Quantifying Gaze Behavior During Real-World Interactions Using Automated Object, Face, and Fixation Detection

Leanne Chukoskie, Shengyao Guo, Eric Ho, Yalun Zheng, Qiming Chen, Vivian Meng, John Cao, Nikhita Devgan, Si Wu, and Pamela C. Cosman, *Fellow, IEEE*

***Abstract*—As technologies develop for acquiring gaze behavior in real world social settings, robust methods are needed that minimize the time required for a trained observer to code behaviors. We record gaze behavior from a subject wearing eye-tracking glasses during a naturalistic interaction with three other people, with multiple objects that are referred to or manipulated during the interaction. The resulting gaze-in-world video from each interaction can be manually coded for different behaviors, but this is extremely time-consuming and requires trained behavioral coders. Instead, we use a neural network to detect objects, and a Viola–Jones framework with feature tracking to detect faces. The time sequence of gazes landing within the object/face bounding boxes is processed for run lengths to determine "looks," and we discuss optimization of run length parameters. Algorithm performance is compared against an expert holistic ground truth.**

***Index Terms*—Computer vision, eye-tracking, face detection, gaze behavior.**

I. Introduction

The EMERGENCE and refinement of social communicative skills is a rich area of cognitive developmental research, but one that is currently lacking in objective assessments of real-world behavior that includes gaze, speech, and gesture. Gaze behavior is especially important during development. Shared or joint attention provides a means by which adults name objects in the child's field of view [1]. As such, joint attention is an important aspect of language development, especially word learning [2]–[4]. Gaze behavior in children with autism spectrum disorder (ASD) is atypical in terms of social looking behavior, joint attention to objects, as well as the basic timing and accuracy of each gaze shift [5]. Recent studies report that 1 in 45 individuals is diagnosed with ASD [6]. Disordered visual orienting is among the earliest signs of ASD identified in prospective studies of infant siblings [7] and it persists across the lifespan [5]. Humans use gaze as one of the earliest ways to learn about the world. Any deficit in this foundational skill compounds, leading to functional difficulties in learning as well as social domains. Difficulty shifting gaze to detect and respond to these behaviors will lead to a lower comprehension of the nuanced details available in each social interaction. Although several different therapies have been designed to address social interaction [8]–[10], methods of assessing the success of these therapies have been limited.

The outcomes of social communication therapies must be evaluated objectively within and across individuals to determine clinical efficacy. They are typically measured by parent questionnaire or expert observation, both of which provide valuable information, but both of which are subjective, may be insensitive to small changes, and are susceptible to responder bias and placebo effect. Other outcome measures such as pencil and paper or computer assessments of face or emotion recognition are objective, but measure only a subset of the skills required for real-world social communication. These measures are also a poor proxy for actual social interaction. These deficits impact both research and clinical practice.

The recent development of affordable glasses-based eye trackers has facilitated the examination of gaze behavior. The glasses worn by the subject contain two cameras, an eye-tracking camera typically located below the eye, and a world-view camera typically mounted above the eyebrows on the glasses frame. The glasses fuse a calibrated point of gaze, measured by the eye-tracking camera, with the world view. For images displayed on computer screens, many studies have used eye-tracking to examine what portions of the images ASD children attend to [11]–[15]. Eye-tracking glasses can be used during dynamic social interactions, instead of simply on computer screens. The quantification of interactions during real-world activities remains challenging. Analysis of the resulting gaze-in-world video can be done manually, but labeling and annotating all the relevant events in a 30-min video takes many hours. Computer vision and machine learning tools can provide fast and objective labels for use in quantifying gaze behavior. In addition to the uses in ASD

Manuscript received October 20, 2017; revised January 24, 2018; accepted March 14, 2018. Date of publication March 30, 2018; date of current version December 7, 2018. This work was supported in part by the National Science Foundation under Grant IIS-1522125 and Grant SBE-0542013, and in part by the National Institutes of Health under Grant R21/R33 MH096967. An early version of this work will appear at ISBI 2018. *(Corresponding author: Leanne Chukoskie.)*

L. Chukoskie is with the Institute for Neural Computation, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: lchukoskie@ucsd.edu).

S. Guo, E. Ho, Y. Zheng, Q. Chen, V. Meng, J. Cao, N. Devgan, S. Wu, and P. C. Cosman are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: pcosman@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2821566

and other social or communication-related disorders, quantification of gaze for a seated subject in a social interaction can be useful for evaluating a student's engagement with educational material or a consumer's engagement with an advertisement. This technology can also be used for training in various kinds of occupations that involve conversational interactions, such as instructors, police interviewers, TSA agents, passport officers, and psychologists.

Here, we report on a system that uses eye-tracking glasses to record gaze behavior in real-world social interactions. The system detects objects and faces in the scene, and processes these sequences together with the gaze position to determine "looks." The results are compared to the laborious manual coding. The closest past work to ours is [16] and [17], which also involves social interactions and eye-tracking glasses. Their setup is different, as in their work the investigator wears the eye-tracking glasses rather than the subject, and can avoid excessive motion blur and maintain both steady depth and orientation (the child's face does not go into and out of the scene). They do not aim at object detection, and have only one face to detect. Because our naturalistic setup experiences a number of frame-level detection failures, our algorithm compensates for these using filtering to bridge gaps in detections. Also the goal in [16] and [17] is different, as they aim to detect eye contact events rather than looks (extended inspections of regions). Other related past work is [18] and [19] which combined automatic tracking of areas of interest [20] with manual curation by human coders to ensure high detection accuracy.

The rest of this paper is organized as follows. The system operation including methods for detecting faces, objects, and looks is described in Section II, while creation of ground truth (GT) and calibration issues are in Section III. We define evaluation metrics and provide results in Section IV, and conclude in Section V.

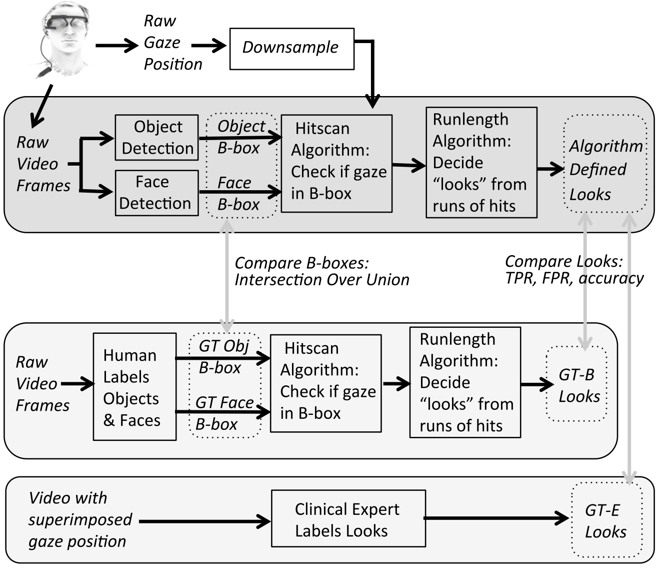

Raw Gaze Position
Downsample
Raw Video Frames
Object Detection
Face Detection
Object B-box
Face B-box
Hitscan Algorithm: Check if gaze in B-box
Runlength Algorithm: Decide "looks" from runs of hits
Algorithm Defined Looks
Compare B-boxes: Intersection Over Union
Compare Looks: TPR, FPR, accuracy
Raw Video Frames
Human Labels Objects & Faces
GT Obj B-box
GT Face B-box
Hitscan Algorithm: Check if gaze in B-box
Runlength Algorithm: Decide "looks" from runs of hits
GT-B Looks
Video with superimposed gaze position
Clinical Expert Labels Looks
GT-E Looks

Fig. 1. Overview of the algorithm and GT formation. B-box stands for a bounding box, GT-B and GT-E are the two types of GT for looks, and TPR and FPR are true/false positive rates.

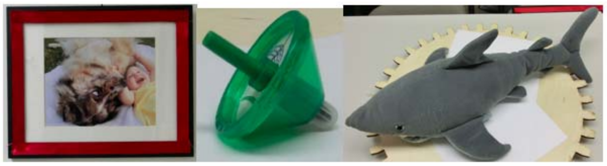


Fig. 2. Objects to be detected: photograph, top, and shark (shown on turntable).

II. Detecting Objects, Faces, and Looks

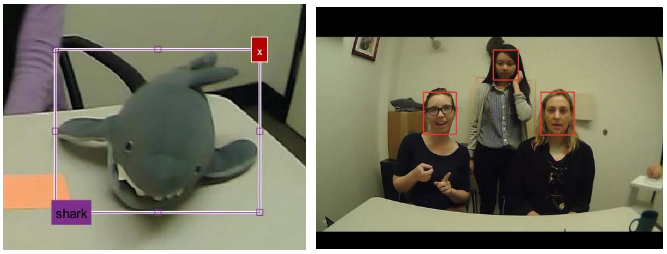
A. System Overview and Data Collection

Fig. 1 presents an overview. The Pupil Labs eye-tracking glasses (Pupil Pro) produce video frames (24-bit color, 720 × 1280, 60 Hz) from the world-view camera and gaze position data at 120 Hz from the eye camera. Gaze data is downsampled to the video frame rate. World-view frames are input to object and face detection modules, whose outputs are sets of bounding boxes. The binary sequence of "hits" and "misses" (corresponding to gaze position inside/outside of the bounding box) is run length filtered to determine looks to an object or face. The lighter gray rectangles depict the formation of the two types of GT. In one approach, humans mark bounding boxes for each object/face in each frame, without gaze position. Boxes and gaze position are then run length filtered to determine looks, called GT-B looks. In the second approach (GT-E looks), an expert neuroscientist directly labels looks by reviewing the video with superimposed gaze position in a holistic way that would be used in clinical practice.

Data were collected to simulate a structured social conversation in a small room with a table and chairs. Each 2.5–3 min interaction began with three undergraduate women (two seated and one standing) across from the participant wearing the gaze glasses. There were five different participants who wore the glasses, all female, and all neurotypical undergraduates, resulting in five videos. After about 15 s, the standing person leaves the room with the glasses wearer following her departure. The remaining three women proceed to play a card game (Taboo) intermixed with looking at each object (frame, shark, and top), and the two faces. The conversation is natural, with laughter. There is a considerable amount of head turning and leaning forward to play cards. Although the glasses wearer is instructed to avoid large, abrupt movements and does not stand up during the interaction, the conversation and interaction proceed naturally otherwise. During the last 30 s, the woman who had previously exited returns to stand behind the two seated participants. The glasses wearer is instructed to look at the person returning to the room.

B. Object and Face Detection

Object detection uses Faster R-CNN [21]. While there are other convolutional neural net approaches to object detection (e.g., [22]) Faster R-CNN is convenient and has good performance. The three objects to be detected are a photograph, a top, and a toy shark (Fig. 2) Training images were collected using world-view frames at different distances and elevation and rotation angles, including occlusions and deformations (for the squeezable shark). In total, there were 15 000



(a) (b)

Fig. 3. (a) Minimum enclosing rectangle for a shark image. (b) Test image with manual GT bounding boxes drawn.

training images. A minimum enclosing rectangle [Fig. 3(a)] was manually placed around each object to serve as the GT during training. We make use of pretrained weights from VGG-16 [23]. We trained each model with 50 000 iterations with a base learning rate of 0.001 and momentum of 0.9. We fine-tuned the models using additional world-view frames. Performance was gauged by the intersection over union (IoU) of the bounding boxes from human labelers and Faster R-CNN outputs.

Face detection and tracking (Fig. 4), executes Viola–Jones face detection [24] once for each video frame, followed by the main function block (which consists of Shi–Tomasi corner detection, eigenfeature tracking with optical flow, tracking points averaging, and adjustment and reinitialization upon failure). The main function block is executed M times for each frame, where M (determined manually) is the number of faces appearing in the video. The Viola–Jones output is a set of bounding boxes that may contain faces. At the start (and again on tracking failure) the Viola–Jones output requires human intervention to select and label a bounding box containing a face. After selection, the corner detection module is triggered, and the tracking loop is engaged. We expect in future work to use a neural net approach such as [25] and [26] to face detection, although unlike the object detection which deploys the same objects during each test session, the face detection algorithm may have different faces in the room in different test sessions.

The Shi–Tomasi corner detector extracts features and scores them [27] using eigenvalues of a characteristic matrix based on image derivatives. The optical flow of each extracted eigenfeature is calculated to track it [28]. The average position of all the trackers is checked against Viola–Jones face boxes for the next frame. If at least 30% of the tracked points are not lost, and if the average tracker position is inside a face box, that is considered a tracking success and the face box is output; the algorithm then continues the tracking loop to the next frame (or moves to a different face if there is another face being tracked). Trackers will continue to function even without a valid detected area and will select the first detected area when available. If 70% of the feature points are lost, or if the average tracker position is not inside any detected face boxes, it is considered a tracking failure. The system then requires human intervention to relocate the face locations, and the algorithm automatically reinitializes the trackers. In a 3-min video consisting of approximately 10 000 frames, there are typically 20–30 reinitializations required (person has to click on the correct face box). The reinitialization typically happens because the subject turns her head and the face exits the field of view, needing reinitialization when it comes back into view, or because the face in the view gets temporarily occluded (e.g., by a hand or object).

C. Definition and Determination of "Look"

Human gaze behavior is typically composed of steady intervals of fixation interposed with fast reorienting movements called saccades. When examined coarsely (granularity of about 2 degrees of visual angle for the viewer), the periods of steady fixation can last between about 200 ms and several seconds depending on the task and level of detail of the object being fixated. At a finer scale, gaze behavior shows a similar pattern of steady fixation and interposed micro-saccades, typically defined as fast orienting movements of less than 1 degree of visual angle. Upon even closer inspection the periods of apparently steady fixation are composed of ocular drift and ocular tremor. Essentially, the eye is constantly moving, as a completely stabilized retinal image fades quickly. The visual system is built to respond to contrast, not stasis. For these reasons, the usual terms associated with the physiology of gaze behavior are not terribly helpful for describing the more cognitive concept of an extended inspection of an object or region. Such an inspection typically involves an aggregated series of fixations and small reorienting saccades or microsaccades. We call this a look; it is defined not in physiological terms, but instead with reference to the object or region under examination. For example, we might usefully describe a look to a face, but might employ a finer scale look to the right eye, when that level of analysis is appropriate.

The object and face detection modules produce bounding boxes around objects and faces in the video frames. For any single object (or face), if the algorithm produces a bounding box in frame i for that object, and if the gaze position is within that bounding box, that frame is considered a hit for that object. Otherwise, it is a miss. The hit sequences for each object/face are run length filtered to determine looks. A run of at least T_1 hits is needed to declare a look, and the first hit position is the start of the look. With a run of T_2 misses in a row, the look is considered terminated, and the last hit position is the end frame of the look. The choice of T_1 and T_2 is discussed in Sections III-C and IV.

III. Ground Truth

GT represents a determination of the true presence of faces and objects and the number and length of looks. GT serves as the basis for evaluating the algorithm results.

A. Ground Truth for Bounding Boxes

GT for face and object bounding boxes was established by manually placing tight axis-aligned enclosing rectangles around each face and object in the image. The protocol for drawing a face box was that the right and left limits should

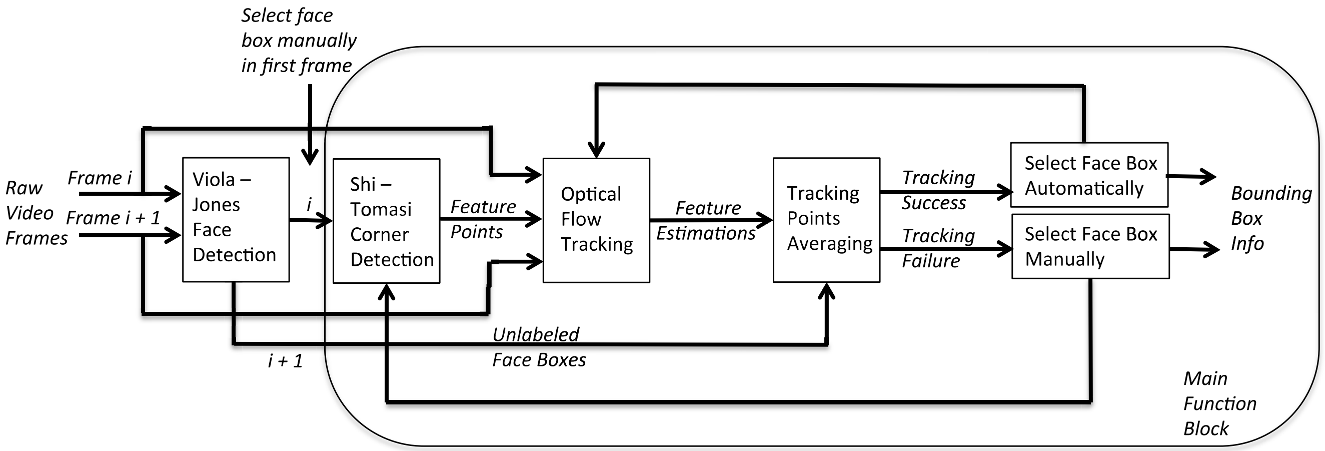


Fig. 4. Block diagram of the face detection module.

include the ears if visible, while the upper limit is at the person's hairline and the lower limit is at the bottom of the chin. An example of manual GT bounding boxes is in Fig. 3(b). A face is not boxed if the face is turned more than 90° away from the camera. A small number of faces were not boxed in the manual GT because the subject wearing the eye-tracking glasses turned his or her head rapidly, so the world-view frames had excessive motion blur [e.g., Fig. 5(a)]. It is possible, however, for the algorithm to detect a face even though it is turned more than 90° or is blurry; such cases would count as false positives since they are not marked in the GT. So the results are slightly conservative on false positives.

For drawing bounding boxes for the top and photograph, the box contains all of the object in the picture, and is drawn only if 50% or more of the object is judged to be present. For the shark object, a box was drawn if 50% or more is present and both eyes are present.

B. Ground Truth for Looks

In one GT approach, an eye-tracking expert determined the GT for looks based on her experience with clinical gaze data, by directly viewing the world-view video with the gaze position dot superimposed on the scene [Fig. 5(b)]. The dot consists of a central red dot (indicating the best position estimate from the eye-tracking glasses) surrounded by a larger green dot (indicating the glasses' estimate of gaze position uncertainty). The expert does not use explicit bounding boxes, but determines holistically, as in clinical or experimental practice, what the subject is looking at. This approach is inherently inferential, and therefore subject to a number of biases. For example, the user may consider that a set of frames corresponds to a single look to a face, despite a short temporal gap in the presence of the gaze dot on the face that may be due to the subject blinking, the subject shifting their head position and producing motion blur, or a reduction in calibration accuracy because of the glasses being jiggled on the subject's head. Indeed, it may happen in practice that the expert notices a calibration error because the gaze dot is consistently slightly low, and so marks a section as a look to an object because they know the subject "intended to look at the object" even though the gaze dot is off. Our videos were calibrated, so this level of subjectivity was not present in the expert GT (called GT-E), but some level of subjective expert judgment is inherent in this process. It is useful to include this type of GT since it is what is actually used currently in analyzing social gaze behavior.

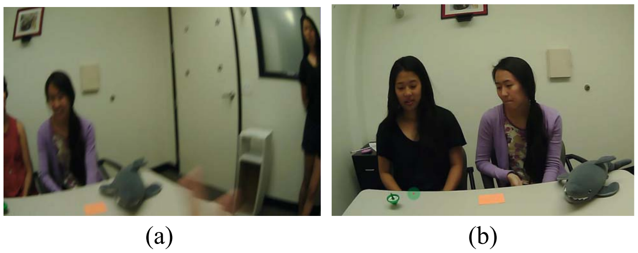


Fig. 5. (a) Example of a motion-blurred image that is not given bounding boxes. (b) Test image with the gaze dot superimposed on the world view.

The other type of GT for looks, referred to as GT-B, uses the manual bounding boxes. As shown in Fig. 1, GT-B is established by putting the gaze position and manually derived bounding boxes for objects/faces through the same run length filtering used on the automatically derived bounding boxes.

C. Entry and Exit Parameters

One approach to choosing entry and exit parameters T_1 and T_2 is based on physiology and eye behavior. At 60 frames/s, five frames represents 83 ms, a reasonable lower bound duration of a single fixation (period of gaze stability between the fast orienting saccadic eye movements). Typically, a fixation duration is about 200–300 ms in standard experimental studies with controlled target appearance and standard screen refresh rates [29]. However, we are measuring gaze behavior in the real world. Human observers typically plan sequences of saccades, especially when scanning a complex object [30] and for those sequences, the fixation duration can be quite short. Depending on task demands and the subject's level of focus, fixations can also be quite long, approaching 2 s. Physiologically speaking, the fixation need only be long enough for the visual system to extract relevant information in high resolution detail before moving to a new spot to examine. Data from visual psychophysics demonstrates that image detail can be resolved with a presentation of only 50 ms, followed by an immediate mask to prevent the use of after images [5]. Given this approximate lower bound, the T_1 value could be even lower than five frames, however in practice, we do not typically see fixations this brief. From the eye physiology point of view, the T_2 exit parameter needs to be long enough to bridge a blink.

The selection of T_1 and T_2 based solely on eye physiology does not connect to our definition of looks as extended inspections, and also ignores the fact that the detection problem is difficult due to occlusions, object deformability, rapid turning of the subject's head, and other reasons. A second approach to choosing these parameters is based on making the algorithm mimic the behavior of the expert neuroscientist, which is the approach we present in Section IV.

D. Camera Calibration and Accuracy Issues

Calibration ensures that the gaze position in the world-view scene corresponds to what the subject is looking at. The calibration markers were bullseye-style markers about 3 inches in diameter. The subject wears the glasses and looks steadily at the target. The markers are detected by the Pupil Capture calibration code (using the Manual Marker Calibration method in Pupil Capture) and were moved sequentially through at least nine points in the calibration plane (approximately at the seated position of the experimenters, 1 m from the subject). The nine points were placed at three vertical levels and three horizontal positions to approximate a grid. This method was used so that we can calibrate a large space in which agents and objects that are part of the social conversation could be reliably detected. Once the calibration routine is completed, we validate it by asking the subject to look at different parts in the scene and confirm that the gaze point represented in Pupil Capture is where the subject reports looking.

Accuracy Issues: The world view camera mounted on the glasses captures the world in the direction the head is facing. Typically, the eyes look forward, and so the gaze position is rarely at the extreme edges of the world view scene. Gaze location data show that the eyes spend less than 5% of the time looking at the area that is within 20% of the edge of the field of view. Furthermore, when the eyes do shift to the side, the glasses have greater gaze position uncertainty. The confidence score for gaze location reported by the Pupil Pro is 98.6% for gazes to the central 10% of the world-view scene, and this sinks to 89.8% confidence for gazes to the outer 10% portion of the scene. For these reasons, bounding boxes that touch the scene border (meaning the object is cut off by the border) are ignored in the performance evaluation. That is, if the GT bounding box coincides with one generated by the algorithm, the performance evaluation does not count this as a true positive. If the algorithm does not output a bounding box for an object at the border, it is not penalized as a false negative.

IV. Results

For a given face or object (e.g., the shark) we first evaluate bounding boxes. We compute for each frame the area of intersection divided by the area of union (IoU) of the algorithmic and manual bounding boxes for that object. The IoU values are averaged over frames, and over five videos, and reported in Table I (last column).

TABLE I
Average Results Across Five Videos Comparing the Algorithm Against GT-B, Where Both Use $T_1 = 5$ and $T_2 = 17$. Den = Number of Frames in the Denominator of (1) Entering Into the Accuracy Computation for Each Face and Object

	Acc.	FPR	FNR	Den	IoU
face1	85.24	9.58	6.3	2825	79.55
face2	77.06	8.22	17.23	2363	64.72
face3	83.23	6.56	11.6	3506	80.77
photo	68.88	21.74	3.6	4753	77.16
shark	81.43	8.18	10.21	3576	78.27
top	71.91	18.53	14.04	1431	69.32
total	77.83	12.0	9.92	18454	76.04

Evaluating the algorithm above the level of bounding boxes, one must choose run length parameters for the filtering. To do this, we examine the accuracy between the algorithm results and GT-E as a function of run length entry and exit parameters T_1 and T_2.

Frame i represents a true positive event for a look to face 1 if frame i is part of a look to that face according to GT and frame i is also part of a look to that face in the algorithm output. Recall that for frame i to be part of a look to a face does not require that the gaze is within the face bounding box for frame i, or even that the face was detected in that frame. If the face was detected and the gaze was inside its bounding box for earlier and later frames, and frame i is part of a sufficiently short gap, then frame i can still be considered part of the look.

A standard definition of accuracy is $A = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN})$ where TP is the number of true positive events, FP is the number of false positive events, FN is the number of false negative events (when a frame is part of a look according to GT but the algorithm does not mark it) and TN represents the number of true negative events (where neither GT nor the algorithm considers a look to be occurring in a given frame). False positive rate and false negative rate are defined as $\text{FPR} = \text{FP}/(\text{FP}+\text{TP})$ and $\text{FNR} = \text{FN}/(\text{TP}+\text{FN})$. Since the subject is often not looking at any of the objects or faces, TN is large, and including it in both the numerator and denominator obscures trends. So we use the definition of accuracy

$$A = \text{TP}/(\text{TP}+\text{FP}+\text{FN}). \tag{1}$$

Fig. 6 shows heat maps in which the color shows the accuracy of results relative to GT-E when using run length entry and exit parameters given by the values on the x- and y-axes. The top heat map is for the accuracy between GT-E and GT-B. The highest accuracy achieved is 71.2% when $(T_1, T_2) = (1,17)$. From Fig. 6(a), we see that the accuracy is generally high for small values of the entry parameter and relatively large values of the exit parameter (e.g., 16–18). The second heatmap in Fig. 6 shows the accuracy between GT-E and the algorithm (with automatic object/face detection and run length filtering). Here, the highest accuracy is 65.1% which occurs when $T_1 = 1$ and $T_2 = 16$.

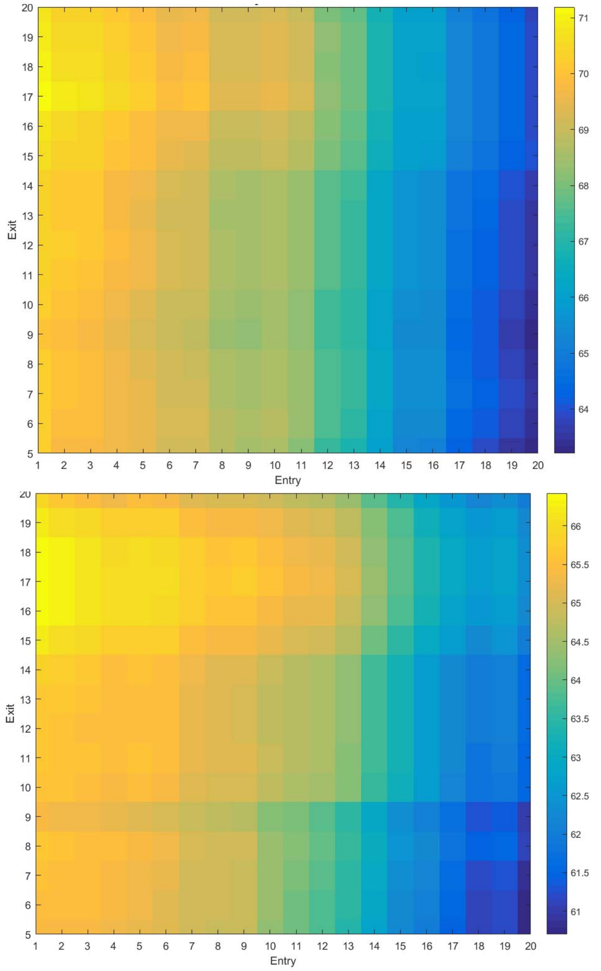


Fig. 6. Heat maps showing accuracy as a function of run length entry (T_1) and exit (T_2) parameters between GT-E and GT-B (top) and GT-E and algorithm (bottom).

To understand the heatmap result, consider first the exit parameter T_2. Suppose the expert neuroscientist observes that the subject gazes at a face, and the face turns away momentarily, and turns back. The expert judges the gaze remains on the face the entire time, a look that spans 100 frames. The algorithm gets 30 frames but loses track when the face turns. The face gets reacquired by the Viola–Jones detection module after a gap of 16 frames. With small values of $T_2 < 16$, this would be considered by the algorithm as two distinct looks with a gap in between. A large value of $T_2 = 16$ bridges the gap; the entire set of frames, including the gap frames, constitute one long look, making for good agreement (many true positives) with GT-E. In short, choosing T_2 somewhat larger than the physiologic causes (e.g., blinks) alone would suggest allows the run length filtering to compensate both for blinks, motion blur, and other deficiencies in the detection modules.

For the entry parameter T_1, the value which maximizes the accuracy is 1, meaning a single hit frame should be considered the start of a look. This is smaller than fixation data would suggest. When the algorithm finds a single frame that has a gaze dot in the bounding box, if there are no further hit frames within a distance of T_2, then this is very unlikely to correspond to a look in GT-E. Yet the accuracy penalty from calling this a look is small, since it is a single false positive frame. On the other hand, if there are other hit frames within a distance of T_2 then the algorithm with $T_1 = 1$ will declare the start of the look and will bridge the gap to the later frames, so all those frames get declared to be part of the look. If this look was marked by the expert in GT-E, then many frames become true positives. In other words, taking small values of $T_1 = 1$ will cause more frames overall to be declared part of looks, increasing both false positives and true positives. In general, the false positive frames will incur little accuracy penalty because they will occur as individual frames, but the true positive frames will usually be part of a larger look event, thereby producing an overall increase in accuracy.

TABLE II
AVERAGE RESULTS ACROSS FIVE VIDEOS, FOR THE COMPARISON OF THE ALGORITHM AND GT-E, WHERE THE ALGORITHM USES PARAMETERS $T_1 = 5$ AND $T_2 = 17$

	Acc.	FPR	FNR	Den
face1	70.34	3.27	27.94	3662
face2	61.01	11.90	33.51	2865
face3	65.42	8.36	30.43	4375
photo	52.75	24.75	16.41	5221
shark	79.21	13.20	6.96	3444
top	68.32	17.34	20.24	1528
total	65.00	13.34	23.23	21095

Based on the second heatmap, we choose to use $T_1 = 5$ and $T_2 = 17$, as these values give nearly the same accuracy (65.0%) as the best parameter set, and using $T_1 = 5$ rather than 1 is close to what would be selected based on physiologic considerations for a fixation. In general, the parameters could be selected based on the application and the need to prioritize either FPR or FNR. It is significant that both the algorithm and the student ground truth GT-B have similar filtering parameters for maximizing accuracy relative to GT-E, and that the accuracy of 71.2% for GT-B is not vastly higher than the best accuracy of 65.1% for the algorithm. This means that even the manual bounding boxes of GT-B often fail to capture the decisions arising in the holistic expert ground truth GT-E, and further improvements will be needed in processing above the level of individual frames or small groups of frames.

Using $(T_1, T_2) = (5{,}17)$ for the algorithm and for GT-B, the values for algorithm accuracy, FPR, and FNR for each object/face, averaged across videos, are in Table I (relative to GT-B) and in Table II (relative to GT-E). Sample results for faces in one video appear in Fig. 7, and for objects are in Fig. 8.

A. Discussion

The faces are labeled 1–3 from left to right, and the middle face (face 2) is usually farther back in the scene. We see that the average IoU values for faces 1 and 3 are very similar (79.55 and 80.77) but the average IoU is worse for face 2. The accuracy results for looks to faces 1 and 3 are also better than those for face 2. The face 2 participant is the one who leaves the room and returns later, and the glasses-wearer is instructed to look at the person returning to the room. The accuracy is

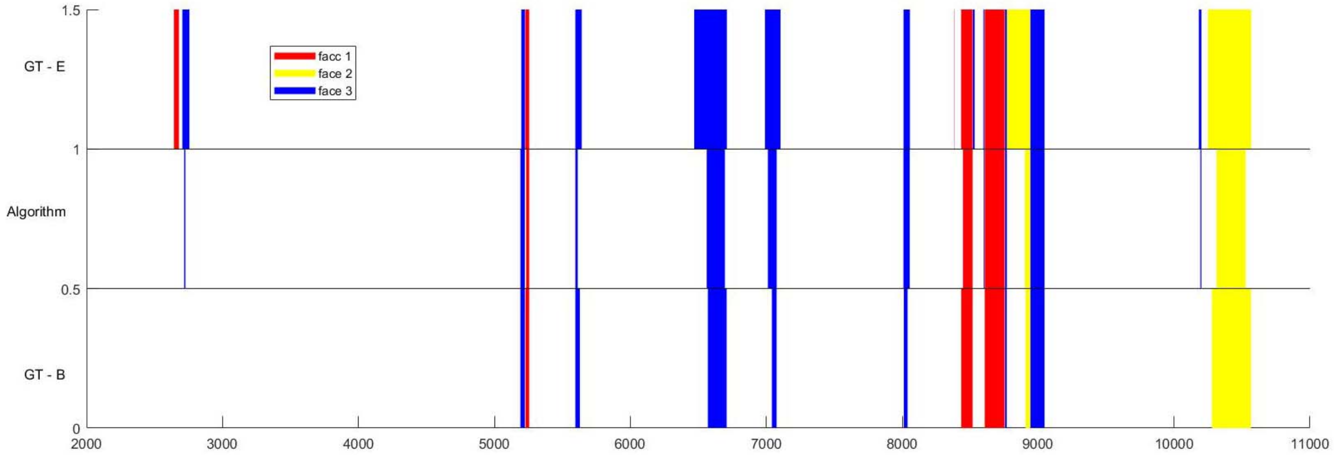


Fig. 7. Example of algorithm results and both GTs for three faces in one video. The *x*-axis shows the frame number. The *y*-axis shows, from top to bottom, GT-E, algorithm looks, and GT-B for faces.

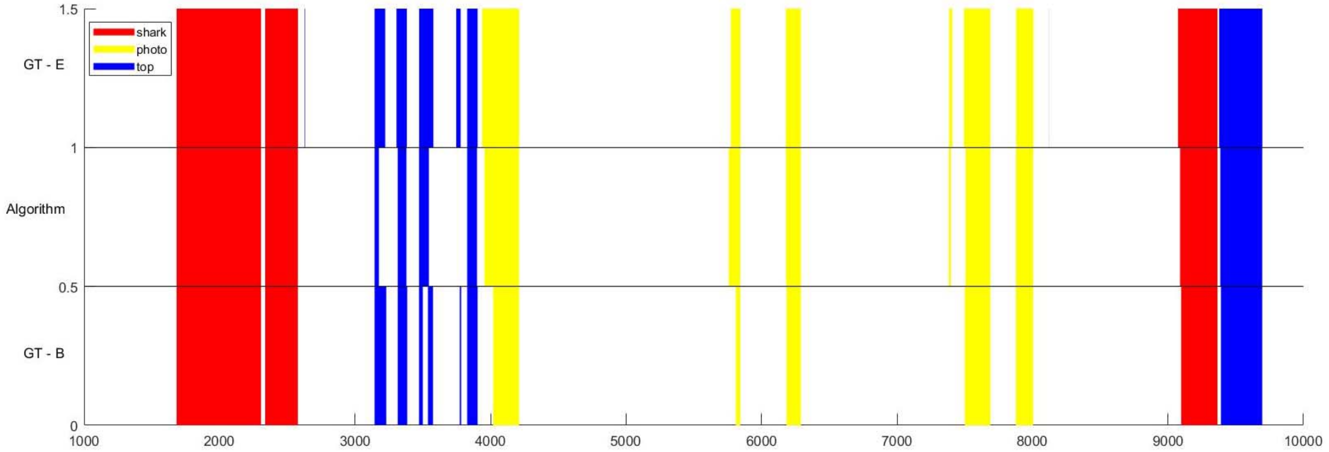


Fig. 8. Example of algorithm results and both GTs for objects in one video. The *x*-axis shows the frame number. The *y*-axis shows, from top to bottom, GT-E, Algorithm results, and GT-B for objects.

lower for face 2 both because of the movement of the participant returning to the room and the changing distance from the glasses-wearer, which means the gaze position calibration is not as accurate for this person.

Among the three objects, the average IoU values are similar for the shark and the photograph (78.27 and 77.16) and the value is lower for the top (69.32) likely because the top is a much smaller object. Of the objects, the photograph has a high FPR. This is driven by the fact that the photograph framing colors (red, black, and white) are common clothing colors worn by the participants, and the photograph itself shows faces, causing nonphotograph items to be detected as photographs, or causing the algorithm's photograph bounding box to be drawn too large. With the exception of the photograph, the accuracy rates for looks are all higher than the IoU measures of bounding box accuracy, suggesting that lack of precision in the bounding boxes can to some degree be compensated for by the run length filtering.

Comparing results between Tables I and II, we see that FPR values are generally similar, whereas FNR values are generally 2 to 4 times higher in Table II than in Table I. Table I compares the algorithm (automatic frame-level detection) against GT-B (manual frame-level bounding boxes), where they both use the same run length filtering. By contrast, Table II compares the algorithm against the holistic GT-E. The fact that there are higher FNR values in Table II means that the expert, in declaring looks, is able to ignore many gaps which come from occlusions, motion blur, blinking and the like, where both the algorithm and GT-B miss frames. The discrepancy is largest for the photograph, possibly because it hangs on the back wall so is the farthest away of all the faces/objects in the scene.

Accuracy results differed substantially among the five participants, with overall accuracy ranging from 46% to 81%. The differences are driven by large differences in the amount of time individual participants chose to look at particular faces and objects. For example, one subject spends about the same amount of time looking at face 1 as at face 2 (notwithstanding the fact that face 1 is in the room the entire time, whereas the face 2 participant leaves and comes back), whereas another subject spends more than four times as much time looking at face 1 as at face 2. Face 2 has lower detection accuracy overall because it is usually farther back in the scene, so the amount of time spent looking at face 2 matters to the overall accuracy of a participant. Similar large differences were found in the time spent looking at specific objects. This points to the need to restrict the set of test objects involved in the interaction to ones that have similar high detection rates, so that the amount of time the subject chooses to look at one compared to another will not have a dramatic effect on the overall accuracy.

A different approach to computing algorithm accuracy could use whole look events, rather than frames within looks, as the basis for correctness. Consider the case where GT-E reports a single long look of 50 frames in the first 100 frames. Suppose the algorithm detects that same look exactly, and also three isolated single frames as being looks. In a frame-based approach to counting correctness, the false positive rate is $3/(50 + 3) = 5.7\%$, whereas in a look-based approach to counting correctness, the false positive rate would be $3/(1 + 3) = 75\%$. In examining the results in Fig. 7, we see that the photograph has five entire "look events" in the GT but six in the algorithm, leading to an FPR of 0.17 if one counts entire look events. However the FPR is different if one counts at the frame level, since several of the look events have extra FP frames at the leading edge of the event. Whether or not it is desirable to count FP and FN events at the level of entire look events or at the level of frames, or indeed whether some completely different metrics are needed, will depend on the application. The best choice of parameters T_1 and T_2 will depend on whether one is optimizing a frame-based accuracy or is using a whole-look-based approach.

B. Comparison With Order Statistic Filters

For comparison with run length filtering, we implemented order statistic filters of various lengths. When filtering the jth element in a sequence with an order statistic filter of (odd) length n, the jth element along with the $(n-1)/2$ preceding elements and the $(n-1)/2$ subsequent elements are arranged in ascending order of magnitude: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, and the filter output $X_{(i)}$ is called the ith-order statistic. Order statistic filters include the min ($i = 1$), max ($i = n$), and median ($i = (n+1)/2$) filters. For binary data, order statistic filters essentially set a threshold for the number of hits needed in the filter window in order to declare the output a hit. Fig. 9 shows a heatmap of the accuracy from (1) between GT-E and the algorithm (where algorithm detections are filtered with the order statistic filter). The x-axis represents filter length (odd values) and the y-axis shows the order. The highest accuracy is 60.5%, achieved with filter length 13 and order 8 [note that (length, order) = (13, 7) is a median filter]. This best filter is slightly biased toward converting a 0 (miss) into a 1 (hit). In general, the heatmap shows a band of bright yellow positions (highest accuracy values) which tend to be the median filters or filters with orders slightly higher than the median filters. The highest accuracy is 60.5%, which is lower than the best of the run length filters. The better performance of run length filtering is likely due to detection misses tending to occur in runs (from blinking, head movement, etc.) and hits also tending to occur in runs (due to periods of fixation with little movement).

C. Consideration of Different Applications

There are research, educational, and clinical applications for which it would be useful to have a system that can automatically identify looks to faces and objects as part of a real-world interaction. These applications vary in their spatial and temporal demands in terms of what constitutes a look, which has a bearing on the values of T_1 and T_2 and on other aspects of the system.

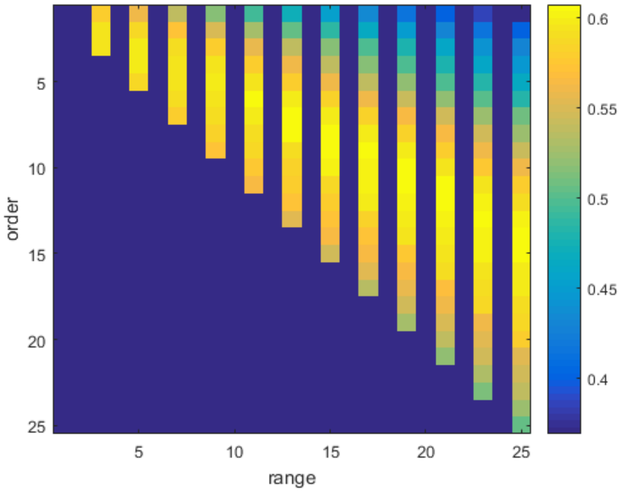


Fig. 9. Heat map showing accuracy, as a function of filter parameters, between GT-E and the algorithm, where detection results are filtered with various order statistic filters. Filter length is on the x-axis (odd values only) and order is on the y-axis.

Consider a child reading a middle school science textbook. One might want to identify when the student is reading the columns of text, approximately at what point in that text the student jumps to a figure box, how long the student spends in the figure box and where the student's gaze goes after the box (ideally back to the point in the text where she left off). Since the primary interest is in mapping gaze onto the textbook, we would want to optimize the spatial accuracy of looks within the book (and not worry about the background). We would not be as concerned about temporal precision in this case. It is useful to know that the child spent about 3.2 s reviewing the figure, but it is not necessary to know that she entered it on frame 80 and left on frame 272. In a clinical example, an adolescent with ASD might wear the gaze glasses and engage in a conversation and a game with two other people. We could identify all looks to faces and quickly calculate the proportion of time spent looking at faces during the interaction as a whole, a potentially useful measure in a social evaluation. For cases such as these where overall time spent looking at a face or object is important but not the number or precise onset of looks, the T_1 and T_2 parameters can be set to the values which optimize the accuracy with GT-E for that task.

It may also be useful to know the *number* of separate looks. If the child looks back and forth ten times between the text and the figure box, it may be a sign that the figure is confusing or has insufficient labeling. When counting the number of looks is the primary goal, the choice of T_1 and T_2 using GT-E would use the count as the optimization goal.

Taking the clinical example further, the adolescent and two research assistants might all wear gaze glasses and the data streams are synchronized. One might like to know how quickly after one assistant turns to the other does the adolescent also turn to look at the assistant. The latency to orient to a social cue is a useful part of a social evaluation since slow orienting behavior can result in missed information. However, whenever we intend to calculate latency, the temporal precision in the onset and offset of a look matters a great deal.

These various applications with various requirements suggest that the algorithm parameters can usefully be tailored for different scenarios. For cases where the spatial precision is important, a restricted region of interest (e.g., the textbook) can be precisely calibrated. Also, allowing some padding region outside the algorithm bounding box for where the gaze location counts as a hit, or conversely, tightening up the region which counts as a hit might allow for greater accuracy optimization between the algorithm and GT-E.

V. Conclusion

This project brings together multiple different technologies to enhance our understanding of gaze behavior in real-world situations. Currently, the use of real-world eye-tracking is limited because the first-to-market glasses-based eye-trackers were expensive, and the resulting gaze-in-world data was difficult to analyze in any automated or semi-automated way. The open source model offered by Pupil Labs has made glasses-based eye-tracking both affordable and customizable. The system developments described here allow us to automate the count and duration estimate of looks to faces and objects during a social interaction. Because of the prevalence of ASD and its social interaction challenges, together with the subjectivity and difficulty in current methods for assessing the success of therapeutic efforts, investing in objective and quantitative social outcome measures can be useful to measure efficacy of social therapies.

One contribution of this paper is the system integration involving both face and object detection in the context of naturalistic social interactions with varied motion of the subject and other participants. But the main contribution is the approach to determining looks, involving a run length algorithm whose parameters are set by optimizing a suitable definition of agreement between the algorithm looks and an expert ground truth. The definition of agreement can be modified depending on the application. The detection accuracy of our modules is already sufficiently high for many clinical or educational evaluation purposes, and superior detection algorithms could be substituted in a modular way for the current methods, retaining the optimized run length approach to determining looks as a postprocessing method after any detection algorithm.

Our long-term goal is to develop a system using gaze glasses and analytic software to assess change in social and communicative behavior in individuals at a range of ages and levels of function. We plan to improve the accuracy of the system through a series of practical changes: switching to objects that are more easily distinguishable, comparing neural net face detection approaches against the current Viola–Jones-based approach, and using glasses that track both eyes. Extending functionality, we plan to include methods for automated sound and voice detection as well as gesture detection. Steps include identifying instances in time (trigger points) from which one might want to calculate latencies. Audio triggers might include a knock on the door, the onset of speech in general, or when a participant's name is spoken. Visually identifiable trigger points include pointing movements, head turns, and other gestures.

Acknowledgment

The authors would like to thank S. Hacker who helped with data collection. This paper extends the preliminary work of the ISBI 2018 conference paper by adding expert ground truth GT-E, algorithm parameter optimization, new results, comparison with order statistics filters, as well as Section IV-C.

References

[1] P. Mundy, "A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder," *Eur. J. Neurosci.*, vol. 47, no. 6, pp. 497–514, Mar. 2018.

[2] D. A. Baldwin, "Understanding the link between joint attention and language," in *Joint Attention: Its Origins and Role in Development*, C. Moore and P. J. Dunham, Eds. Hillsdale, NJ, USA: Erlbaum, 1995, pp. 131–158.

[3] M. Hirotani, M. Stets, T. Striano, and A. D. Friederic, "Joint attention helps infants learn new words: Event-related potential evidence," *Neuroreport*, vol. 20, no. 6, pp. 600–605, Apr. 2009.

[4] B. R. Ingersoll and K. E. Pickard, "Brief report: High and low level initiations of joint attention, and response to joint attention: Differential relationships with language and imitation," *J. Autism Dev. Disord.*, vol. 45, no. 1, pp. 262–268, Jan. 2015.

[5] J. Townsend, E. Courchesne, and B. Egaas, "Slowed orienting of covert visual-spatial attention in autism: Specific deficits associated with cerebellar and parietal abnormality," *Dev. Psychopathol.*, vol. 8, no. 3, pp. 563–584, 1996.

[6] B. Zablotsky, L. I. Black, M. J. Maenner, L. A. Schieve, and S. J. Blumberg, "Estimated prevalence of autism and other developmental disabilities following questionnaire changes in the 2014 national health interview survey," *Nat. Health Stat. Rep.*, no. 87, pp. 1–21, Nov. 2015.

[7] L. Zwaigenbaum *et al.*, "Behavioral manifestations of autism in the first year of life," *Int. J. Dev. Neurosci.*, vol. 23, nos. 2–3, pp. 143–152, Apr./May 2005.

[8] L. Schreibman and B. Ingersoll, "Behavioral interventions to promote learning in individuals with autism," in *Handbook of Autism and Pervasive Developmental Disorders*, vol. 2, F. Volkmar, A. Klin, R. Paul, and D. Cohen, Eds., New York, NY, USA: Wiley, 2005, pp. 882–896.

[9] E. A. Laugeson, F. Frankel, C. Mogil, and A. R. Dillon, "Parent-assisted social skills training to improve friendships in teens with autism spectrum disorders," *J. Autism Dev. Disord.*, vol. 39, no. 4, pp. 596–606, 2009.

[10] P. J. Crooke, L. Olswang, and M. G. Winner, "Thinking socially: Teaching social knowledge to foster Social behavioral change," *Topics Lang. Disord.*, vol. 36, no. 3, pp. 284–298, Jul./Sep. 2016.

[11] M. Chita-Tegmark, "Social attention in ASD: A review and meta-analysis of eye-tracking studies," *Res. Dev. Disabil.*, vol. 48, pp. 79–93, Jan. 2016.

[12] K. Chawarska and F. Shic, "Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder," *J. Autism Dev. Disord.*, vol. 39, no. 12, pp. 1663–1672, 2009.

[13] M. Hosozawa, K. Tanaka, T. Shimizu, T. Nakano, and S. Kitazawa, "How children with specific language impairment view social situations: An eye tracking study," *Pediatrics*, vol. 129, no. 6, pp. e1453–e1460, 2012.

[14] K. Pierce, D. Conant, R. Hazin, R. Stoner, and J. Desmond, "Preference for geometric patterns early in life as a risk factor for autism," *Archives Gen. Psychiat.*, vol. 68, no. 1, pp. 101–109, 2011.

[15] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives Gen. Psychiat.*, vol. 59, no. 9, pp. 809–816, 2002.

[16] E. Chong *et al.*, "Detecting gaze towards eyes in natural social interactions and its use in child assessment," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 3, p. 43, Sep. 2017.

[17] Z. Ye *et al.*, "Detecting eye contact using wearable eye-tracking glasses," in *Proc. UbiComp*, Pittsburgh, PA, USA, Sep. 2012, pp. 699–704.

[18] S. Magrelli *et al.*, "Social orienting of children with autism to facial expressions and speech: A study with a wearable eye-tracker in naturalistic settings," *Front. Psychol.*, vol. 4, p. 840, Nov. 2013.

[19] B. Noris, J. Nadel, M. Barker, N. Hadjikhani, and A. Billard, “Investigating gaze of children with ASD in naturalistic settings,” *PLoS ONE*, vol. 7, no. 9, Sep. 2012, Art. no. e44144.

[20] B. Noris, K. Benmachiche, J. Meynet, J. P. Thiran, and A. G. Billard, “Analysis of head-mounted wireless camera videos,” in *Computer Recognition Systems 2*. Heidelberg, Germany: Springer, 2007, pp. 663–670.

[21] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[22] S. Gidaris and N. Komodakis, “Object detection via a multi-region & semantic segmentation-aware CNN model,” in *Proc. ICCV*, 2015, pp. 1134–1142.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv: 1409.1556v6[cs.CV]*, Apr. 2015.

[24] P. Viola and M. J. Jones, “Robust real-time face detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[25] H. Jiang and E. Learned-Miller, “Face detection with the faster R-CNN,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, 2017, pp. 650–657.

[26] S. Farfade, M. J. Saberian, and L.-J. Li, “Multi-view face detection using deep convolutional neural networks,” in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, Shanghai, China, 2015, pp. 643–650.

[27] J. Shi and C. Tomasi, “Good features to track,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 1994, pp. 593–600.

[28] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, Vancouver, BC, Canada, Aug. 1981, pp. 674–679.

[29] D. D. Salvucci and J. H. Goldberg, “Identifying fixations and saccades in eye-tracking protocols,” in *Proc. Symp. Eye Track. Res. Appl. (ETRA)*, Palm Beach Gardens, FL, USA, 2000, pp. 71–78.

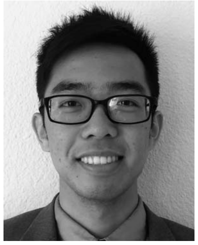
[30] G. T. Buswell, *How People Look at Pictures: A Study of the Psychology of Perception in Art*. Chicago, IL, USA: Univ. Chicago Press, 1935.

Leanne Chukoskie received the Ph.D. degree in neuroscience from New York University, New York City, NY, USA.

She trained as a Post-Doctoral Fellow with the Salk Institute for Biological Studies, La Jolla, CA, USA. She spent three years in the science leadership at Autism Speaks after which she accepted a Research Faculty position with the University of California at San Diego, La Jolla, where she is currently the Associate Director for the Research on Autism and Development Laboratory and the Director of the Power of NeuroGaming Center, Qualcomm Institute. She is also a Professor of social science with Minerva Schools at KGI, San Francisco, CA, USA. She was a Research Scientist with Qualcomm Institute and the Institute for Neural Computation, allowing her to engage in interdisciplinary research with clinicians, engineers, and educators. Her current research interests include sensory-motor behavior, especially eye movement behavior and its neural correlates across both typical and atypical development. This focus has evolved from early studies of basic visual and eye movement processes combined with an interest and experience working with individuals on the autism spectrum. She seeks new tools and new ways to analyze data that might be used to assess outcomes of interventions.

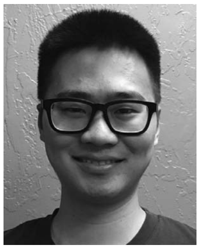
Shengyao Guo received the B.S. degree in electrical engineering from the University of California at San Diego, La Jolla, CA, USA, in 2016, where he is currently pursuing the M.S. degree.

His current research interests include neural network-based machine learning applications in image and video processing, website development, recommendation systems, cloud computing, and big data analytics.



Eric Ho received the B.S. degree in electrical engineering from the University of California at San Diego, La Jolla, CA, USA, in 2017, where he is currently pursuing the M.S. degree in electrical engineering with a focus on machine learning and data science.

His current research interests include computer vision, deep neural networks, and robotics.



Yalun Zheng received the B.S. degree in electrical and computer engineering from the University of California at San Diego, La Jolla, CA, USA, in 2018, where he is currently pursuing the master’s degree in electrical engineering.

His current research interests include machine learning and deep learning.

Qiming Chen, photograph and biography not available at the time of publication.

Vivian Meng, photograph and biography not available at the time of publication.

John Cao received the B.S. degree in electrical engineering from the University of California at San Diego, La Jolla, CA, USA, in 2017, where he is currently pursuing the M.S. degree in electrical engineering with a focus in photonics.

His current research interests include machine learning and photonics.

Nikhita Devgan, photograph and biography not available at the time of publication.

Si Wu is currently pursuing the Computer Engineering degree with the University of California at San Diego, La Jolla, CA, USA.

Her current research interests include software engineering, algorithms, cryptography, machine learning, and psychology.

Pamela C. Cosman (S’88–M’93–SM’00–F’08) received the B.S. degree (Hons.) in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1987 and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1993.

In 1995, she joined the faculty of the Department of Electrical and Computer Engineering, University of California at San Diego (UC San Diego), La Jolla, CA, USA, where she is currently a Professor. She has published over 250 papers in the areas of image/video processing and wireless communications, as well as one children’s book entitled *The Secret Code Menace*, which introduces error correction coding and other concepts in wireless communications through a fictional story. Her past administrative positions include serving as the Director of the Center for Wireless Communications, UC San Diego, from 2006 to 2008 and the Associate Dean for Students of the Jacobs School of Engineering, UC San Diego, from 2013 to 2016.

Dr. Cosman was a recipient of the Globecom 2008 Best Paper Award, the HISB 2012 Best Poster Award, the 2016 UC San Diego Affirmative Action and Diversity Award, and the 2017 Athena Pinnacle Award (Individual in Education). She has been a member of the Technical Program Committee or the Organizing Committee for numerous conferences, including the 1998 Information Theory Workshop in San Diego, ICIP from 2008 to 2011, QOMEX from 2010 to 2012, ICME from 2011 to 2013, VCIP 2010, PacketVideo from 2007 to 2013, WPMC 2006, ICISP 2003, ACIVS from 2002 to 2012, and ICC 2012. She is currently the Technical Program Co-Chair of ICME 2018. She was an Associate Editor of the IEEE COMMUNICATIONS LETTERS from 1998 to 2001 and the IEEE SIGNAL PROCESSING LETTERS from 2001 to 2005. For the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, she served as a Senior Editor from 2003 to 2005 and from 2010 to 2013, and the Editor-in-Chief from 2006 to 2009. She is a member of Tau Beta Pi and Sigma Xi.