Analyzing Gaze Behavior Using Object Detection and Unsupervised Clustering

Pranav Venuprasad
pvenupra@ucsd.edu
University of California San Diego
La Jolla, California

Li Xu
l3xu@eng.ucsd.edu
University of California San Diego
La Jolla, California

Enoch Huang
enochihuang@gmail.com
Canyon Crest Academy
San Diego, California

Andrew Gilman
A.Gilman@massey.ac.nz
Massey University
Auckland, NZ

Leanne Chukoskie
lchukoskie@ucsd.edu
University of California San Diego
La Jolla, California

Pamela Cosman
pcosman@ucsd.edu
University of California San Diego
La Jolla, California

ABSTRACT

Gaze behavior is important in early development, and atypical gaze behavior is among the first symptoms of autism. Here we describe a system that quantitatively assesses gaze behavior using eye-tracking glasses. Objects in the subject's field of view are detected using a deep learning model on the video captured by the glasses' world-view camera, and a stationary frame of reference is estimated using the positions of the detected objects. The gaze positions relative to the new frame of reference are subjected to unsupervised clustering to obtain the time sequence of looks. The clustering method increases the accuracy of look detection on test videos compared against a previous algorithm, and is considerably more robust on videos with poor calibration.

CCS CONCEPTS

• **Applied computing** → **Consumer Health**; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

eye-tracking, gaze behavior, unsupervised clustering, object detection, computer vision, Kalman filtering

ACM Reference Format:

Pranav Venuprasad, Li Xu, Enoch Huang, Andrew Gilman, Leanne Chukoskie, and Pamela Cosman. 2020. Analyzing Gaze Behavior Using Object Detection and Unsupervised Clustering. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers)*, June 2–5, 2020, Stuttgart, Germany. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3379155.3391316>

1 INTRODUCTION

Humans use gaze as one of the most basic and frequently engaged ways to learn about and understand the world around them. Gaze

behavior in children with autism spectrum disorders (ASD), a neurodevelopmental disorder affecting approximately 2% of all children [Baio et al. 2018], is atypical in terms of social looking behavior, joint attention to objects, and the basic timing and accuracy of each gaze shift, as well as covert shifts of attention [Bruinsma et al. 2004; Keehn et al. 2013; Miller et al. 2014; Takarae et al. 2004; Townsend et al. 1996]. In fact, disordered visual orienting is among the earliest signs of ASD identified in prospective studies of infant siblings [Zwaigenbaum et al. 2005] and it persists across the lifespan [Miller et al. 2014]. Since humans use gaze as one of the earliest ways to learn about the world around us, this foundational deficit compounds, leading to functional difficulties in learning as well as social domains. However, the building blocks of gaze behavior—accuracy of the metrics and their timing—are malleable. For example, Chukoskie and colleagues [Chukoskie et al. 2018b] have demonstrated that the timing and accuracy of gaze shifts can be trained through playing gaze-contingent video games. The success of these games in changing overt and covert attention has been evaluated on a computer screen [Snider and Chukoskie 2019]. However, it is equally, if not more important to evaluate how training benefits might transfer to real-world behavior. Unfortunately, measuring real-world gaze and related social behaviors accurately and quantitatively is more challenging and demands new tools for objective assessment. Many other different therapies have been designed to address social interaction such as those described in [Crooke et al. 2016; Laugeson et al. 2009; Schreibman and Ingersoll 2013], yet methods of quantitatively assessing the success of these therapies have been limited. These interventions are typically evaluated by parent questionnaire or expert observation, both of which provide valuable information, but both of which are qualitative and time-consuming.

The recent development of affordable eye trackers embedded in glasses has facilitated the examination of real-world human gaze behavior, making it a potentially valuable technique to assess social gaze behavior in individuals with ASD. This paper presents a new approach to detecting looks to objects of interest, based on unsupervised clustering of gaze location data together with neural network based object detection. Here, looks are defined as intervals of time where a subject is focusing their gaze on a single object. The most similar prior work is [Chukoskie et al. 2018a], which presents a system that quantitatively analyzes a subject's gaze behavior using eye-tracking and object detection with runlength filtering of the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7133-9/20/06...\$15.00

<https://doi.org/10.1145/3379155.3391316>

gaze-in-bounding-box results. The current method improves on that work in several ways: (a) We improve the calibration approach and smooth the gaze data stream as done in [Venuprasad et al. 2019], (b) we use a Faster-RCNN face detector [Ren et al. 2015], and (c) most importantly, we develop a novel approach based on unsupervised clustering which considerably improves the detection of looks. In addition to the development of a new approach to look detection in head-mounted gaze tracking, we also introduce a new set of metrics for evaluating the performance. Previous results focused only on frame-level accuracy, which is useful for analyzing, for example, the proportion of time spent by a subject in looking at different objects. Here we examine look-level results as well, which can have value for assessing the timing and naturalness of social gaze behavior. The current study uses objects from the COCO dataset [Lin et al. 2014], demonstrating that the system works well with common real-life objects.

2 PROPOSED METHOD

Here we describe a system that can analyze the pattern of looks of a subject wearing eye-tracking glasses. Our methods for analyzing looks are based on the looking behavior literature in young children where forced-preferential looking paradigms, formerly scored via video but now more frequently with eye gaze data, are prevalent (see for example [Richards and Cronise 2000].) Standard analyses applied to gaze data measured on screens with high spatial- and temporal-resolution remote eye tracking technology typically involve smoothing and differentiating the gaze sample trace to identify saccades and periods of steady fixation [Nyström and Holmqvist 2010]. Fixations, or small collections of fixations, are the "looks" that occur between these fast reorientations. Lower spatial and temporal resolution data does not lend itself to standard saccade detection methods. In addition, these standard saccade identification methods are particularly challenging in mobile head-mounted eye tracking scenarios and demand additional methods and ground-truthed datasets [Kothari et al. 2020; Larsson et al. 2016; Steil et al. 2018]. For this stage of the project, we define a look as an amalgam of fixations (and intervening saccades) that fall within an object's boundaries, obtaining intervals of looks to objects by finding clusters of gaze samples' spatial coordinates in the larger gaze trace.

The intuition is that when a person looks at an object, their gaze position tends to dwell or localize in a region of space. A challenge inherent in head-mounted eye tracking is that the gaze position is relative to the origin of the world-view image obtained from the world camera. As this is a non-stationary frame of reference due to the subject's head movement, we are at risk of not observing the localization phenomenon corresponding to looks. Hence we detect objects in the scene and use their position (centers) to obtain a stationary reference frame. Estimating the position of a single origin using the objects present in the scene is erratic since video frames will generally not have all the objects in the scene. Hence we use separate coordinate systems to detect looks to each object.

2.1 System Overview

We use the Pupil Labs eye-tracking glasses (PupilPro) [Kassner et al. 2014]. World-view frames are obtained from the world camera at 60fps with 1280×720 resolution. The gaze data is obtained from the

pupil camera at 120fps but is then downsampled in time by a factor of two by computing the mean of the gaze points which correspond to each world-view frame. Figure 1 illustrates the detection of looks in the recorded data for each object of interest, and the steps will be described in detail in the next subsections. Objects in the world-view images are detected using a neural network object detector, and the position of the filtered gaze location relative to the center of each object in each video frame is calculated. For each object, clusters close to the origin (object center) are obtained, and the (temporal) connected components are filtered to obtain candidate looks to each object. Conflicting looks are resolved to obtain the final looks to objects. Note that in Fig. 1, the boxes labeled "Object detection" and "Kalman filter, Blink removal" are not new contributions, whereas the remaining elements of the block diagram are new contributions of this paper.

2.2 Object and Face Detection

From the COCO [Lin et al. 2014] dataset, we chose 6 real-life objects: keyboard, stop-sign, cup, bowl, bottle, monitor. We aim to detect the looks of the subject to these objects, as well as to human faces. For estimating tight axis-aligned bounding boxes around these objects in the world video, we use a Faster-RCNN network [Ren et al. 2015] pretrained on the COCO dataset using the tensorflow object detection api [Huang et al. 2016]. The COCO dataset has 80 classes corresponding to common objects, but we ignore the detections of other objects. For face detection, we use a Faster-RCNN model pretrained on the WIDERFACE [Yang et al. 2016] dataset. Even though COCO has a 'person' object class for detecting people in images, we use a separate face detector since the COCO class is not specific to faces and, given the importance of looks to faces in assessing individuals with ASD, we are interested in obtaining tight bounding boxes around human faces. The detections obtained from the two networks are combined. Henceforth, we will refer to a human face as an object in the list of 7 objects of interest. We threshold the classification score of the detected objects, and retain only the detections with a score of at least 50%. Also, if multiple instances of the same object are detected in an image, we retain only the one with the higher score. This is done in order to remove duplicate detections of the same object and other random detections.

2.3 Kalman Filtering and Blink Removal

To account for jitter and noise in the gaze stream, we use the filtering technique described in [Venuprasad et al. 2019]. Since blinking causes incorrect detection of the gaze location which appears as sudden huge shifts in the detected gaze position, we apply a threshold to the movement of the gaze location between consecutive frames to detect blinks. Gaze data corresponding to blinks is marked as unreliable. A Kalman filter with an 8-dimensional state vector and 2-dimensional observation vector is implemented with an initial noise estimate of $\eta = 0.2$ and delta time estimate of $dt = 2.0$. The Kalman filter takes in the observed gaze data as input and computes the filtered gaze position. For frames where the gaze data has been marked as unreliable, only the time update (prediction step) is carried out and the measurement update (correction step)

Figure 1: Block diagram of look detection

is skipped. The filtered gaze stream is given as input to the look detection algorithms.

2.4 Coordinate System Conversion

As described previously, we cannot observe the localization of the gaze dot corresponding to looks to objects if the frame of reference is moving. So we detect objects in the world images and use their center positions to obtain new coordinate systems as follows:

For each world image j :

For each object i :

If object i is present in image j :

$$x_{rel}[j, i] = x_{dot}[j] - x_{center}[j, i]$$

$$y_{rel}[j, i] = y_{dot}[j] - y_{center}[j, i]$$

Else :

$$x_{rel}[j, i] = y_{rel}[j, i] = \infty$$

Here $(x_{center}[j, i], y_{center}[j, i])$ is the detected position of the center of object i in frame j , $(x_{dot}[j], y_{dot}[j])$ is the position of the filtered gaze dot in frame j , and $(x_{rel}[j, i], y_{rel}[j, i])$ is the estimated position of the gaze dot relative to object i in frame j . Here, we assume that, if the subject is looking at a particular object, then that object is present in the frame and hence its center can be detected. If the object is not present in the frame, then the subject is not looking at it and hence we do not need to process that as a possible look (the relative position is set to ∞).

2.5 Unsupervised Clustering

For clustering we use the DBSCAN algorithm [Ester et al. 1996] since it does not require the number of clusters to be pre-specified, and it is inherently density-based, which is suitable for our case. Once the clusters are obtained, we only need to retain the clusters in a small region centered at the origin since we perform a separate

clustering to find the looks to each object. For this, we examine the detections of object i across all frames and find the dimensions of the maximum bounding box around the object ($bbox_{max}[i]$). We then only retain clusters whose centroid falls inside a region given by $0.9 \times bbox_{max}[i]$ centered at the origin.

The parameters for DBSCAN chosen here are $\epsilon = 25$, $Minpts = 10$. $Minpts$ is chosen manually based on domain knowledge, as recommended in [Ester et al. 1996]. The ϵ value is fixed based on the frame-level precision-recall analysis of looks on a different set of recordings, as illustrated in Figure 2. The elbow of the precision-recall curve occurs at $\epsilon = 25$, and also the maximum accuracy is observed at this point. Here, the videos used for the precision-recall analysis are not re-used in the performance analysis.

Figure 2: Precision-recall curve for DBSCAN parameter ϵ

2.6 Connected Components and Filtering Looks

The clustering uses only spatial information on the gaze location and the object locations, and it ignores temporal information. Suppose a person looks at a portion of the bottle at the start of the video, then looks at the keyboard and the bowl and back to the bottle. The clustering algorithm will lump together the earlier frames and the later frames associated with the gaze dot focused on the bottle. So once we obtain the clusters corresponding to the object looks, we can distinguish the multiple distinct looks to the same object by considering the time-connected components of the gaze samples belonging to these clusters. However, these components need some further processing. Sometimes a component is very short in time, for example when the subject's gaze passes through object B on the way from object A to C. Also, sometimes two looks to the same object are separated by a very short saccade or look outside the object. To account for these cases, we ignore looks that are less than 100ms in duration and we combine looks to the same object that are separated by a duration of less than 100ms. At the end of this step, we obtain the look streams to each individual object.

2.7 Resolving Conflicts In Looks

It is possible to have cases where bounding boxes around two objects intersect and the gaze dot falls inside the area of intersection. Such frames can be classified as looks to both of the objects since the algorithm up to this point finds the looks to each object independently. To resolve these cases of looks to multiple objects, we compare the distance from the gaze dot to the centers of objects of conflict and assign the frame as a look to the object whose center is the closest to the gaze dot. Figure 3 shows a case where the gaze dot is inside the bounding boxes for both the mug and the keyboard; the "resolve conflicts" step declares this frame to be a look to the mug since the dot is closer to the mug's box center.

In our study, the stop sign, mug, and bowl as well as human faces are fairly compact objects, so the axis-aligned bounding boxes fit them rather tightly even though the mug and bowl were frequently rotated into different orientations. The bottle and monitor are somewhat more elongated objects, however they always were placed upright and so axis-aligned bounding boxes generally fit tightly for them as well despite rotations around the vertical axis. This was not true for the keyboard, a relatively flat elongated object that, depending on orientation, could cause an axis-aligned bounding box to fit very loosely, as shown in Fig. 3. Because of this, the problem of conflicts in looks was particularly an issue for the keyboard. While the simple conflict resolution we used solved this problem in our case, there are several other ways to solve this problem, including avoiding the use of elongated objects, ensuring that elongated objects are never close to other objects, or using minimum enclosing rectangles.

3 EXPERIMENT DESIGN

The experimental setup is shown in Fig. 4. The subject wears the eye-tracking glasses. The other person is across the table. The objects of interest in the scene (and their approximate distances) are the stop sign (7ft), monitor (5ft), keyboard (2ft), bottle (1.5ft), bowl (1.5ft), and mug (1.5ft), as well as the person's face (3.5ft). Note that the

Figure 3: A typical conflict in looks

Figure 4: Experimental setup

positions of objects were occasionally shifted on the table during each recording.

Five adult volunteers were used as subjects for the experiments. A data session begins with calibrating the glasses; we use the manual marker calibration method described in [Venuprasad et al. 2019]. A 3-inch bullseye marker (provided by Pupil Labs) is placed at different positions (at least 9) in the world view, and the subject is asked to look at it while keeping their head still. In calibration mode, the Pupil Labs software detects the marker and adjusts the gaze dot given that the subject is looking at the marker. We use a confidence threshold of 0.5 for calibration; the system rejects points which it deems to have accuracy below this threshold.

Following calibration, the subject does not need to hold their head still, but can freely look around the room, shifting their gaze from one object of interest to another and also to other objects in the room.

For obtaining the ground truth of looks, an eye-tracking analyst inspects the recorded world-view video with the gaze dot superimposed on the scene and marks the intervals of looks. The detected bounding boxes are not displayed; rather the analyst determines holistically, as in clinical practice, where the subject is looking. For

example, the gaze dot might briefly fall just outside some object region between intervals where the gaze dot is clearly on the object, and the analyst could mark the entire section as a look to the object, considering the error to be caused by calibration inaccuracy, error in blink removal, or other noise based on her experience. We would like our system to account for these errors and hence the comparison is done with such a manual ground truth.

Note that the baseline work [Chukoskie, et al. 2018a] used both a holistic expert ground truth (referred to as GT-E) which involves a person directly labeling looks by reviewing the video with superimposed gaze position in a holistic way that would be used in clinical practice, as well as a ground truth (GT-B) in which the bounding boxes were manually marked and then processed into looks using a runlength algorithm. GT-B allowed one to separate out the effects of bounding box detection accuracy and of runlength processing, and this is not needed in the current work because the neural net models pre-trained on the COCO dataset (for objects) and on the WIDER-FACE dataset (for faces) have high detection accuracy. Hence we only use one type of ground truth, the holistic manual one, as a comparison for the end-to-end algorithm performance.

We recorded a total of 55 minutes of data with 5 adult subjects spread across 11 recordings. The statistics of the ground truth looks in the data are summarized in Table 1.

Table 1: Recorded dataset: Statistics of looks

Object	# of looks	Total duration of looks (min:sec)	Longest look (sec)	Shortest look (sec)	Average duration of looks (sec)
Stop Sign	115	06:34	10.34	0.30	3.42
Bottle	95	05:22	13.75	0.51	3.39
Cup	99	05:46	21.18	0.33	3.49
Bowl	103	05:38	10.65	0.56	3.28
Monitor	86	05:02	11.06	0.55	3.51
Keyboard	106	06:29	08.83	0.43	3.67
Face	96	03:49	07.93	0.31	2.39
Overall	700	38:43	21.18	0.30	3.31

To evaluate the system performance, we compare the system’s detection of looks to a manual ground truth of looks and to the runlength algorithm described in [Chukoskie et al. 2018a]. Here we compare the performances of the two algorithms at the look level and frame level. At the frame level, we define TP to be the number of detected true look frames, FP to be the number of false positive frames, that is the number of frames detected by the algorithm as looks which are not in the ground truth, and FN to be the number of false negative frames, that is frames that are in the ground truth as part of a look but are not detected by the algorithm. The evaluation metrics at the frame level are

- (1) Precision = $\frac{TP}{TP+FP}$
- (2) Recall = $\frac{TP}{TP+FN}$
- (3) Accuracy = $\frac{TP}{TP+FN+FP}$

Here we are using a modified version of the standard definition of accuracy. This is because the number of true negative frames (TN)

is very large for each object, and including the TN term in both the numerator and denominator obscures trends as explained in [Chukoskie et al. 2018a].

While the baseline work only considered frame-level analysis, here we also consider look-level analysis, for which we define the following terms, illustrated in Figure 5. A *first look* is a look detected by the algorithm that has non-zero intersection with a ground truth look, and which also is the earliest in time of all detected looks that have non-zero intersection with that particular ground truth look. An *extra look* is a look detected by the algorithm that has non-zero intersection with a ground truth look, and which is not a first look. Using these definitions, we can define the quantities:

- (1) FL = Number of first looks. This is also the number of ground truth looks that have non-zero intersection with at least one of the looks for the same object detected by the algorithm.
- (2) EL = Number of extra looks. This is the number of detected looks that have non-zero intersection with at least one the ground truth looks for the same object minus FL.
- (3) FPL = Number of false positive looks. This is the number of detected looks that have zero intersection with ground truth looks for the same object.
- (4) FNL = Number of false negative looks. This is the number of ground truth looks that have zero intersection with detected looks for the same object.
- (5) Onset Latency = Difference of starting frame timestamp between detected and ground truth looks for first looks. We examine the mean and standard deviation of onset latency.

Figure 5: Illustration: Look-level analysis

The runlength (RL) algorithm described in [Chukoskie et al. 2018a] detects looks based on two threshold parameters: Entry ($T1$) and Exit ($T2$). If there are at least $T1$ consecutive frames where the gaze dot is inside an object bounding box, the algorithm calls it a start-of-look. Once a start-of-look is detected, the end-of-look occurs when there are at least $T2$ consecutive frames where the gaze dot falls outside the object bounding box. For each of the five subjects, the runlength parameters for that subject’s recordings were determined using the videos of the other four subjects. For example, the videos from subjects 1, 2, 3, and 4 were considered a training set; the processed data yielded a single set of optimal runlength parameters for framewise accuracy (relative to ground truth) for that training set. Those runlength parameters were applied to the data from Subject 5, treated as the test set, to get final results for Subject 5. This process was repeated for each subject’s

data. In this way, the algorithm parameters for a given subject were independent of the data for that subject.

4 RESULTS

4.1 Frame-level Results

For both the clustering and RL methods, the mean frame-level precision, recall and accuracy values of look detection for each object across the videos are given in Table 2. Results for each video are in Table 3. Comparing the clustering approach with the RL approach, the precision is slightly lower for the clustering approach but the recall is significantly higher, and the overall accuracy is higher. Examining the output streams revealed that the RL algorithm misses certain intervals of looks to objects completely in cases where the calibration drifts or where the gaze dot hovers around the object bounding box. For faces, this difference is even more prominent as can be seen from the recall values in Table 2. This is because the subject sometimes looks at a person's hair or neck which is outside the face bounding box. This phenomenon is illustrated in Figure 6. Here a portion of video #5 is analyzed; the look streams from the two methods are plotted along with the ground truth. The video has poor calibration. As the gaze dot hovers near the edge of the bounding box for many frames in the video, the RL algorithm is not able to detect certain looks in their entirety. The clustering algorithm is able to detect these looks since it considers the centroid formed by the gaze locations and considers the maximum bounding box computed across the frames. This trend is even more evident from Table 3. For videos 3,5,6,9 where the calibration drifts, the clustering method performs considerably better than the RL algorithm. For the other cases where the calibration is good, the two methods have similar performance.

Table 2: Frame-level look detection results

Object	Precision (%)		Recall(%)		Accuracy(%)	
	Clust.	RL	Clust.	RL	Clust.	RL
Stop Sign	84.43	95.11	94.58	87.37	81.24	84.43
Bottle	95.60	94.59	96.70	86.46	92.48	82.89
Cup	93.66	93.77	96.17	93.37	90.26	88.43
Bowl	94.08	96.10	96.98	83.25	91.41	80.41
Monitor	92.61	95.20	96.99	95.10	90.14	91.14
Keyboard	84.98	93.44	96.46	89.52	83.49	85.48
Face	92.74	94.10	92.14	78.17	85.97	74.97
Overall	91.04	94.64	95.61	87.53	88.96	83.02

For the objects bottle, cup, bowl, monitor and face, the clustering and runlength methods have similar values of precision. However for the stop sign and keyboard, the runlength algorithm significantly outperforms the clustering. This is because in the experimental setup, there are other objects, which are not of interest to this experiment, placed close to the stop sign and the keyboard. Since the subject is free to look around, there are looks to these nearby objects which are being detected as looks to the stop sign or keyboard because of the maximum bounding box considered. This is illustrated in Figure 7 which shows the look streams for video #8. The subject looks at the stop sign for about 8 seconds, but the

Table 3: Frame-level results for each video

Video #	Length (min:sec)	RL param (Entry, Exit)	Accuracy(%)	
			Clust.	RL
1 (Subject: 1)	5:08	(1,56)	92.15	92.04
2 (Subject: 1)	4:38	(1,56)	88.54	84.26
3 (Subject: 2)	3:16	(2,57)	96.12	73.89
4 (Subject: 2)	5:31	(2,57)	97.54	97.89
5 (Subject: 3)	4:25	(1,57)	71.45	60.41
6 (Subject: 3)	4:12	(1,57)	89.95	70.31
7 (Subject: 3)	5:13	(1,57)	90.74	92.45
8 (Subject: 4)	5:53	(2,57)	83.81	91.45
9 (Subject: 4)	6:19	(2,57)	89.85	81.49
10 (Subject: 5)	5:46	(2,57)	94.46	95.75
11 (Subject: 5)	5:17	(2,57)	85.58	85.41
Overall	5:03	(1.5,56.8)	89.11	84.12

Figure 6: Illustration of look streams: Video #5

Figure 7: Illustration of look streams: Video #8

clustering method detects this as a 16-second look since the subject shifts his gaze to a nearby object, as observed from the world video.

Table 4: Look-level detection results

	Number of GT looks	Total number of looks detected		FL		EL		FPL		FNL	
		Clustering	RL	Clustering	RL	Clustering	RL	Clustering	RL	Clustering	RL
Stop Sign	115	139	190	114	113	2	22	23	55	1	2
Bottle	95	101	152	93	92	1	15	7	45	2	3
Cup	99	120	163	96	98	6	25	18	40	3	1
Bowl	103	116	164	101	102	5	33	10	29	2	1
Monitor	86	98	139	86	86	5	19	7	34	0	0
Keyboard	106	126	185	104	103	12	29	10	53	2	3
Face	96	101	140	93	88	0	8	8	44	3	8
Total	700	801	1,133	687	682	31	151	83	300	13	18

4.2 Look-level Results

The look-level detection results are summarized in Table 4. The numbers of false negatives are low, 1.8% for the clustering approach and 2.6% for the RL approach, indicating that most of the ground truth looks are being detected by both algorithms. For faces, somewhat more looks are missed by the RL algorithm, likely due to the subject looking around the face as described previously. The number of extra looks is much higher for the runlength algorithm, which means it is splitting some ground truth looks into multiple looks to the object. The number of false positives is also much higher for the runlength algorithm. Overall the clustering algorithm is considerably more accurate at detecting looks.

4.3 Latency Results

The latencies of look detection for each object are given in Table 5. The latency is defined as the difference in starting frame between the ground truth look and the algorithm’s corresponding first look (if there is a detected look that overlaps that ground truth look). This type of metric would be appropriate for cases where there is an event that might trigger a gaze shift. For example, a knock on the door might trigger a look to the door, a ring on a phone might trigger a look to the phone, and a mention of some object might trigger a look to that object. In the context of assessing individuals with ASD, a protocol could have the timing of such trigger events embedded in the data stream, and the goal is then to detect the onset of the first look after the trigger event, in order to measure the latency between the trigger event and the look onset. In such a case, the existence of extra looks is not problematic, since the onset would be calculated using the first look after the trigger event. False negative events would be problematic, since the estimate of latency would be far off if the look occurs but is not detected.

In Table 5, columns 2 and 3 show the mean latency for the clustering and RL algorithms. The fact that it is negative for the clustering algorithm means that, on the average, the clustering algorithm is detecting the start of a look earlier than the ground truth. This slight bias is unsurprising, as the algorithm uses a maximum bounding box, so as the gaze moves close to an object it can already be counted as a gaze to the object. While the centroid of a gaze cluster is inside the bounding box for the object, a look can start earlier than when the gaze dot itself is inside the object bounding box. This

Table 5: Latency (Lat.) of look detection

Object	Mean(Lat.) (# frames)		Mean(Lat.) (# frames)		Stdev(Lat.) (# frames)	
	Clust.	RL	Clust.	RL	Clust.	RL
Stop Sign	-6.74	2.98	10.25	7.53	46.32	29.20
Bottle	-0.95	1.53	4.63	6.85	9.58	27.35
Cup	0.33	-1.77	4.75	3.08	9.82	3.83
Bowl	-3.80	-1.98	6.17	3.00	10.35	4.13
Monitor	-4.95	-3.11	7.07	3.39	17.85	4.55
Keyboard	-4.64	1.47	16.5	6.46	32.04	14.85
Face	-6.03	0.26	9.96	6.25	35.52	21.25
Overall	-3.49	0.013	8.63	5.27	27.91	18.74

is not the case for the runlength algorithm, since it cannot start a look until a gaze dot is actually inside the object bounding box.

Columns 4 and 5 show the mean of absolute value of latency, which is 0.14 seconds for the clustering algorithm and 0.087 seconds for the runlength algorithm. The runlength algorithm slightly outperforms the clustering, but both methods have generally low values of latency. Among the objects, the stop sign and keyboard have relatively large mean absolute latencies. As discussed earlier, the stop sign and keyboard have other objects near them which are not in our set of objects of interest, but which can attract the subject’s gaze. Also, the keyboard is an elongated object which often runs along the diagonal of a bounding box that is aligned with the world-view coordinate axes, so there are false positive frames associated with these looks. These observations can help inform the design of protocols for ASD assessment that would lead to higher accuracy in latency measurements (e.g., objects which are being used as objects of interest should be compact rather than elongated, and should not be placed too close to other objects).

4.4 Limitations and Future Work

One limitation of this study is the accuracy of the ground truth; a single eye-tracking analyst manually determined the ground truth. In future work, having multiple people undertake this task would both yield a more accurate ground truth, and also the inter-rater reliability would provide useful context for evaluating the performance of the algorithm (e.g., perhaps the difference between the

algorithm and a human is no larger than the difference between one human and another).

A limitation of the algorithm is complexity. We evaluated the processing time of each step running on a i7-7700K CPU @ 4.20GHz $\times$ 8 processor. The detection and filtering take 0.897sec and 0.023sec respectively per frame. All frames of the video are needed for the clustering step, which took approximately 18min for a 5min video with 60 fps. The object and face detection is the computational bottleneck in the pipeline. While the system is intended for post-processing analysis rather than real-time operation, in future work, processing time could be decreased by switching to a lower complexity neural network for the detection.

Another limitation to be considered is our reliance on computer vision-based object detection to calculate looks and its impact on the generalizability of our approach. Our methods are in development with the intention of making them available broadly, including use by non-experts. At the current time, we are still at a point where the system needs handling by individuals who are well-trained in the methods. As far as the available trained object classes, we will seek to use the most recent methods to improve ease of adding robustly detected objects to the library, but the first version of our tools that would be made widely available will likely include a shopping list for specific objects. Note that this is standard practice for many norm-referenced psychological tests, such as the Autism Diagnostic Observational Schedule, which comes with a large tote full of toy and book-based testing materials [Lord et al. 2012].

Object detection failures will have a greater or lesser impact depending on the nature of the analyses we seek. When we are seeking to characterize overall time spent looking to particular objects, a failure to detect an object completely would create significant impact, while a delay in object detection would contribute minimally. On the other hand, if we are examining the latency of gaze orientations to objects, then we would be impacted more by a delay in missed frames of gaze orientation to an object, but an entirely missed object would simply reduce the number of samples over which to calculate an average gaze orientation latency for each object.

5 CONCLUSION

This project describes the use of unsupervised clustering to robustly analyze human gaze behavior. There are many situations in which this technology might be beneficial. Although our goal is analysis of social interaction during typical and atypical development, this technology can be utilized to examine shopper looking patterns in a store, patterns of gaze during driving behavior, and student behavior engaged in reading and instructional tasks, for example. In social interactions with autistic children wearing eye-tracking glasses, good calibration cannot always be guaranteed due to practical difficulties. In these real-life scenarios, the technique described can be especially helpful to provide a robust analysis of the looking patterns of the subject. As a tool for analyzing gaze behavior, this work can be extended to analyze the latency of human subjects to external stimuli such as an audio trigger like a door knock or a spoken command to look at an object. The experiment can be repeated at various time intervals to analyze the progress of children over time. To improve the look detection further, we can consider

optimizing the parameters for a weighted metric obtained by a combination of the metrics described here and this would require a careful examination of the desired end goal of the system.

ACKNOWLEDGMENTS

We are grateful for support from the National Science Foundation under Grants IIS-1522125, DUE-1928604, and SMA 1640909. We also gratefully acknowledge NVIDIA Corporation's donation of the Titan Xp used for this research.

REFERENCES

- Jon Baio, Lisa Wiggins, Deborah L Christensen, Matthew J Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, et al. 2018. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries* 67, 6 (2018), 1.
- Yvonne Bruinsma, Robert L Koegel, and Lynn Kern Koegel. 2004. Joint attention and children with autism: A review of the literature. *Mental retardation and developmental disabilities research reviews* 10, 3 (2004), 169–175.
- Leanne Chukoskie, Shengyao Guo, Eric Ho, Yalun Zheng, Qiming Chen, Vivian Meng, John Cao, Nikhita Devgan, Si Wu, and Pamela C. Cosman. 2018a. Quantifying Gaze Behavior During Real-World Interactions Using Automated Object, Face, and Fixation Detection. *IEEE Transactions on Cognitive and Developmental Systems* 10 (2018), 1143–1152.
- Leanne Chukoskie, Marissa Westerfield, and Jeanne Townsend. 2018b. A novel approach to training attention and gaze in ASD: A feasibility and efficacy pilot study. *Developmental neurobiology* 78, 5 (2018), 546–554.
- Pamela Crooke, Michelle Winner, and Lesley Olswang. 2016. Thinking Socially: Teaching Social Knowledge to Foster Social Behavioral Change. *Topics in Language Disorders* 36 (09 2016), 284–298. <https://doi.org/10.1097/TLD.0000000000000094>
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-based Algorithm for Discovering Clusters: A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, 226–231. <http://dl.acm.org/citation.cfm?id=3001460.3001507>
- Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2016. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR* abs/1611.10012 (2016). arXiv:1611.10012 <http://arxiv.org/abs/1611.10012>
- Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. <https://doi.org/10.1145/2638728.2641695>
- Brandon Keehn, Ralph-Axel Müller, and Jeanne Townsend. 2013. Atypical attentional networks and the emergence of autism. *Neuroscience & Biobehavioral Reviews* 37, 2 (2013), 164–183.
- Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (2020), 1–18.
- Linnéa Larsson, Andrea Schwaller, Marcus Nyström, and Martin Stridh. 2016. Head movement compensation and multi-modal event detection in eye-tracking data for unconstrained head movements. *Journal of neuroscience methods* 274 (2016), 13–26.
- Elizabeth A. Laugeson, Fred Frankel, Catherine Mogil, and Ashley R. Dillon. 2009. Parent-Assisted Social Skills Training to Improve Friendships in Teens with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders* 39, 4 (01 Apr 2009), 596–606. <https://doi.org/10.1007/s10803-008-0664-5>
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 <http://arxiv.org/abs/1405.0312>
- C Lord, M Rutter, P DiLavore, S Risi, K Gotham, and S Bishop. 2012. Autism diagnostic observation schedule—2nd edition (ADOS-2). *Los Angeles, CA: Western Psychological Corporation* (2012).
- Michael Miller, Leanne Chukoskie, Marla Zinni, Jeanne Townsend, and Doris Trauner. 2014. Dyspraxia, motor function and visual-motor integration in autism. *Behavioural brain research* 269 (2014), 95–102.
- Marcus Nyström and Kenneth Holmqvist. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods* 42, 1 (2010), 188–204.

- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99. <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf>
- John E Richards and Kim Cronise. 2000. Extended visual fixation in the early preschool years: Look duration, heart rate changes, and attentional inertia. *Child Development* 71, 3 (2000), 602–620.
- Laura Schreibman and Brooke Ingersoll. 2013. *Behavioral Interventions to Promote Learning in Individuals with Autism*. John Wiley & Sons, Ltd, Chapter 34, 882–896. <https://doi.org/10.1002/9780470939352.ch8> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470939352.ch8>
- Joseph Snider and Leanne Chukoskie. 2019. Gaze-based video games for assessment of attention outside of the lab. In *IEEE EMB Special Topic Conference on Healthcare Innovations and Point-of-Care Technologies (HI-POCT'19)*. 4.
- Julian Steil, Michael Xuelin Huang, and Andreas Bulling. 2018. Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–9.
- Yukari Takarae, Nancy J Minshew, Beatriz Luna, Christine M Krisky, and John A Sweeney. 2004. Pursuit eye movement deficits in autism. *Brain* 127, 12 (2004), 2584–2594.
- Jeanne Townsend, Naomi Singer Harris, and Eric Courchesne. 1996. Visual attention abnormalities in autism: Delayed orienting to location. *Journal of the International Neuropsychological Society* 2, 6 (1996), 541–550.
- Pranav Venuprasad, Tushal Dobhal, Anurag Paul, Tu N. M. Nguyen, Andrew Gilman, Pamela Cosman, and Leanne Chukoskie. 2019. Characterizing Joint Attention Behavior During Real World Interactions Using Automated Object and Gaze Detection. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19)*. ACM, New York, NY, USA, Article 21, 8 pages. <https://doi.org/10.1145/3314111.3319843>
- Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lonnie Zwaigenbaum, Susan Bryson, Tracey Rogers, Wendy Roberts, Jessica Brian, and Peter Szatmari. 2005. Behavioral manifestations of autism in the first year of life. *International journal of developmental neuroscience* 23, 2-3 (2005), 143–152.