2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

USING FACE AND OBJECT DETECTION TO QUANTIFY LOOKS DURING SOCIAL INTERACTIONS

Shengyao Guo, Eric Ho, Yalun Zheng, Qiming Chen, Vivian Meng,
John Cao, Si Wu, Leanne Chukoskie, and Pamela Cosman*

Dept. of Electrical and Comp. Engineering, *Institute for Neural Computation
University of California, San Diego, La Jolla, CA, 92093, USA

ABSTRACT

Quantifying gaze is important in various realms, such as evaluating atypical social looking behavior in autism spectrum disorder. This paper reports on a system that uses eye-tracking glasses and object/face detection to quantify looks. The algorithms use Viola-Jones face detection with feature point tracking and Faster-RCNN object detection trained for three objects, followed by a runlength algorithm to declare the start and end of looks. Results are presented in terms of bounding box overlap and accuracy of looks compared to a manual ground truth. The system can be useful for quantifying gaze behavior during dynamic social interactions.

***Index Terms*—** Eye-tracking, gaze behavior, face detection, object detection

1. INTRODUCTION

Gaze behavior is important during development. Adults name objects in a child's field of view, so joint attention helps in learning words [1, 2, 3]. Gaze behavior in children with autism spectrum disorder (ASD) is atypical in terms of social looking behavior, joint attention to objects, and the timing and accuracy of gaze shifts [4, 5]. Objective assessments of real-world gaze behavior would be useful to determine efficacy of social communication therapies. The success of such therapy is typically measured by questionnaire or observation, both of which are subjective and 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.

Glasses-based eye trackers can facilitate the study of gaze behavior during dynamic social interactions. The glasses fuse a calibrated point of gaze (measured by a camera below the eye), with the world view (from a camera above the eye). Quantification of this videorecorded interaction by manual labeling of events is time-consuming. Here, we report on a system that uses eye-tracking glasses with automated quantification of looks during real-world social interactions. Defining a "look" is challenging. Human gaze behavior is composed of steady intervals of fixation interposed with fast reorienting movements called saccades. At a coarse granularity (e.g., 2 degrees of visual angle for the viewer), the periods of steady fixation can last between approximately 200ms and several seconds depending on the task. A similar pattern of steady fixations and interposed micro-saccades emerges at a finer scale. The usual terms of fixations, saccades, and micro-saccades associated with gaze physiology do not well describe the more cognitive concept of an extended inspection of an object or region, which typically involves an aggregated series of fixations and small re-orienting saccades or micro-saccades. We call each extended inspection a "look".

Some past work [6, 7] has also focused on gaze quantification and social orienting in naturalistic settings, making use of software for automatic tracking of areas of interest [8] but in those works the goal was not development of tools for automating gaze analysis, and all output was reviewed by human coders to ensure high detection accuracy. The closest past work to ours is [9, 10], which aims at developing automatic methods for detecting faces and specifically eye contact events, but in a scenario where the investigator wears the eye-tracking glasses rather than the subject, thereby allowing steady gaze orientation (the child's face does not go in and out of the field of view) and maintenance of distance and avoidance of motion blur.

2. SYSTEM OPERATION

A test session begins with calibration, involving having the subject wear the glasses and look steadily at a bullseye target (in several positions) that is recognized by the Pupil Capture software routine. The Pupil Labs eye-tracking glasses (Pupil Pro) produces video frames (24-bit color, 720×1280, 60Hz) from the world-view camera and raw gaze position data at 120Hz from the eye camera. After calibration, data were recorded. Data were collected to simulate a structured social conversation in a small room. Each approximately 2.5

This work was partially supported by NSF grants IIS-1522125 and SBE-0542013 and by NIH grant R21/R33 MH096967.

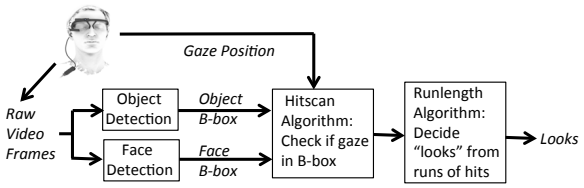

Fig. 1. Algorithm overview. B-box is a bounding box.

minute interaction began with three undergraduate women (two seated and one standing) across from the participant wearing the gaze glasses (also an undergraduate woman, and a different one in each of the 5 videos). After about 15 seconds, the standing person leaves the room (returning during the last 30 seconds). The remaining students play a card game and talk, intermixed with looking at objects and people. The glasses wearer is instructed to avoid large, abrupt movements, to remain seated, and to look at the person leaving and returning, but the interaction proceeds naturally otherwise.

In subsequent (non-real-time) processing, the system detects objects and faces, and determines the presence and duration of looks to these objects and faces. Figure 1 presents an overview of the system operation. The world-view video frames are input separately to object and face detection modules, whose outputs are sets of bounding boxes. The hitscan algorithm takes as input a bounding box and a gaze position (downsampled to 60Hz), and puts out a binary result of whether the gaze position is inside the bounding box; this binary sequence is runlength processed to determine the presence and duration of a look to an object or face.

2.1. Object Detection

The object detection module is based on the Faster R-CNN deep neural network [11] which classifies and localizes multiple objects in a single image. The objects to be detected are a photo, a top, and a toy shark (see Fig. 2). For training the neural network, images were collected using world-view video frames. At distances of 40cm and 80cm, and elevation angles of 0, 30, and 60 degrees above the table, images were taken of the object on a turntable at 10 degree rotations. During testing, the photo hangs on the wall and is not occluded, but the top and shark might have occlusions, so top and shark images with occlusions were included in the training set, as were images of the shark being squeezed. In total, there were 15,000 training images.

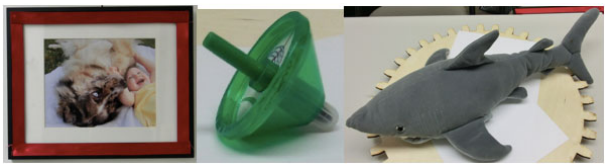

Fig. 2. Objects: photo, top, shark (shown on turntable).

In each training image, a minimum enclosing rectangle was manually placed around the object. The object class and box coordinates serve as the ground truth during training. We exploit transfer learning [12] making use of pre-trained weights from VGG-16 [13]. We used end-to-end training on each object individually. Tuning the model [14] to adapt to our custom dataset consisted of modifying the outputs of the last fully-connected layers. We trained each model with 50,000 iterations with a base learning rate of 0.001 and momentum of 0.9. After this initial training, we fine-tuned the models using additional data consisting of frames from the world-view camera where some of the objects are present. Ground truth bounding boxes were drawn for the objects, and the model performance was gauged by the intersection over union (IoU) of the bounding boxes from the human labelers and those from the Faster R-CNN outputs. If IoU $< 75\%$, we tuned the hyperparameters, such as the learning rate and the mini-batch size, and re-trained that object model.

2.2. Face Detection

For each frame extracted from the input video, the Viola Jones [15] algorithm extracts Haar features and searches each subregion for possible matches. The output is a set of bounding boxes that may contain faces. At the start of the video (and again upon a tracking failure) the Viola-Jones output requires human intervention to select a bounding box containing a face. After box selection, the corner detection module is triggered, and the tracking loop is engaged. The Shi-Tomasi corner detector extracts features and scores them [16]. The optical flow of each extracted eigenfeature is calculated to track it in subsequent frames using the Lucas-Kanade method [17]. The average position of all the trackers is calculated and checked against the face detection boxes output from Viola-Jones for the next frame. If at least 30% of the tracked points are not lost, and if the average position of the tracked points is inside one of the detected face boxes, that is a tracking success and the face bounding 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 of the detected face boxes, that is a tracking failure and human intervention is needed to re-locate the face. The algorithm then automatically re-initializes the trackers. In a 3-minute video consisting of approximately 10,000 frames, there are typically 20-30 face re-initializations required (person has to click on the correct face box). The re-initialization typically happens because the subject turns her head and the face exits the field of view, needing re-initialization when it comes back into view, or because the face in the view gets temporarily occluded (e.g., by a hand or object).

2.3. Determination of Looks

For any single object (or face), if the object detection module produces a bounding box in frame i for that object and the gaze position is in that bounding box, the hitscan algorithm considers this frame a "hit" for that object. Otherwise, it is a "miss". The runlength algorithm processes this binary sequence; a runlength of at least T_1 hits is considered a look, and the first hit position is the start of the look. Due to eye blinks and system noise, a small runlength of misses does not end the look. A runlength of T_2 misses ends the look, and the last hit position is the end frame of the look. The parameters T_1 and T_2 were set equal to 6. At 60 fps, 6 frames represents 100 msec, a reasonable value for declaring a gaze fixation [18, 19, 5] which begins a "look."

2.4. Ground Truth

Ground truth (GT) represents a determination of the true presence of faces and objects and the number and length of looks. 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. Face boxes were drawn to 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. A face is not boxed if the face is turned more than 90 degrees away from the camera. A small number of faces were not boxed in the manual GT because the subject turned his or her head rapidly, so the world-view frames had excessive motion blur. Boxes for the top and photo objects were drawn to contain all of the object in the picture, and were 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 also both eyes are present in the picture. An example of manual GT bounding boxes is in Fig. 3. Bounding boxes which touch the outer edge of the frame (meaning the object is at the edge of the field of view) are ignored both for the algorithm and the GT. This is because the subject's eyes rarely do an extreme sidelong look, and the glasses have lower gaze position accuracy when they do. GT for looks is established by putting the gaze position and GT bounding boxes for objects/faces through the same hitscan and runlength algorithms that are applied to the automatically-derived bounding boxes. An expert neuroscientist viewed a subset of videos with gaze position superimposed as a means of verifying the reasonableness of the GT looks.

3. RESULTS

First, the face and object detection modules are evaluated by comparing their bounding boxes against the manually derived ones. For each frame in each video, we compute the area of intersection divided by the area of union (IoU) of the algorithm and manual bounding boxes for a given face or object. The IoU values are averaged over frames, and over five videos, and reported in the last column of Table 1.

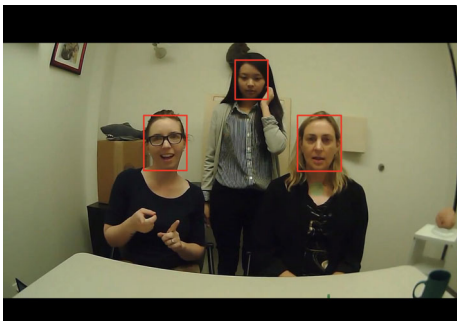

Fig. 3. Image with manual ground truth bounding boxes

	Acc.	FPR	FNR	Den	IoU
face1	85.66	8.5	6.94	2741	79.55
face2	77.07	8.64	16.88	2128	64.72
face3	83.81	7.4	10.18	3317	80.77
photo	66.81	33.08	0.23	4568	77.16
shark	80.36	9.67	12.08	3416	78.27
top	71.16	17.52	16.16	1283	69.32
total	77.22	16.13	9.31	17453	76.04

Table 1. Comparison of the algorithm and ground truth, average results across five videos. Den is the number of frames in the denominator of Equation (1) entering into the Accuracy computation for each face and object.

We next evaluate the algorithm at the level of frames within looks. A sample of the results for one video appears in Figure 4, which is intended to give a qualitative feel for how a subject's eyes move around between the various faces and objects in the scene. 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 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 frame i. From the runlength algoritm, if the face was detected and the gaze was inside its bounding box for at least 6 earlier frames and at least one later frame, and if frame i is part of a sufficiently short gap, then frame i is still considered part of the look.

Let TP = total number of true positive events for a video and object. A false positive event occurs when frame i is not part of a look to that object according to GT but is part of a look to that object for the algorithm. FP and FN denote the number of false positive and false negative events. We define $Accuracy = TP/(TP + FP + FN)$. Table 1 shows the values, averaged across videos, for Accuracy as well as False Positive Rate ($FPR = FP/(FP+TP)$) and False Negative Rate ($FNR = FN/(TP + FN)$).

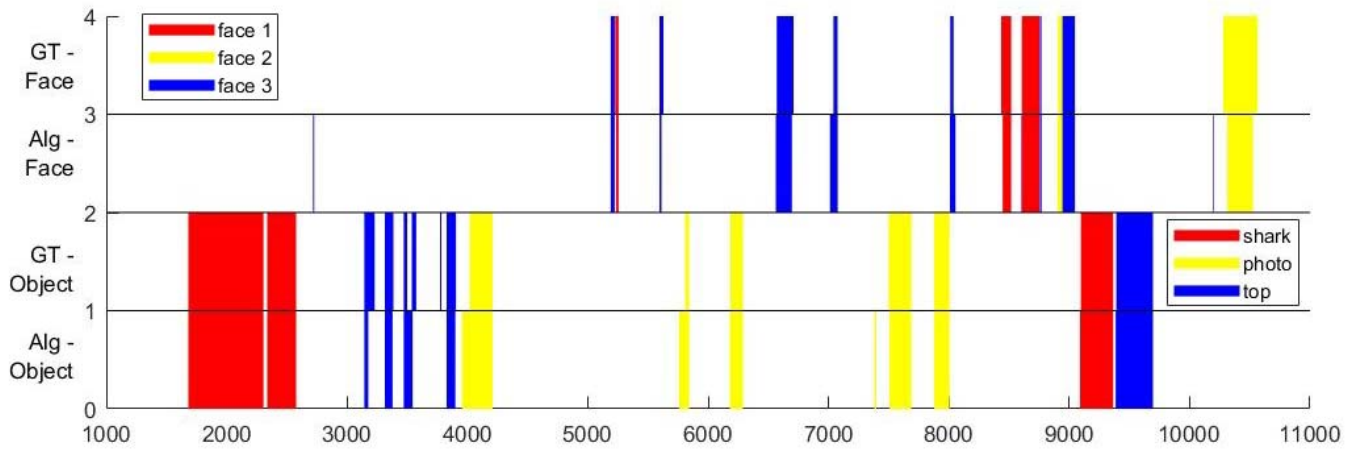

Fig. 4. Example of algorithm results and GT for looks for one video. The x-axis shows the frame number. The y-axis shows, from top to bottom, GT looks to faces, algorithm looks to faces, GT looks to objects, and algorithm looks to objects.

3.1. Discussion

The average IoU values for faces 1 and 3 are very similar (79.55 and 80.77) but they are worse for face2 (the standing person, usually farther back in the scene). Among the objects, IoU values are similar for the shark and photo (78.27 and 77.16) and lower for the top (69.32) likely because it is a much smaller object. Accuracy results for looks to faces 1 and 3 are again better than those for face 2. Of the objects, the photo has a high FPR. This is driven by the fact that the photo framing colors (red, black, white) are common clothing colors worn by the participants, and the photo itself shows faces, causing non-photo items to be detected as photos, or causing the algorithm's photo bounding box to be drawn too large. With the exception of the photo, 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 runlength algorithm that declares looks.

Counting FP and FN events at the level of entire looks, rather than, as we do, at the level of frames within looks, would change the numeric results. For example, in Figure 4, the photo has 5 entire "looks" in the GT but 6 in the algorithm, leading to a FPR of 0.17 if one counts entire looks. However the FPR is different if one counts at the frame level, since several of the looks 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 looks or at the level of frames, or indeed whether some completely different metrics are needed, will depend on the application. There are many different research, educational, and clinical applications for which it would be useful to have a system that can automatically identify looks to faces and objects in real-world interactions, and these applications vary in their spatial and temporal demands in terms of what constitutes a look.

For a child reading a science textbook, one might want to identify when the student is reading the columns of text, where in that text the student jumps to a figure box, how long the student spends in the figure and where the student's gaze goes after it. Spatial accuracy of looks within the book would be important, but temporal precision less so. In another application, one might want to identify all looks to faces and calculate the proportion of time an ASD child spent looking at faces during a whole interaction, a potentially useful measure in a social evaluation. One might want to know the latency between a knock on the door and a look to the door. Temporal precision in the onset of the gaze would be important.

These applications with various requirements suggest that the algorithm parameters, (including T_1 and T_2 or the padding around an object in a bounding box) can be tailored for different scenarios. We believe that the current system is already sufficiently accurate for some of these applications but not others. In particular, the system already can be useful for cases where overall time spent looking at an object is important but the precise onset of looks is not.

Conclusions: The open source model offered by Pupil Labs has made glasses-based eye-tracking affordable and customizable. Our algorithms provide fully automatic detection of looks to objects, and semi-automatic detection of looks to faces, allowing us to study gaze behavior in real-world situations. Evaluation of the usefulness of this type of system will require further development of accuracy metrics that are tailored to particular scenarios, and the algorithms themselves can be tailored to the scenario. Because of the prevalence of ASD and its social interaction challenges, and the difficulty with current methods for assessing therapeutic efforts, objective and quantitative social outcome measures can benefit all social therapies that demonstrate real-world efficacy.

Acknowledgment: We thank Ms. Sarah Hacker who assisted with data collection and troubleshooting.

4. REFERENCES

[1] D.A. Baldwin, "Understanding the link between joint attention and language," In C. Moore & P.J. Dunham (Eds.) *Joint attention: Its origins and role in development,* (pp.131-158). Hillsdale, NJ: Erlbaum, 1995.

[2] M. Hirotani, M. Stets, T. Striano, and A.D, Friederic, "Joint attention helps infants learn new words: event-related potential evidence," Neuroreport. 2009 Apr 22;20(6):600-5. doi: 10.1097/WNR.0b013e32832a0a7c.

[3] 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," *Journal of Autism and Developmental Disorders,* 45(1):262-8, January 2015. doi: 10.1007/s10803-014-2193-8.

[4] P. Mundy, "A Review of Joint Attention and Social-Cognitive Brain Systems in Typical Development and Autism Spectrum Disorder," *European Journal of Neuroscience,* Sep 18. doi: 10.1111/ejn.13720, 2017.

[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," *Development and Psychopathology,* 8(3): 503-584, 1996.

[6] S. Magrelli, P. Jermann, B. Noris, F. Ansermet, F. Hentsch, J. Nadel and A. Billard, "Social orienting of children with autism to facial expressions and speech: a study with a wearable eye-tracker in naturalistic settings," *Frontiers in Psychology,* Vol. 4, Nov. 2013. doi: 10.3389/fpsyg.2013.00840.

[7] B. Noris, J. Nadel, M. Barker, N. Hadjikhani, and A. Billard, "Investigating Gaze of Children with ASD in Naturalistic Settings," *PLOS One,* Vol. 7, Issue 9, Sept. 2012

[8] B. Noris, K. Benmachiche, J. Meynet, J.P. Thiran, and A.G. Billard, "Analysis of head mounted wireless camera videos," *Comp. Recognit. Syst. 2*, 663670, 2007. doi: 10.1007/978-3-540-75175-5_83

[9] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R.M. Jones, A. Rozga, and J.M. Rehg, "Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Vol. 1, No. 3, Article 43, September 2017.

[10] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G.D. Abowd, and J.M. Rehg, "Detecting Eye Contact using Wearable Eye-Tracking Glasses," UbiComp 2012, Sep. 5-8, Pittsburgh, USA.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Neural Information Processing Systems Conf., Montreal, Canada, Dec. 2015.

[12] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both Weights and Connections for Efficient Neural Network," Neural Information Processing Systems Conf., Montreal, Canada, 2015.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv: 1409.1556v6[cs.CV], Apr. 2015.

[14] "Fine-tuning CaffeNet for Style Recognition on Flickr Style Data" Internet: http://caffe.berkeleyvision.org/gathered/examples/finetune_flickr_style.html

[15] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," International Journal of Computer Vision 57(2), pp. 137-154, 2004.

[16] J. Shi and C. Tomasi, "Good Features to Track", IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, June 1994.

[17] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in Proc. 7th international Joint Conference on Artificial Intelligence (IJCAI), Vancouver, B.C., pp. 674–679, August 24-28, 1981.

[18] D. D. Salvucci and J. H. Goldberg, Identifying Fixations and Saccades in Eye-Tracking Protocols, ETRA 00 Proc. 2000 Symp. on Eye Tracking Research & Application, pp. 71-78, 2000.

[19] G.T. Buswell, *How People Look at Pictures: a Study of the Psychology of Perception in Art*, Chicago, IL: University of Chicago Press; 1935.