are interacting with will be visible at all times. It produces a lower resolution image of the child, but is still useful for localizing the child in 3D. The second camera is inconspicuously located in a pair of glasses (Pivothead SMART) worn by the examiner. The glasses can be filled with a prescription or the lenses can be removed to provide an unobstructed view. Since the examiner naturally maintains an orientation to the child at all times, this camera provides continuous high resolution capture of the child's face. Moreover, the position of this camera facilitates the detection of eye contact via the method of [5]. In order to support continuous calibration of the head-worn camera, a single poster board with black and white patterns (AR Tags) printed on it is attached to one wall so that it is visible to both cameras. Note that the addition of the wearable camera and poster board are all that is needed to convert a standard psychology assessment into our face plus context scenario.

B. Camera Pose Estimation

In order to reliably estimate camera pose from different views, we use a marker-based pose estimation approach using an open-source Augmented Reality (AR) software called ARToolKit.[3] This utilizes a square pattern (marker) designed so that it can be detected easily and its 6-DOF pose relative to the camera can be computed reliably. Since the pattern can be occluded by the participants, we use 8 square marker patterns, printed on a single foam board which is mounted on the wall. Fig. 3 illustrates this step. Pose can be estimated from one or more detected markers. This method requires knowledge of the camera's intrinsic parameters. We perform checkerboard calibration of the intrinsics at the start of the session. For the wearable camera, calibration can be done once and reused across different sessions. The room camera intrinsics need to be recalibrated if the camera zoom or other settings change.

We considered using a structure-from-motion approach and we evaluated several different SLAM methods [33], [34] on

[3]https://artoolkit.org/

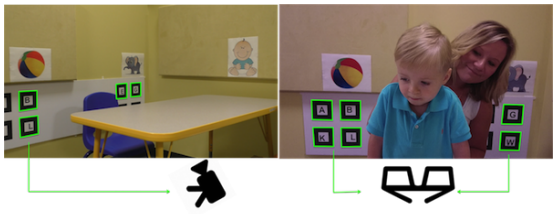

Fig. 3. **Camera pose estimation**. Green boxes show detected markers used for pose estimation of a room camera and a wearable camera.

our dataset. Since the videos captured by the POV camera contain abrupt motions, a narrow camera field-of-view, and a dynamic scene, these structure-from-motion methods had difficulties in identifying stable features, and primarily tracked points on the moving humans, which is not useful for camera pose estimation.

According to the reported accuracy of marker-based pose estimation by ARToolKit [35], our setup is within the range of permissible error (2 degrees). In future work, we could explore the use of bundle adjustment with the initial pose estimate to improve the accuracy further.

C. Face Detection and Head Pose Estimation

In each view, faces are detected using the Omron OKAO library.[4] For each detected face, we use IntraFace [8] to find and track facial landmarks and we use the Perspective-n-Point algorithm to estimate head pose relative to the camera coordinate frame. The result is a set of 5-DOF measurements, 3 for the head pose R_w^{face} in (1), and 2 for the face ray $X(\lambda)$ in (2), since the depth is unknown:

$$R_w^{face} = R_w^c \times R_c^{face} \tag{1}$$

gives the head pose R_w^{face} in world coordinates based on the estimated camera pose R_w^c and the estimated face pose w.r.t. camera R_c^{face}.

$$\begin{aligned} P^+ &= P^T(PP^T)^{-1} \\ X(\lambda) &= P^+ x + \lambda C \end{aligned} \tag{2}$$

gives the ray $X(\lambda)$ passing through the face center (i.e. face ray), based on the camera center C, the camera projection matrix P, and the image pixel location of the face center x.

D. Data Association and Tracking

One of the challenges in analyzing social interactions is to track each person's head consistently without confusing one participant for another (i.e. ID-switch errors). We now describe the multi-target tracking framework we developed to solve this problem. We start by describing the 6-DOF head state model which defines the state space for tracking and supports the fusion of observations from both cameras over time.

[4]https://www.omron.com/ecb/products/mobile/okao01.html

1) Head State Model: For each head, we define a state

$$\begin{aligned} S = (&x, y, z, \dot{x}, \dot{y}, \dot{z}, \ddot{x}, \ddot{y}, \ddot{z}, \\ &q_1, q_2, q_3, q_4, r_1, r_2, r_3, \dot{r_1}, \dot{r_2}, \dot{r_3})^T, \end{aligned} \tag{3}$$

where x, y, z is location, $\dot{x}, \dot{y}, \dot{z}$ is velocity, $\ddot{x}, \ddot{y}, \ddot{z}$ is acceleration, q_1, q_2, q_3, q_4 is rotation in quaternion, r_1, r_2, r_3 is angular velocity, and $\dot{r_1}, \dot{r_2}, \dot{r_3}$ is angular acceleration. Then, each state is tracked with an Extended Kalman Filter following the process update model in (4) and measurement update model in (5). The process update model is

$$\begin{aligned} S_t^- &= f(S_{t-1}, w_{t-1}) \\ P_t^- &= A_t P_{t-1} A_t^T + W_t Q_{t-1} W_t^T \\ A_t &= \frac{\partial f}{\partial s} \\ W_t &= \frac{\partial f}{\partial w}, \end{aligned} \tag{4}$$

where S_t^- is predicted state, P_t^- is predicted state covariance, w is process noise, Q is process noise covariance, f is the function that projects the positional data and angular velocity linearly, except for q that is updated through a quaternion multiplication with d, the difference caused by angular velocity. The measurement update model is

$$\begin{aligned} h(S_t, v_t) &= (x, y, z, \frac{q}{|q|}) + v_t \\ K_t &= P_t^- H_t^T (H_t P_t^- H_t^T + V_t R_t V_t^T)^{-1} \\ S_t &= S_t^- + K_t(m_t - h(S_t^-, 0)) \\ P_t &= (I - K_t H) P_t^- \\ H_t &= \frac{\partial h}{\partial s} \\ V_t &= \frac{\partial h}{\partial v}, \end{aligned} \tag{5}$$

where h is the function that extracts position and rotation from the state vector, v is measurement noise, K is Kalman gain, R is measurement noise covariance, m is taken measurement, P is updated covariance matrix. Note that we let x, y, z, q be measurables, and how this measurements are generated and associated is described in detail in the following section.

2) Data Association: Initially, per view and per detected face, we have a measurement of 5 degrees of freedom,

$$m = (X(\lambda), q1, q2, q3, q4), \tag{6}$$

where q is quaternion of R_w^{face} in (1) and $X(\lambda)$ is from (2). For instance, if two faces are detected in camera 1 and two faces are detected in camera 2, there should be four individual m's at that moment. Then, we define a cost function $C(m, S)$ between a pair of m's and a state S, considering both geometric and appearance constraints as follows:

$$C(m, S) = w_g \times C_g(m, S) + w_a \times C_a(face, S), \tag{7}$$

where w_g and w_a are weight parameters, C_g is Mahalanobis distance between m and S, which is calculated using P^- in

(4), and C_a is classification score of a detected face to the state.[5] With this cost function, all m's are assigned to a state S independently per camera. Meanwhile, an m with a cost above a certain threshold is discarded to remove spurious detections, incorrect pose estimations, etc. Note that this matching is done per view, such that an m from a given view is associated with one state S exclusively (or discarded), but each state can have multiple m's from different views. As a result, if one state is assigned two m's, they are triangulated to obtain 3D position, and the rotation is interpolated using the slerp algorithm. If only one m is assigned to a state, the closest 3D point on the line is selected. This is the m_t used in the measurement update model (5). To reflect the confidence of the final measurement, the measurement noise covariance R in (5) is reduced as the number of m's used increases.

E. 3D Scene Estimation

In addition to 3D head tracking, we also estimate the location of the table and the toy. We estimate table pose by using the same approach as in camera pose estimation. In the beginning of each session, a known square pattern is put on the table for a few seconds. This is sufficient to retrieve table pose as the table does not move throughout session. Additionally, the play protocol we utilized (Sec. IV-B) incorporates a set of toys presented at one location on the table (Fig. 5), for which the estimated table pose can be used as well. Additional work could be done to refine the toy locations further, for example using pose estimation from a 3D model [36] or triangulation with a custom toy-object detector.

IV. Data

In this section, we describe the data used in the paper and how it was collected and processed.

A. Participants

Participants were recruited and data was collected at Georgia Tech (GT) and Weill Cornell Medicine (WC). Our dataset consists of eight sessions from typically developing (TD) children and eight sessions from children with autism. Eight TD children (3 female) with no known diagnosis of social, developmental, or communication delays were recruited at GT via community advertising and a parent mailing. Eight children with a diagnosis of ASD (4 females) were recruited by WC. The diagnosis of ASD was confirmed prior to participation by a licensed clinician. TD participants were between 20 and 36 months of age (mean age = 30.8 months), and ASD participants were between 32 and 60 months of age (mean age = 43.8 months). All participants completed play-based assessments during a single visit.

B. Play Protocol

All participants completed a modified version of the Early Social Communication Scales [37], a semi-structured, examiner-directed assessment of nonverbal communication skills in young children. The child is seated, sometimes on a caregiver's lap, at a small table across from the examiner. The examiner presents several different toys and activities to the child, selected because of their potential to elicit joint attention (using gaze and gestures to share the experience of objects or events with a social partner) and requesting (using nonverbal behaviors to elicit aid in obtaining objects or events). The toys include: a) three small wind-up mechanical toys, b) three hand-operated toys, c) a small car and a ball that will roll easily across the table, d) a book with large distinct pictures on its pages, and d) colorful posters positioned on the walls to the left, right and behind the child. The ESCS administration takes about 15-25 minutes to complete.

The present analysis focused on the object spectacle toys, which include three unique wind-up toys and three hand-operated toys, including a trapeze monkey, a balloon, and a spintop. For each of the six object spectacle tasks, the examiner places one of the six toys in the corner of the table to her left, activates the toy for about 5-10 seconds, then allows the toy to remain inactive for about 5-10 seconds. Per the scoring rules in the ESCS manual, any time the child shifts their gaze from the active toy to the examiner's eyes and then back to the toy, they are credited with engaging in initiating joint attention. When the toy becomes inactive, similar shifts of attention are credited as initiating behavior regulation (i.e., requesting).

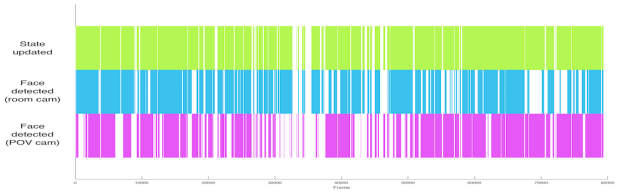

Fig. 4. State update frequency and face detection frequency.

C. Annotation

Videos of the ESCS assessments from the room camera and the POV camera were used for manual coding by trained raters to identify each moment when the child was looking at the toy or making eye contact with the examiner. The start and the end of each spectacle toy presentation was coded as well. Coders used Mangold International's Interact annotation software[6] to identify these moments and mark the onsets and offsets at the video frame level.

V. Results

In this section, we report quantitative and qualitative evaluations of our approach.

A. Head Tracking Statistics

We first evaluated the performance of our system by calculating overall tracking statistics. We ran the tracker on 15-25

[5] In the very first frame, each state is initialized by assigning a face to S_{parent} as -1 or to S_{child} as 1, and a linear regressor is trained online subsequently.

[6] https://www.mangold-international.com/en/software/interact

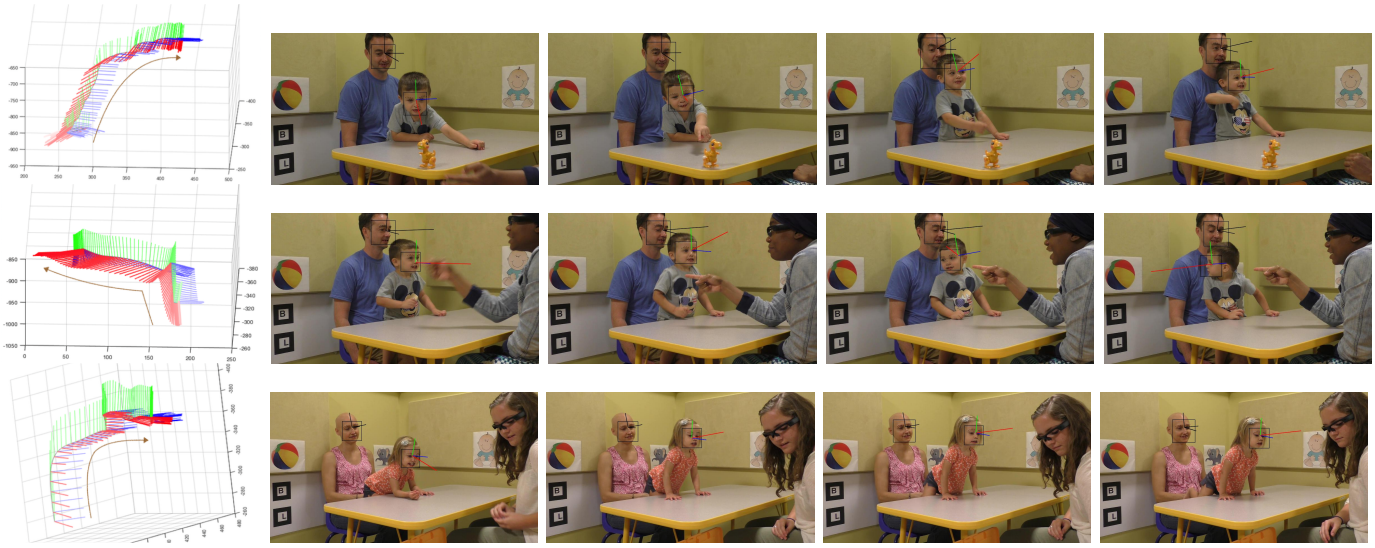

Fig. 5. **Social signals captured by our system**. Three axes of red, green, and blue represent child's head pose. First column shows the 3D head trajectory during 4 seconds in the direction of the arrow. Row 1: Wind-up toy is presented and the child is requesting the examiner to give it to him by making eye contact. Row 2: Examiner is pointing to a poster and the child is following. Row 3: Examiner is choosing a toy and the child is peeking over the table.

minute-long sessions of 16, which resulted in 1,152,600 total frames. Among these, the child head state model has been updated with new measurements 87.6% of the time, meaning that the child's face was successfully detected in 87.6% of the frames. Within this detection rate, 80.1% of face detection is from the room camera and 67.5% is from the POV camera. The reasons why it is lower in the POV camera are: A) limited field of view, B) camera viewpoint change, and C) motion blur. While B) naturally occurs following wearer's head movements, A) and C) can be improved with the advances of wearable camera technology. Despite these challenges, the inclusion of a wearable camera adds great value to the system overall. This can be seen in Fig. 4, which identifies frames in which face detection failed for the room camera (vertical white lines in the middle row) but succeeded for the POV camera (last row) leading to greatly increased state updates (top row is more dense than either face detection row alone). This finding confirms the effectiveness of our face plus context setup.

B. Qualitative Results

Children show a wide range of head movements in the course of ESCS. This can be observed by visualizing the 3D head trajectories and their reprojections on the input videos. As shown in Fig. 5, children use a broad range of head and body movements to communicate with people and achieve their goals. As a consequence, our videos comprise a rich, dynamic, and densely-sampled (60 Hz frame rate) database of social motion in a variety of contexts.

To further evaluate our method, we collected videos during a short ESCS session in which an IMU sensor was tightly-attached to head. The results from this experiment (see Fig. 6) indicate that the head motions measured by the IMU sensor and follow a very similar pattern to our head tracking system (mean error = 0.12, variance = 0.08, in radians).

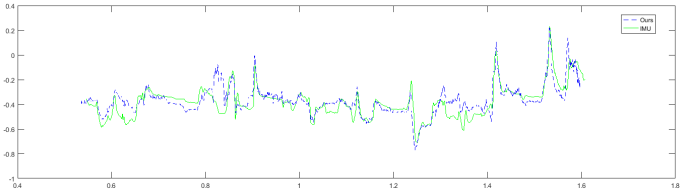

Fig. 6. IMU vs. video-based head tracker.

C. Detection of Gaze Shift for Joint Attention

To assess the viability of using head pose estimates to measure gaze shifts, we devised a simple detector for moments when the child looks at a spectacle toy and subsequently makes an eye contact with the examiner. This type of gaze shift is known as initiating joint attention (IJA). IJA is more frequent in TD than in ASD and is correlated with language outcomes. During ESCS object spectacle tasks, toys are presented at a known 3D location at the corner of the table. We trained a classifier to analyze the head tracking data and detect the child's shifts of attention from the toy to the examiner. Specifically, we identified time segments in which the child's yaw angle decreased (segments from cyan up to red point of Fig. 7-2), and created features using the distances from child's head z-vector to the examiner and toy (green and red curve of Fig. 7-1). Each segment was labeled as an IJA shift or not based on the ground truth annotations. 10% of the 1665 total segments are IJA shifts. We trained a binary SVM classifier using 20% of the samples, obtaining the ROC curve in Fig. 7-3. The detector performed equally well on both diagnostic groups (AUC score 0.78 for TD, 0.8 for ASD). Moreover, the predicted total number of gaze shifts per subject is correlated with the ground truth gaze shift counts (p-values: TD < 0.005, ASD < 0.0077, All < 0.0007).

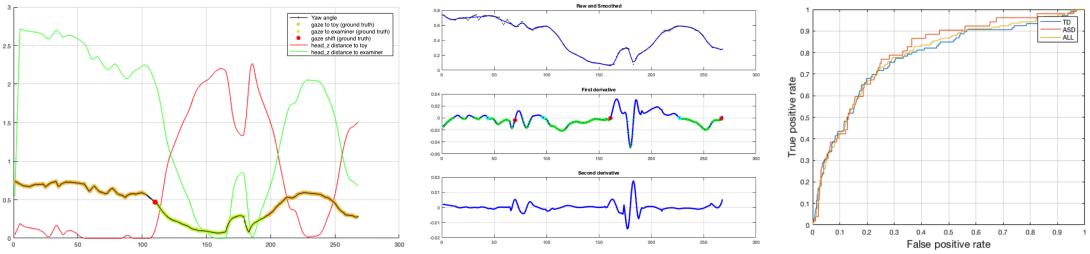

Fig. 7. **Gaze shift detection**. Left figure shows how measurements change over time when there is a gaze shift from toy to examiner. Middle figure shows how a segment is selected at testing time for gaze shift detection. Right figure is an ROC curve showing our gaze detector's performance.

D. 3D Attention Map

We developed a method for generating a 3D attention density map to support additional visualization and understanding of our collected measures. We begin by creating a 3D volumetric scalar field to represent the gaze density at any 3D point in the interaction space. We utilize a gaze model similar to [38], wherein the gaze vector is assumed to lie in a cone-shaped distribution emanating from the center of two eyes, capturing the uncertainty in head pose and eye gaze. The head pose estimate from our system directly gives a 3D vector (Z axis of the head coordinate system) emanating from the center of the eyes in the direction of the front of the head. The uncertainty in the actual gaze direction with respect to the head pose estimate is represented by a Gaussian distribution on a plane normal to the head direction vector.

For computational feasibility, the volumetric scalar field is discretized into 3D voxels (which are analogous to pixels in a 2D image). For speed and memory efficiency, we use a 512^3 voxel array to represent the entire the 3D interaction space. Each voxel stores a scalar score representing the likelihood of gaze at that voxel. The scores are initialized to zero and recursively updated for all head pose estimates. With our cone-shaped gaze distribution model, each voxel's gaze likelihood score is updated according to the voxel's 3D location with respect to the gaze distribution. The scores are simply aggregated across head pose estimates to produce the final cumulative 3D gaze likelihood. Given this representation, we can compute a heat-map on any 3D surface by extracting a slice through the attention density map. An advantage of this 3D volumetric approach is that it can accommodate any other arbitrary gaze model.

Fig. 8 gives an example of the application of our attention model to a sequence in which a child is reaching for a toy. The cone-shaped gaze distribution in 3D space and its reprojection on the image are shown. Fig. 9 illustrates the process of cumulative map generation over a period of time and the final cumulative volumetric map sliced along the table surface, reprojected on a room camera image, and color-coded in heat map color scheme. This example corresponds to a toy presentation period, explaining the high density at the corner where the toy is observed, and a smaller peak in the vicinity of the examiner.

VI. Conclusion

We have presented a novel method for automatically capturing children's head motion in face-to-face naturalistic social interactions. Our flexible camera setup and automated tracking framework makes our system especially suitable for the large-scale capture of children's social interactions. Our method has been successfully applied to 16 sessions that include typically developing children and children with autism, during naturalistic play interactions with an adult examiner. Our experimental results demonstrate that our 3D head tracking approach is effective in measuring children's social behavior, and we present promising results for detecting gaze shifts based on head motion.

Acknowledgment

This work was supported by grant 288028 from the Simons Foundation.

References

[1] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *IEEE Pervasive Computing*, vol. 13, no. 2, pp. 84–87, 2014.

[2] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[3] C. Breazeal, "Role of expressive behaviour for robots that learn from people," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3527–3538, 2009.

[4] J. F. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots–a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 110–125, 2014.

[5] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg, "Detecting bids for eye contact using a wearable camera," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference on*, vol. 1. IEEE, 2015, pp. 1–8.

[6] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 30–35.

[7] D. S. Messinger, M. H. Mahoor, S.-M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant–mother interaction: A pilot study," *Infancy*, vol. 14, no. 3, pp. 285–305, 2009.

[8] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[9] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] J.-S. Jang and T. Kanade, "Robust 3d head tracking by online feature registration," in *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

Fig. 8. **Visualization of our gaze distribution model when a child reaching for a toy.** Bottom row shows it in the reconstructed 3D space and top row shows it by reprojecting the model on actual image.

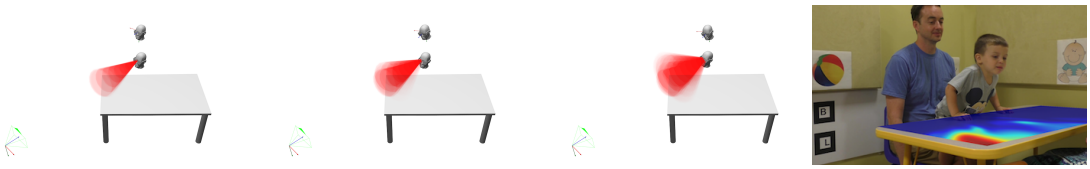

Fig. 9. **Cumulative attention map during a toy presentation period.** The cumulative map generation process and the final heat map along the table plane.

[11] S. Choi and D. Kim, "Robust head tracking using 3d ellipsoidal head model in particle filter," *Pattern Recognition*, vol. 41, no. 9, pp. 2901–2915, 2008.

[12] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[13] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 101–116, 2011.

[14] L. Dong, H. Di, L. Tao, G. Xu, and P. Oliver, "Visual focus of attention recognition in the ambient kitchen," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 548–559.

[15] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[16] E. De Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel, "Marker-less deformable mesh tracking for human shape and motion capture," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[17] J.-G. Wang and E. Sung, "Em enhancement of 3d head pose estimated by point at infinity," *Image and Vision Computing*, vol. 25, no. 12, pp. 1864–1874, 2007.

[18] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[19] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[20] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[21] S. Tulyakov and N. Sebe, "Regressing a 3d face shape from a single image," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 3748–3755.

[22] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 5. IEEE, 2004, pp. V–881.

[23] J.-M. Odobez and S. O. Ba, "A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose," in *International Conference on Multi-Media & Expo (ICME07)*, no. LIDIAP-CONF-2007-033, 2007.

[24] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," in *International Conference on Advances in Visual Information Systems*. Springer, 1999, pp. 765–772.

[25] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3457–3464.

[26] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations." in *BMVC*, vol. 2, 2011, p. 4.

[27] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1212–1229, 2008.

[28] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1226–1233.

[29] H. Soo Park and J. Shi, "Social saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4777–4785.

[30] N. J. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[31] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1470–1474.

[32] P. Wang, G. D. Abowd, and J. M. Rehg, "Quasi-periodic event analysis for social game retrieval," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 112–119.

[33] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[34] R. Mur-Artal, J. Montiel, and J. D. Tardós, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[35] D. F. Abawi, J. Bienwald, and R. Dorner, "Accuracy in optical tracking with fiducial markers: An accuracy function for artoolkit," in *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2004, pp. 260–261.

[36] Y. Konishi, Y. Hanzawa, M. Kawade, and M. Hashimoto, "Fast 6d pose estimation from a monocular image using hierarchical pose trees," in *European Conference on Computer Vision*. Springer, 2016, pp. 398–413.

[37] 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.

[38] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Advances in Neural Information Processing Systems*, 2012, pp. 431–439.