SCIENTIFIC REPORTS
nature research

OPEN

Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities

Rakshit Kothari[1*], Zhizhuo Yang[2], Christopher Kanan[1,3], Reynold Bailey[2,3], Jeff B. Pelz[1,3] & Gabriel J. Diaz[1,3]

The study of gaze behavior has primarily been constrained to controlled environments in which the head is fixed. Consequently, little effort has been invested in the development of algorithms for the categorization of gaze events (e.g. fixations, pursuits, saccade, gaze shifts) while the head is free, and thus contributes to the velocity signals upon which classification algorithms typically operate. Our approach was to collect a novel, naturalistic, and multimodal dataset of eye + head movements when subjects performed everyday tasks while wearing a mobile eye tracker equipped with an inertial measurement unit and a 3D stereo camera. This Gaze-in-the-Wild dataset (GW) includes eye + head rotational velocities (deg/s), infrared eye images and scene imagery (RGB + D). A portion was labelled by coders into gaze motion events with a mutual agreement of 0.74 sample based Cohen's κ. This labelled data was used to train and evaluate two machine learning algorithms, Random Forest and a Recurrent Neural Network model, for gaze event classification. Assessment involved the application of established and novel event based performance metrics. Classifiers achieve ~87% human performance in detecting fixations and saccades but fall short (50%) on detecting pursuit movements. Moreover, pursuit classification is far worse in the absence of head movement information. A subsequent analysis of feature significance in our best performing model revealed that classification can be done using only the magnitudes of eye and head movements, potentially removing the need for calibration between the head and eye tracking systems. The GW dataset, trained classifiers and evaluation metrics will be made publicly available with the intention of facilitating growth in the emerging area of head-free gaze event classification.

Human visual behavior can be viewed as a sequence of periods of stable visual input, punctuated by saccades to new locations within the visual environment. Although saccadic targeting during visual search may demonstrate an influence of visual salience[1], the effect is overwhelmed in the presence of a task, when motor execution requires that attention be directed towards information-rich, task-relevant locations within the visual environment[2–4]. As a result, the dynamics of gaze coordination in natural contexts are affected by a variety of extra-retinal properties of the task, the agent, the environment, and by their interaction. These include the spatial distribution of information in the natural environment[5], cognitive resources related to memory or higher order reasoning[6], motor constraints that determine the dynamics of gaze shifts[7–10], and biomechanical constraints that influence visual strategies for foot placement during locomotion[11].

Despite the importance of extra-retinal influences upon gaze behavior during visually guided action, surprisingly little attention has been dedicated to the study of gaze behavior in more natural contexts. For instance, head movements are often suppressed through the use of a chin-rest, or by constraining target movement to only a small portion of the subject's visual field. Furthermore, target motion is often restricted to two dimensions, and sometimes viewed monocularly. In part, the study of strategies for coordination of the eyes, head, and body has been limited by a lack of suitable technology. Successful tracking of the coordination between the head and eyes in unconstrained settings requires advances in two parallel domains: the instrumentation to jointly monitor the direction of the eyes and head ("eye + head tracking"), and the algorithms to parse and categorize key oculomotor events in the rapid stream of data (i.e. "event detectors").

[1]Chester F. Carlson Center for Imaging Science, RIT, Rochester, NY, USA. [2]Golisano College of Computing and Information Sciences, RIT, Rochester, NY, USA. [3]These authors contributed equally: Christopher Kanan, Reynold Bailey, Jeff B. Pelz and Gabriel J. Diaz. *email: rsk3900@rit.edu

The present study aims to develop new event detectors for the study of eye and head coordination during natural behavior. This involves both the development of a custom eye + head tracker, and the capture of a novel dataset of head-free gaze behavior - the Gaze-In-Wild dataset (GW). GW was collected from 19 participants engaged in everyday activities using spatially and temporally calibrated equipment comprised of a hardhat with an inertial measurement unit (IMU), eye tracking glasses, and a stereo-based RGB-D (RGB imagery plus depth) sensor. A custom-made software tool which facilitates the efficient hand-labelling of the captured data was used to label a significant portion (approx. 2 hours and 15 minutes) of the GW dataset (see Section 3). We then use this labelled data for supervised training and assessment of automated event detectors.

This work builds upon a variety of techniques previously used to track head orientation during natural behavior. Published studies have demonstrated the use of rotational potentiometers and accelerometers[8], magnetic coils[12], or motion capture[13] for the sensing of head orientation[6]. Perhaps the highest precision eye + head tracker which allowed body movement leveraged a 5.8 m^3 custom-made armature capable of generating a pulsing magnetic field. The subject was outfitted with a head-worn receiver capable of measuring head position and orientation within its operational region[14]. Despite the high accuracy, this solution, and all solutions involving optical motion-capture are limited in that they constrain the user to a predefined capture volume. Several systems have adopted video based head motion estimation using egocentric video[10,15] and demonstrated promising results, but are too computationally expensive for real-time use, and are prone to irrecoverable track loss especially during periods of rapid head movement, occlusion of tracking features or degradation of image quality due to motion blur. Recent approaches have involved the use of head-mounted IMUs. For example, Larsson *et al.* used a head-mounted IMU in a study where subjects were asked to perform visual tracking tasks when watching pre-rendered stimuli projected onto a 2D screen[16]. They established that compensating for head movements results in a reduced standard deviation for the eye position signal. More recently, Tomasi *et al.* used two IMUs for tracking eye and head orientation relative to heading direction[17] and reported a 7.1° average angular error ($\sigma = 5.2°$). However, estimates of orientation using IMU will accrue error over time. Although solutions which fuse IMU pose estimates with head tracking based on egocentric video are promising, they have not yet been adopted in the context of eye tracking, and to do so is beyond the scope of the current work.

This work also builds upon a long history of methodologies for the automated detection of gaze events within an eye tracking signal. The simplest methods use threshold based filters and numerous descriptive features for classification[18]. Threshold based techniques require parameter tuning for each test scenario as well as being sensitive to noise and sample rate. A better solution is to use machine learning to learn a model for classifying gaze events. These algorithms have been shown to work well when the head is fixed. Pekkanen *et al.* proposed the Naive Segmented Linear Regression (NSLR) model[19] which segments a time sequence into distinguishable events which are then classified using continuous Hidden Markov Models (HMM). While earlier work used hand-crafted features[20], more recent methods have employed recurrent neural networks (RNN)[21] which enable algorithms to directly learn what features are relevant to the task.

Classification Scheme and Nomenclature

Gaze classification requires distinct and separable classes that are identifiable in our daily activities. There has been some disagreement in the research community about the specific criteria for establishing a taxonomy of gaze events[22,23]. For example, one approach is to classify events based upon specific oculomotor movements, such as the two major retinal image stabilizing mechanisms: the vestibular-ocular response (VOR), and the opto-kinetic response (OKR). In VOR, the semicircular canals of the inner ear measure head rotation acceleration which results in eye movements in the opposite direction with near unity gain (*i.e.*, the ratio of eye and head velocity is ~1, see Fig. 1). OKR is generated by retinal motion which in turn leads to compensatory eye movements to reduce retinal blur[8,24–26]. It is difficult to derive a classification scheme based solely on these stabilizing mechanisms, because they may be used in isolation, or in combination, for either fixation of a target that is stationary in the exocentric frame, or pursuit of a moving target.

Our approach is to adopt an exocentric classification scheme and to discuss its applicability in classifying a broad range of coordinated head and eye movements. We define movement categories by the functional role of the eye movement, as well as the motion of an object within an exocentric frame of reference. As a result, events in our dataset is classified as follows:

1. **Gaze fixation (GF)** - Gaze fixation may be brought about through stabilization of the eyes and head, or during movements of the eyes and head that are compensatory and, as a result, produce a stable gaze vector on a stationary object in the world coordinate frame. Stabilized retinal image motion lies near to the range of 0.5 to 5°/*s*, a limit above which the target image starts to blur[24]. Hence, a wide range of miniature head compensated eye movements can be termed as gaze fixation. In our taxonomy, gaze fixations may be further categorized as:

- **Tremors** - The resting eye and head rarely display perfect stability. Skavenski *et al.* identified that despite instructing subjects to remain as stationary as possible, tremor was observed in the head and eyes ($<1°/s$, $10\,Hz$)[27]. Furthermore, the characteristics of tremor is known to vary based on the nature of the instrumentation[28] and type of restraint[27].
- **Drift** - Drifts are slow motions of the eye that are often punctuated by microsaccades and aid in maintaining crisp visual features across the retina. While there is some disagreement on the range of drift motion, they usually display amplitudes within 0.25° and velocities less than 0.5°/*s* when the head is fixed[28].
- **Microsaccades** - Small, rapid eye movements that occur in between fixations are termed as microsaccades and usually last about 25 *ms* with a velocity range capped at 50°/*s*[28].

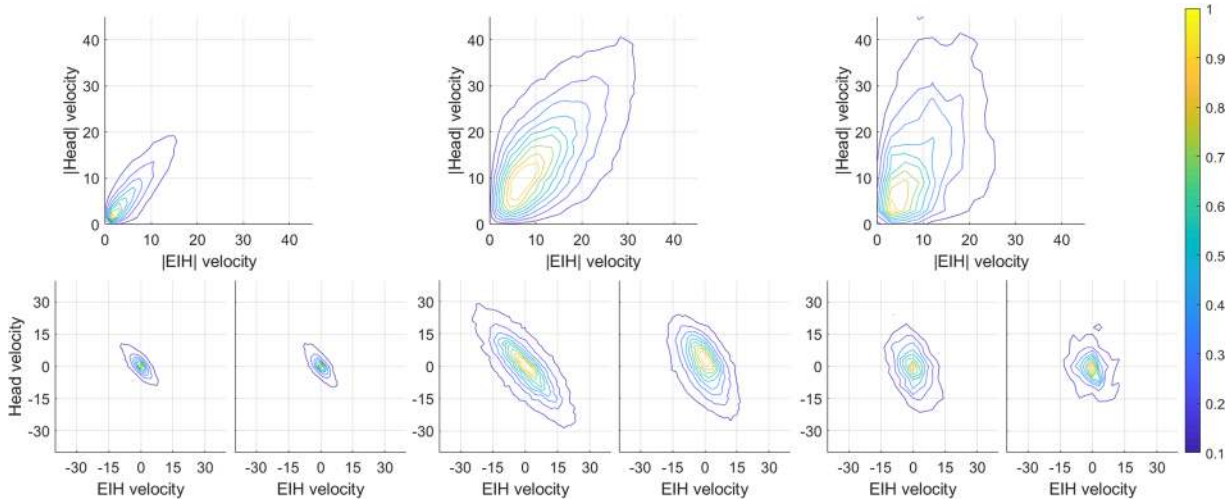

Figure 1. Eye and head movement statistics. The top row signifies absolute eye and head velocity. The bottom row signifies the distribution of eye and head velocity in the azimuth and elevation direction. The left column illustrate fixations when subjects were stationary. The middle column illustrates fixations when subjects were in translatory motion. The right column illustrates pursuit behavior. The scale on the right shows the normalized concentration of samples and is used in all figures.

- **Fixation by rotational vestibular-ocular reflex (rVOR)** - When the subject and target are stationary in the world reference frame, rotational motion of the head is compensated using rVOR. Fixations are maintained by the VOR system because it has a significantly lower response lag as compared to OKR[24]. Generally, a rVOR event displays near unity gain unless it is modulated due to other compensatory mechanisms such as OKR or pursuit.
- **Fixation by translational vestibular-ocular reflex (tVOR)** - When a target is stationary in the world reference frame, image stability at the fovea during self motion or passive displacements is achieved by tVOR[29]. Unlike rVOR, wherein a counter rotation of the eye in head rotation can stabilize the entire retinal image, tVOR cannot accommodate for the entire visual field due to the large range of optic-flow motion experienced at different depth planes. Primarily a foveal image adjustment mechanism, it follows that properties of tVOR motion depend on the gaze direction and can be difficult to differentiate with pursuit movements[30]. OKR augments VOR to help maintain a stable image over stationary targets. Fixations are maintained by a combination of gain modulation and optokinetic stimulation[24,26,29,31]. While microsaccades may be triggered for retinal image adjustment, larger saccades during fixations signify shifts in attention or an inability of gain adjustment to compensate for motion such as observed during nystagmus. These visually driven eye movements work in synergy with tVOR[30] making them difficult to observe in everyday activities as opposed to controlled experiments which are designed to isolate their behavior.

2. **Gaze pursuit (GP)** - Also known as smooth pursuit movements[32], gaze pursuit is the visual tracking of an object that is moving through the world frame using the eyes or a combination of the eyes and head by augmenting over our compensatory systems[24]. Gaze pursuit is often interrupted by catch-up saccades in compensation of retinal error[7]. While it is somewhat trivial to identify GP events using visual imagery, it may become difficult to differentiate them with GF (for more information refer to Supplementary Fig. 1).
3. **Gaze shift (S)** - A rapid shift of gaze to a new location in the world (i.e. a saccade) using the eye or eye and head in combination.

To illustrate our nomenclature, consider a situation where a person under fore-aft motion attempts to pursue a moving target. In situations such as these, the effects of stepping are compensated using VOR in the elevation direction. Relative distance and gaze angle modulates the tVOR to maintain target image at the fovea. The moving target's retinal image motion elicits a pursuit signal punctuated by predicative saccades. The pursuit motion augments over translational VOR by modulating its gain. If the eye and head pursuit movement can be distinctly identified in their velocity traces, we would consider such an event as a gaze pursuit. However, a distant or slow moving target may induce a small pursuit signal which may not be easily identifiable over opto-kinetic stabilization of the retinal image. In these situations, we would consider the event as a gaze fixation. Note that the exocentric nomenclature enables us to define multiple concurrent coordinate systems and thus requires that we specify the reference system under analysis. In this work, the reference system is chosen during the calibration process (described in section 2.2).

Figure 2. Task selections in the GW dataset. Left to right → Indoor navigation, ball catching, visual search and tea making.

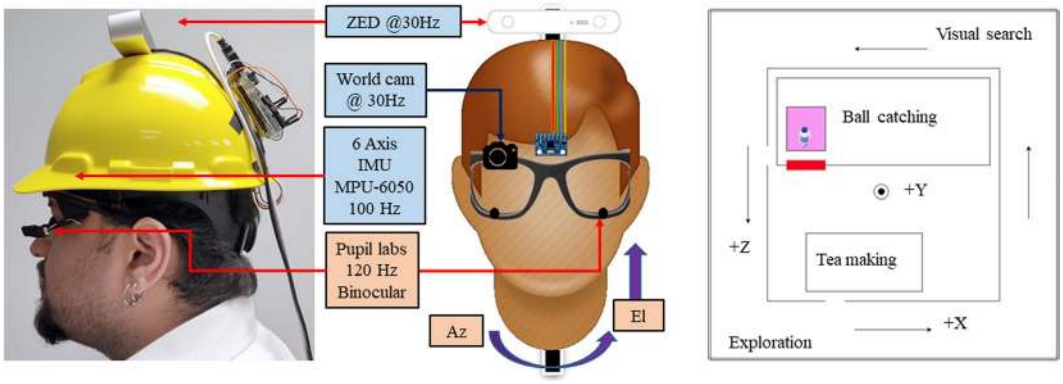

Figure 3. (left) side-view, (middle) front view of hardware setup. (right) Top view of all trajectories within our world coordinate system. The red box indicates the position of the calibration pattern. The purple box signifies the region where subject stood during calibration.

The Gaze-in-Wild Dataset

The aim of this work is to generate a dataset that captures complex ocular-motor strategies during natural tasks (see Fig. 2). We recruited 19 participants (7 female, age $\mu = 28$, $\sigma = 12.52$). Informed consent was obtained from all participants prior to hardware setup to anonymously share all data recorded from them. Identifiable people in this manuscript consent to publicly share their information as presented. All methods were carried out in accordance with relevant guidelines and regulations as approved by the Institutional Review Board at Rochester Institute of Technology, FWA-00000731. Participants were tasked with performing up to four activities while wearing an eye tracker, a hardhat instrumented with sensors, and a backpack with a laptop computer (see Fig. 3). Since task demands and interpretation have been known to guide eye movements[2], care was taken to ensure all participants received a standard set of instruction read aloud by the experimenter. Subjects were instructed to stand 1 to 2 meters away from a calibration chart within a predefined rectangular area. Once a participant was within the calibration region and facing the chart, they performed two calibration routines. After calibration was complete, participants proceeded to complete the given task. Table 1 in the Supplementary lists the calibration accuracy, tasks recorded, and the labelling status of each observer. Tasks were selected to create a wide range of head and eye poses as seen in Fig. 4. Upon completion of a task, participants returned to the calibration area to prepare for the next task. The following tasks were chosen:

- **Indoor navigation**: Subjects were instructed to walk around an indoor corridor loop twice. Indoor navigation was chosen to elicit coordinated eye and head movements that occur naturally during walking. We observed various gaze shifts to objects such as text on posters, signboards, people walking by etc. As expected, subjects made very few to no gaze shifts towards the ground due to lack of terrain complexity[11] and very little attention demands[6] for foot placement accuracy[33]. While some of the subjects were familiar with the indoor

	Fixational samples		Gaze-pursuit samples		Saccade samples	
	μ	σ	μ	σ	μ	σ
κ	0.74	0.04	0.73	0.05	0.75	0.04
p/r	0.94	0.03	0.77	0.12	0.79	0.10
F_1	0.94	0.02	0.75	0.04	0.78	0.03

Table 1. Sample based Cohens κ, precision/recall p/r and F_1 score between labellers. Note that the precision and recall values are identical (see section 3 for details).

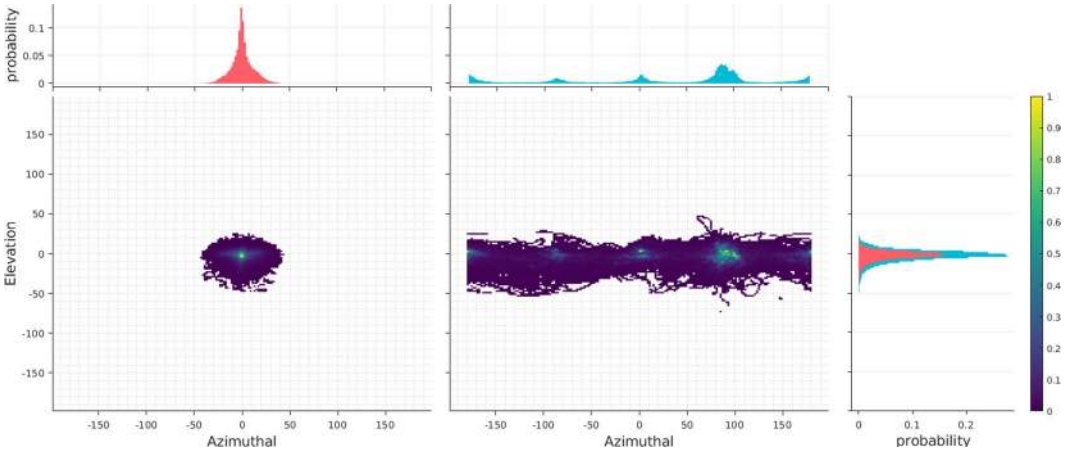

Figure 4. Head pose (right) and cyclopean eye distribution (left) in the azimuthal and elevation direction. The cyclopean eye distribution is reported in the Eye-in-Head coordinate system. Angles are provided in degrees. Note that head distribution peaks occur at 90°/*s* intervals.

corridor layout, we did not observe any noticeable difference in their behavior compared to subjects unfamiliar with the environment.

- **Ball catching**: The purpose of this task was to induce gaze pursuit behavior by asking participants to play catch with the experimenter. The experimenter would change throwing strategies in the middle of the task by either bouncing the ball on the floor, passing the ball to another experimenter or rolling the ball on the ground towards the participant. The subjects tracked the ball as a series of gaze fixations and predictive catch-up/look-ahead saccades and occasionally pursued the ball during a specific period of the ball trajectory.
- **Object search without prior subject-object interaction**: Subjects were tasked to locate and count as many objects with geometrical shapes (such as triangles, rectangles etc.) as they could find in a predetermined closed circuit corridor. This task was chosen to elicit visual search behavior in a head-free setting without biasing a subject with a particular object or shape.
- **Tea making**: As a validation for the classic tea making paradigm[34], we instructed subjects to go to the kitchen and make themselves a cup of tea. For this task, due to the close proximity of objects, relevant information sometimes fell outside the field of view.

Hardware setup and error categorization. To collect naturalistic data, we instrumented participants with an MPU-6050 6-axis Inertial Measurement Unit (IMU) mounted under a hardhat, an ATMega Arduino attached behind the hardhat, a 120 Hz binocular Pupil Labs eye tracking glasses (ETG)[35] and a ZED stereo camera (see Fig. 3). To ensure its applicability in a wide variety of domains, the Gaze-in-Wild dataset provides easy access to depth of the real world stimulus calibrated from the person's FoV. Contrary to a two IMU system[17], we chose a single IMU system to avoid using a body worn device since many applications of eye tracking are predominately head-mounted. The hardware setup weighed 700 gms (excluding laptop weight), which is similar to previous setups[8]. To reduce slippage, the hardhat was equipped with an adjustable knob to tighten its hold on a subject's head.

Pupil labs eye tracking glasses (ETG). Binocular eye trackers usually contain two eye cameras and a single world camera (which captures the scene in front of a person). Eye tracking solutions require some form of eye feature (derived from images captured from the eye camera) to Point of Regard (PoR - pixel position on the world camera) mapping to provide an accurate gaze estimate. This process is also known as eye tracker calibration. Mapping functions often vary from polynomial regression to multi-layer perceptron regression. Despite calibration, angular error tends to remain low near to the calibration region and increases radially outwards. Furthermore, there exist many sources of error which degrade the quality of gaze tracking[35], particularly in unrestrained settings. The Pupil Labs eye tracker estimates the approximate center of eye ball rotation and a 3D pose of the pupil (modeled as a 3D disc). This enables the extraction of 3D gaze vectors with respect to the Eye-In-Head (EiH) coordinate

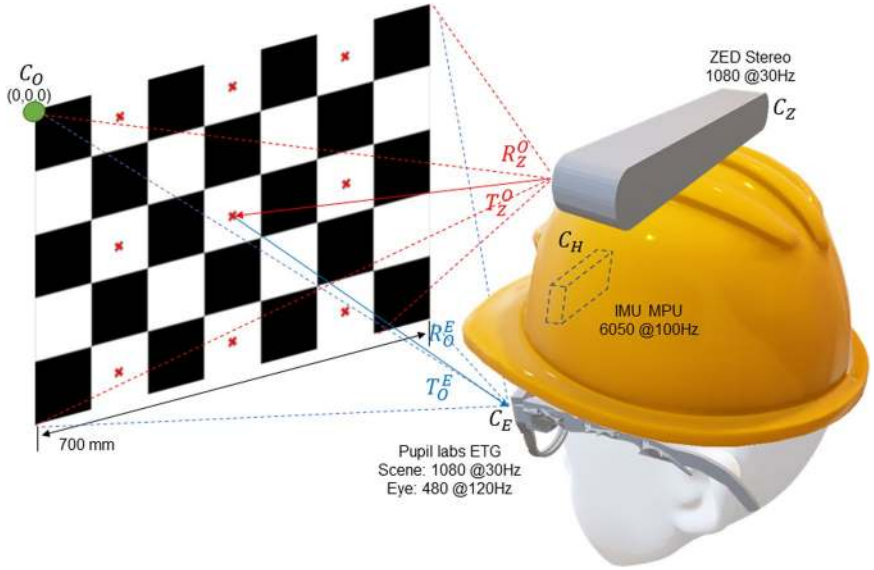

Figure 5. Checkerboard pattern placed in front of a participant during calibration phase. The red cross marks are used to calibrate the eye tracker in routine 1. The checkerboard corners are used for a 2-way multiview calibration between the ZED stereo camera C_Z and the eye tracker world camera C_E. $(R_Z^O, T_Z^O) \rightarrow$ Transformation needed to move from C_Z to the calibration chart's coordinate system, C_O. $(R_O^E, T_O^E) \rightarrow$ Transformation needed to move from C_O to C_E.

system C_E. The Pupil Labs software (version number 1.8.26) also provides a confidence value for each gaze sample which can be interpreted as a reliability measure. It is calculated as a ratio of the number of support pixels to the number of pixels on the ellipse fit of an imaged pupil. Support pixels are the edge points within a threshold distance away from the pupil ellipse fit. All gaze samples with confidence below 0.3 were discarded from analysis.

Inertial measurement unit (IMU). The MPU-6050 is a low cost 6-axis IMU that integrates a 3-axis accelerometer and a 3-axis gyroscope to estimate its pose relative to its initial position at the onset of data acquisition. The IMU is connected to an Arduino placed behind the hardhat, which in turn, is connected to the laptop backpack. The Digital Motion Processor inside the IMU provides its pose estimate at 100 *Hz*. The I^2 Cdevlib open source library was used to extract information from the IMU[36]. Pose estimates using an IMU sensor are known to drift due to error accumulation making it necessary to offset the IMU regularly to avoid drift in orientation measurements. Calibrating the IMU's offset at the beginning of data collection and fine tuning during post processing ensures accurate head pose within 7° ($\sigma = 8.34°$) of error for short recordings. Longer recordings may incur significant error in pose estimates unless externally corrected or reduced using a secondary sensor. Frequent head turns may also lead to an increase in head pose error so it is a good practice to reset the IMU following a few head turns[17]. While we do not hinder participants mid task, pose estimates for certain recordings (marked with γ in Supplementary Table 1) were manually corrected by a rotation operation before and after each heading change during post processing. Head angular drift and deviation in orientation are measured for all participants by the difference in head pose at the beginning and end of a task. We evaluated the sensor drift to be 0.021°/*s* ($\sigma = 0.035$) on average. Per participant drift can be found in Supplementary Table 1.

ZED Stereo camera. The ZED stereo camera provides a 1080p point cloud at 30 *Hz* which is calibrated and mapped onto the ETG coordinate system C_E from its own coordinate system C_Z. We found the error in depth measurement to be proportional to the distance under consideration. The euclidean 3D error was found to be less than 0.5 *m* at a distance of ~10 *m* (beyond that is considered to be infinity), which is in agreement with other independent analysis[37].

System calibration. All measurements in the GW setup are reported in reference to a modified checker chart which is fixed in the world coordinate system (see Fig. 5). Prior to data collection, we instructed the participants to perform two calibration routines before each task.

Routine 1 - This is the native offline calibration routine offered by Pupil Labs version 1.8.26 (i.e. *calibration using natural features*) following 3D pupil detection and gaze mapping. This routine required that subjects looked sequentially at red calibration targets placed in alternating boxes on the modified checkerboard chart.

Routine 2 - In the second routine, participants were asked to maintain a comfortable head pose while fixating on one of the calibration targets. They were then asked to move their heads horizontally or vertically while maintaining fixation at that point, thus inducing a vestibular ocular response. This routine performed a system calibration by aligning all hardware components to a common world coordinate system.

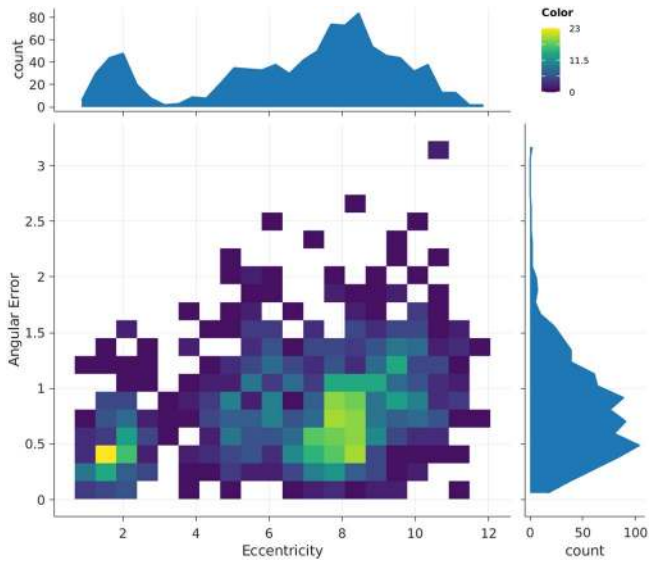

Figure 6. Eye tracker accuracy vs eccentricity from the center of the calibration pattern. Color scale indicates the number of calibration samples from all subjects.

Pupil labs eye tracker calibration. The angular error between the gaze POR and the location of the red calibration target within the world camera imagery is presented in Fig. 6. This measure reflects $\angle(k_e^{-1}P_x,\ k_e^{-1}P_c)$, where P_x and P_c are the homogeneous coordinates of the red calibration target and gaze PoR. k_e is the intrinsic matrix of the ETG world camera. We evaluated the calibration accuracy to be within 1° of error within 10° from the center of the calibration pattern. Individual participant eye tracker calibration error can be found in Supplementary Table 1. The ETG eye camera has manual focus lenses which were readjusted for every participant to ensure sharp visual features.

Temporal alignment. Each individual component of our system has a fixed temporal offset from each other. This temporal offset is removed using normalized cross-correlation of the angular velocity traces between the IMU, ETG and the ZED stereo camera. Since the ZED camera utilizes visual odometry to derive a pose estimate, it is not uncommon to observe a poor pose estimate during the VOR calibration routine. In those situations, we tracked the checkerboard corners in the ZED and ETG world camera to derive a velocity estimate for each corner point. In the absence of ZED pose information, these velocity estimates were used to compute the offset between ETG and ZED using cross-correlation as described in the next section. It should be noted that there exists an inherent latency between head and eye movements during a VOR[38]. However, we remove all latency while correcting for temporal offsets (including biological latency).

ETG-IMU calibration. Initially, the IMU and ETG are defined in their own respective coordinate systems, C_H and C_E. When participants were fixated at a point on the calibration chart during Routine 2, their eye and head pose was defined as the Z axis of our new world coordinate system C_W using rotation operations. The IMU is placed approximately 1-2 *cm* above the cyclopean gaze origin (an imaginary point midway on the line joining both eye centers). Instead of correcting for translation offset (which can vary by subject), we choose to align C_H and C_E to C_W solely using rotation matrices R_H^W and R_E^W. These matrices were initially derived using vector rotations and manually fine tuned until the coordinate systems were satisfactorily aligned (Gaze-in-World (GiW) velocity, *i.e.*, the head compensated cyclopean eye velocity is minimized). Once the head and eye orientation are defined in C_W, we rotate the EiH vector using the updated head pose to obtain the GiW vector.

ETG-ZED calibration. Calibrating the ETG and ZED is required to register the depth point cloud from ZED's coordinate system C_Z to the ETG scene camera C_E, to obtain calibrated depth values of the visual field. The visual field is defined from the center of the world camera, hence we choose to superimpose the depth map onto the world imagery. Since the distance between each checkerboard corner point is known, we can produce a grid of corner points in world units (*mm*) defined in the checkerboard coordinate system C_O. This grid can be aligned and projected on C_E and C_Z using extrinsic parameters (R, T). Corner points extracted from time synced ETG world and ZED left camera images were used to find R and T. The extracted image points and the checkerboard grid are related using $x_Z = k_Z(R_O^Z X_O + T_O^Z)$ and $x_E = k_E(R_O^E X_O + T_O^E)$. Here, X_O is the 3D checkerboard grid defined in C_O. k_Z and k_E are the left ZED and world camera intrinsic matrices. For detailed information regarding this process, we refer the reader to single camera calibration, part 1, multiview geometry by Hartley *et al.*[39]. The transformations required to align C_Z to C_E can be derived as $R_Z^E = R_O^E R_O^{Z-1}$ and $T_Z^E = T_O^E - R_Z^E T_O^Z$, which are used to transform the depth point cloud from C_Z to C_E. Once we have an aligned depth map, we trace a ray from the ETG world camera center to a subject's PoR and intersect it with the transformed point cloud to derive a 3D PoR in *mm*.

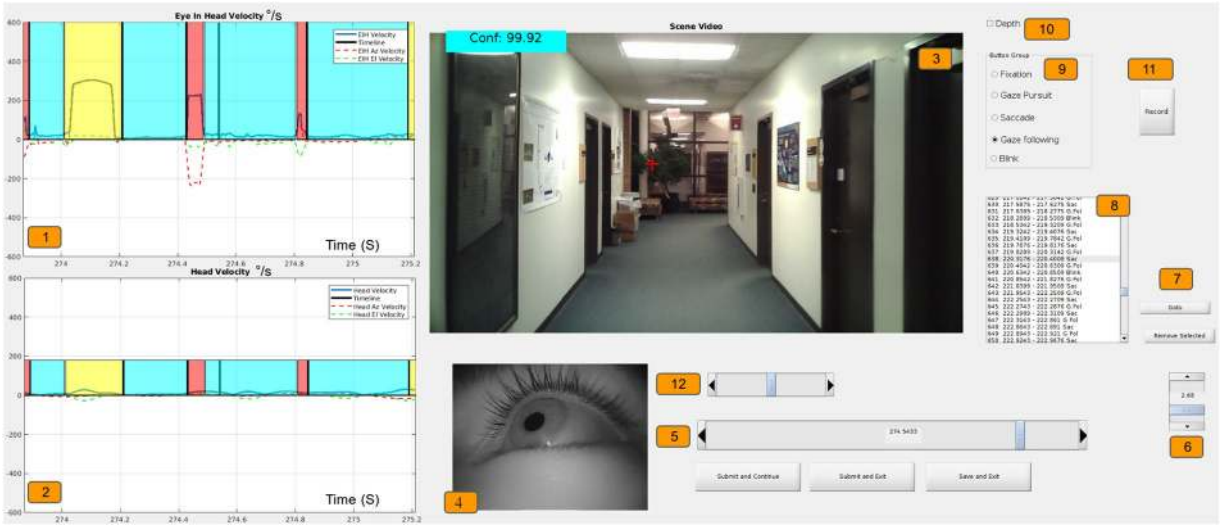

Figure 7. Custom made GUI for labelling. 1: Magnitude of EiH velocity with *Az* (Azimuthal) and *El* (Elevation) velocity traces (°/*s*). 2: Magnitude of Head velocity with *Az* and *El* velocity traces (°/*s*). 3: World-view overlaid with the Point-of-Regard (PoR) and confidence score. 4: Eye-view. 5: Slider to move a window through a recording temporally. 6: Slider to change the window width. 7: *Go-to* and *Remove button* for labelled regions. 8: Interactive list of labels in a session. 9: Radio buttons to select event type and mark across a region. 10: Toggle scene and depth view. 11: Record a 10 second clip of GUI starting at the current sample. 12: Slider to shift labels forward or backward.

Operations. All absolute angular velocity measurements (i.e. magnitudes) are calculated using a modified Two-Point Central Difference algorithm (2-P)[40]. The angular velocity ω_v can be derived as $\delta\theta/\delta t$, where $\delta\theta$ is given by $\angle(v_{n+1}, v_{n-1})$. Here, v_n is a normalized unit direction vector while t_n is the timing associated with sample n. $\delta\theta$ is the angular displacement within the elapsed time. For a fixed sampling rate f_s, $\omega_v = f_s\delta\theta/2$.

Pupil tracking is usually performed in the near infrared because the human iris, regardless of color in the visible spectrum, reflects well in the near infrared. This ensures adequate contrast between the iris and the pupil, which is dark when illuminated off axis. However, noise may be introduced while tracking the pupil due to many external and internal factors such as varying illumination conditions, algorithmic artifacts, lack of contrasting eye features, occluded pupils etc. These artifacts may result in high frequency noise in the pupil positional signal. Consequentially, several steps were taken to filter the gaze signal. Since the eye was imaged with a sampling frequency of f_s, frequencies higher than the Nyquist frequency ($f_n = f_s/2$) were aliased into our signal as noise. To avoid aliasing, we introduced a low pass filter to suppress all frequencies higher than f_n (Kaiser window, cut-off: 58 ± 2 *Hz*), a limit well beyond that where typical saccades exhibit significant power[41]. Furthermore, the 2-P central difference algorithm results in gain suppression near f_n without exhibiting phase shifts as opposed to other non-symmetric techniques wherein signal delay is not constant. Phase offsets due to anti-aliasing filters were removed by performing Zero-Phase filtering[42]. To further reduce noise, we utilized Bilateral filtering[43] since it provides an optimal trade-off between noise removal while maintaining characteristics of eye movements (such as preserving peak saccade velocity). Non-adaptive techniques such as Gaussian filtering suppressed saccade velocity peaks while increasing their duration and potentially produce misleading characteristics which could lead to misinterpretation of eye movements. The optimal parameters for bilateral filtering were empirically derived (window length 50 *ms*, $\sigma_t = 18$ *ms*, $\sigma_r = 8.75$°/*s*). The azimuthal and elevation velocity components are calculated using small angle approximations because of numerical stability during quadrant changes. That is, $\omega^{Az} = \delta\theta^{Az}/\delta t$. $\delta\theta^{Az}$ is approximated as $sin\ \delta\theta^{Az}$. Small angle approximation results in 1 % error in measurement at 14°. Assuming a maximum human angular velocity of 900°/*s*, the upper limit for human angular displacement cannot exceed ~8° within a sample at our sampling rate of 120 *Hz*, which is within 1% measurement error.

Labelling

Training and evaluating a gaze event classification model requires labelling our dataset which is one of the major contributions of this work. The GW dataset was hand-labelled by five annotators who were trained to identify head-free gaze events. They produced over 140 minutes of hand-labelled head-free gaze behavior data. The dataset contains approximately 19,000 detected fixation events, 18,000 saccades, 1,300 pursuit events, and 3,500 blinks. Using a custom labelling tool (see Fig. 7), labellers had access to eye images, scene images with PoR cross-hair, and the individual head and eye velocity traces. Using our tool, one minute of recorded data requires 45–60 minutes of annotator time. While it is possible to develop tools that allow faster labelling[44], they may bias the labeller with automated suggestive labels. Each labeller made decisions independently and they were encouraged to leave sequences where they were uncertain of the classification untouched. These sequences, along with low confidence samples (confidence below 0.3, see section 2.1.1), were treated as unlabelled and were not used to compute statistics or to train/evaluate models. While we do observe saccades as low as 15°/*s*, we do not label microsaccades

	Fixational events		Gaze-pursuit events		Saccade events	
	μ	σ	μ	σ	μ	σ
l_2	12.84	2.25	13.39	2.82	12.96	1.99
O_r	0.91	0.01	0.92	0.02	0.74	0.03
F_1	0.86	0.04	0.75	0.04	0.89	0.04
κ	0.71	0.09	0.54	0.05	0.79	0.09
κ*	0.54	0.14	0.47	0.09	0.61	0.15

Table 2. Inter-labeller event based metrics. All metrics are reported by their mean σ and inter-subject standard deviation σ. l_2 distance of the start and end time (expressed in *ms*) of matched events using ELC. O_r is the overlap ratio between matched events using ELC. F_1 score as proposed by Hooge *et al.*[46]. Event κ proposed by Zemblys *et al.*[21]. Event κ* found using ELC event matching. For more information on each metric, please refer to Section 4.

	Fixational samples		Gaze pursuit samples		Saccade samples	
	μ	σ	μ	σ	μ	σ
Fixational samples	0.94	0.03	0.02	0.02	0.03	0.02
Gaze pursuit samples	**0.20**	0.10	0.77	0.12	0.03	0.03
Saccade samples	**0.19**	0.08	0.02	0.02	0.79	0.10

Table 3. Normalized sample based confusion matrix (created by normalizing the confusion matrix with the number of samples for each event type in the ground truth) across every recording with multiple labellers.

	Matching technique	Timing offsets	Confusion matrix	Symmetric	Threshold dependency	Reliability of timing offsets
Majority vote[48]	Sample-level majority vote	×	✓	×	×	N/A
Event F1[46]	Earliest overlapping event	✓	×	×	×	low
Event kappa[21]	Largest overlapping event	✓	✓	✓	×	low
Event error rate[21]	N/A	×	×	✓	×	N/A
ELC	Window-based matching	✓	✓	×	✓	high

Table 4. Comparison of event level error metrics.

or post saccadic oscillations due to system accuracy limitations (head compensated gaze tremor was found to be $\mu = 0.55°/s$, $\sigma = 0.32°/s$). To provide maximum flexibility to researchers, we labelled stable fixations (caused due to tremors, drift and micro-saccades) and rVOR as a single gaze event type, stationary fixation, while labelled fixations due to tVOR and optokinetic stimulation as another gaze motion category, fixation under translation (labellers used *gaze following* as a pseudonym). This enables researchers to isolate the influence of compensatory mechanisms using a variety of statistical methods.

Cohen's Kappa κ is a measure of the overall agreement between two raters classifying items into a given set of categories[45]. For a given gaze event category, *precision p* is the fraction of accurately detected samples over all retrieved samples while *recall r* is the fraction of accurately detected samples over all relevant samples in the groundtruth. These measures, along with the F_1 score (the harmonic mean of p and r) are applied by iteratively calculating agreement between each labeller and the rest of the group, and then reporting the average value. Note that the described iterative strategy results in p and r holding the same value. The average overall value of Cohen's Kappa κ was $\tilde{\kappa}$ of 0.74 ($\sigma = 0.03$, median $= 0.74$), and Cohen's Kappa is reported for each event type in Tables 1–3. Previous studies have shown that human coders exhibit a performance above 0.85 $\tilde{\kappa}$ while classifying head fixed eye movements, with a very low inter-rater variance[46]. While we have not managed to replicate such a high level of agreement, we can offer insights as to why. First, head-free gaze behavior is significantly more complex with a wide range of behaviors to be classified into the previously mentioned labelling scheme in Section 1.1. For instance, consider classification of head-free gaze behavior while attempting to catch a ball into periods of gaze fixations, saccades and pursuit. Subjects engaged in head-free gaze pursuit for a very small portion of the ball trajectory, primarily relying on a series of fixations and predictive saccades to track the moving ball. This distinction is not straightforward and can easily be overlooked during labelling. Secondly, relying on a single source of information such as visual imagery or gaze signals could lead to incorrect coding (see Supplementary Fig. 1). Signal filtering and interpolation produces artifacts which may be interpreted differently by each rater[47]. Despite the fact that we have provided multiple sources of information, it is not uncommon for a human labeller to make erroneous decisions. Lastly, while it is accepted that human coders may change their labelling strategy over time[46] and the start and end times of coded events may vary, lack of holistic task awareness could result in data misinterpretation.

	κ	G.Fix κ	G.Pur κ	Sac κ
RF	0.63	0.63	0.28	**0.74**
fRNN	0.54	0.54	0.29	0.68
biRNN	0.61	0.61	**0.37**	0.69
Human	0.74	0.74	0.73	0.75

Table 5. Sample based Cohen's Kappa score κ for each optimized classifier.

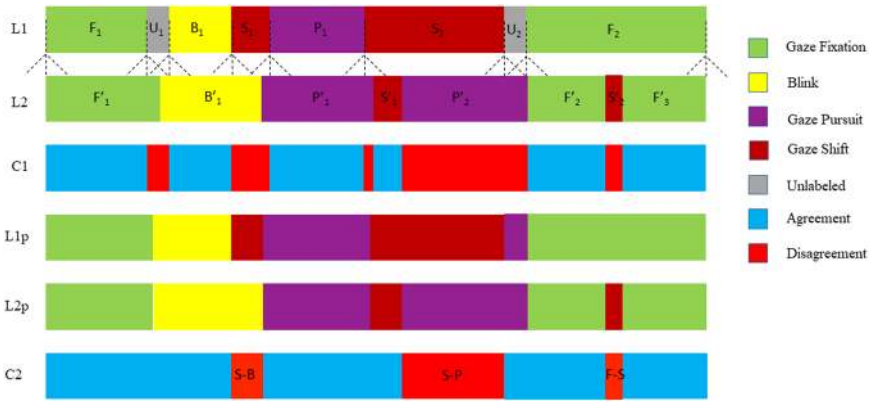

Figure 8. Illustration of the ELC metric on handcrafted test and reference sequences. (L1) Labels provided by labeller 1. (L2) Labels provided by labeller 2. Colors indicate the event type and whether labellers are in agreement. Dotted lines from L1 into L2 indicate the time window used in ELC for each transition point. (C1) Direct sample-sample comparison between labeller 1 and labeller 2. (L1p, L2p) Results of applying ELC to labels provided by labeller 1 and labeller 2 respectively. (C2) Event-level comparison between labeller 1 and labeller 2. Unmatched regions are given specific labels describing the misclassification type. For example, 'S-B' means that labeller 1 labelled the data as gaze shift whereas labeller 2 labelled the same data as blink.

Training labellers. Our labelling team was trained using lectures on eye movements, gaze interpretation and eye-head coordination from the literature to thoroughly understand the labelling nomenclature used in GW. They were then asked to label a common, very small subset of the dataset that was then analyzed and discussed as a group with the authors. The labellers began manually annotating the GW dataset following this group exercise. Individual weekly meetings with the authors were set to discuss periods of uncertain data.

Data cleaning and post-processing. To remove erroneous labels, we adapt the approach proposed by Zemblys *et al.*[20]. For our dataset, fixational events with $<0.5°$ separation between them and within 75 *ms* of each other were combined into a single event. Fixations less than 50 *ms* and saccades greater than 150 *ms* in duration were automatically removed. Finally, labelled events with duration less than 10 *ms* were automatically removed.

Error Metrics

Evaluating the performance of automated classification systems or human labellers is not straightforward. Traditional error metrics give sample-level measurements (*e.g.* percentage of individual samples correctly classified) and evaluate performance on a global basis, thus oblivious to the inherent structure of the data. For instance, metrics such as accuracy, precision, recall and F_1 score are widely used to evaluate the performance of head fixed gaze classification algorithms[20,47,48]. For evaluating agreement level among labellers or classifier performance with unbalanced data (large variation in the number of samples per class), accuracy based error metrics suffer from the *Accuracy Paradox*[21] which means that a predictive model with high sample level scores might have a lower event prediction ability. Powers observed that symmetric kappas (*e.g.* Cohen's kappa), which are designed for inter-rater metrics, may not be directly suitable for automated classifiers[49]. Sample based measures fail to account for any temporal structure and may not reflect the severity of misclassifying a few, albeit structurally important, samples. Furthermore, it is more intuitive to reason in terms of correctly/incorrectly classified collections of continuous samples of the same class, or *events*.

Event based metrics were designed to compensate for the limitations of sample based evaluation methods. Hoppe *et al.* provided the percentage of correctly classified events by comparing the samples within the bounds of each groundtruth event. The category with the highest number of samples was matched with the reference event[48]. Hooge *et al.* proposed a set of evaluation metrics such as the event-level F_1 score, the relative timing offset (RTO) and the relative timing deviation (RTD) between matched events[46]. To compute the F_1 score for a particular gaze movement category, they treat every other category as a common opposite category. However, this operation removes all inter-category confusion. The first overlapping testing event of the same category as the groundtruth is considered as matched. Temporal offsets between event start and end times are calculated for

all matched events, providing the added benefit of a measure for temporal alignment quality. Zemblys *et al.* proposed the event error rate (EER), which is a length normalized Levenshtein distance between event sequences[21]. Zemblys *et al.* also proposed the event-level Cohen's kappa measure, an extension of the event-level F_1 score[21]. These proposed event level metrics use the standard available measures (F_1, Cohen's κ) but vary in their event *matching* scheme. Differing from Hooge *et al.*, Zemblys *et al.* proposed that a testing event with the highest overlap ratio with a groundtruth event is to be treated as a match. Note that events of differing categories may also be considered as *matched*. This results in an event level confusion matrix which is used to generate an overall and per category Cohen's kappa score. Existing event level metrics improve the way we evaluate the performance of temporal classifiers but have their own individual shortcomings for varying scenarios. For instance, the majority vote method gives no penalty to unexpected short events that split longer events, and significantly influence the statistical distribution[21,48]. The event level F_1 score also does not support multi-class evaluation[21], and the EER measure does not match events and treats all event sequences as strings. It does not consider or provide insight into temporal offsets. Furthermore, it also suffers from the *Accuracy Paradox* and only returns a single value as an overall rating. Last but not least, different event-matching procedures significantly affect the RTO and RTD measurements. Zemblys *et al.* identified that the RTO and RTD measures will be compromised when using the largest overlapping event-matching strategy[21]. Similar situations may occur when utilizing the earliest overlapping matching strategy. For example, when onset of the earliest overlapping testing event is close to the offset of a reference event. Various event based metrics are summarized in Table 4. To address some of the shortcomings of previous approaches, we devised the Event Level Cross-Category Metric (ELC) as described below.

Consider the following taxonomy:

- Reference sequence - groundtruth sequence of labels.
- Testing sequence - predicted sequence of labels, usually the output of an automated classification process.
- Matched event - two events are considered matched when their start and end position roughly align in a predetermined window and meet the matching criterion (discussed below). As an example, consider sequences L1 and L2 in Fig. 8. All fixation events in L1 (marked in green) are considered as matched.
- Unmatched event - All events which do not satisfy our matching criterion are considered as unmatched. Both saccades in Fig. 8, are considered as unmatched.
- Detached event - We often find unmatched events in our ground truth which completely overlap with another test event and belong to the same gaze category. These type of events are considered to be detached. For example in Fig. 8, the blink in L1 (marked in yellow, the start point is matched whereas the end point has no match) is considered as a detached event. Researchers may safely consider detached events as matches per their strictness requirements and application (this operation would inflate the performance score of a classifier).
- Transition point - It is assumed that all event boundaries touch each other at their transition points. Transition points have samples of different gaze behavior adjacent to it. In case event boundaries do not touch, we assume the period between them to be the *none* class. All entries pertaining to *none*, i.e blinks and unlabelled periods, are removed from consideration. Note that all events have two transition points.

1. **Window-based matching:** First, we identify every transition point in the reference and testing sequence. For every transition point in reference sequence, we extend a window of a certain size (*e.g.* 50 *ms*) onto the test sequence and find all transition points within. The reference transition point is matched with the first (in time) testing transition point within the match window which satisfies a particular matching criterion. For onset transition points, the event type on the right should match the reference event type. Similarly, offset transition points are matched if the event type on its left matches the reference event type. An event is matched if both its start and end transition points are matched.
2. l_2 **distance calculation for matched events:** Following window-based matching, overall timing offsets are calculated for matched events. Unlike RTO and RTD, which are calculated separately for start and end points of events, we calculate the l_2 distance ($\sqrt{(start_1 - start_2)^2 + (end_1 - end_2)^2}$ where $start_1$ and $start_2$ is the start positions of two events, end_1 and end_2 are end positions of two events) for each event. The mean and standard deviation of all calculated l_2 distances (per class and overall) are used as indicators of alignment quality between two labelled sequences.
3. **Overlap ratio calculation for matched events:** Since events of different categories have various ranges of duration, the severity of temporal misalignment could be different for individual event types having the same timing offset values. Therefore, we calculate the *overlap ratio*[50] O_r ($O_r = n_1 \cap n_2 / n_1 \cup n_2$), where n_1 and n_2 are samples belonging to two matched events. The overlap ratio reflects the temporal alignment quality of two events. As with the l_2 distance, the mean and standard deviation are calculated and reported.
4. **Timing offsets correction:** Once the l_2 distance and overlap ratio is calculated, we remove the effects of misalignment by correcting timing offsets in both sequences. This correction is applied on all matched transition points regardless of an event's match status. For each matched transition point, timing (sample index) of two points are averaged to create a single representative transition point. If the original point is shifted away from the event center, the displaced samples are assigned the event's category. Likewise, if the transition point moves inwards, the displaced samples are assigned the external event's gaze category. If the displaced samples are unlabelled in a particular sequence, they are assigned the same gaze movement class as the corresponding sequence.
5. **Event level confusion matrix:** Comparing two labelled sequences leads to a collection of matched and unmatched events, *i.e.*, a confusion matrix, which describes inter-category event classification performance. Owing to the timing offsets correction step, event mismatches within the preset threshold are eliminated.

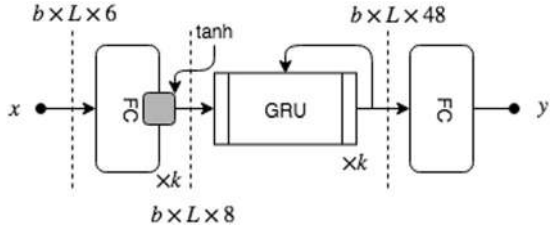

Figure 9. Bidirectional recurrent network model architecture. The model takes the magnitude, azimuthal and elevation eye and head velocity (6 features) as its input, passes through k fully connected feature extraction layers. These features are fed into a stack of k GRU layers which learn temporal patterns to classify a sample x_t. The forward variant (fRNN) outputs a 24 dimensional vector instead of 48 before being reduced to 3 at the final FC layer.

Standard metrics such as Cohen's κ and F_1 score can be derived from the confusion matrix for deeper insights or to summarize performance.

6. **Applying previous steps in both directions**: ELC is an asymmetrical event matching technique. It can be applied twice by interchanging the testing and reference sequences to find an average performance measure along with a sense of metric agreement. For instance, if the number of detached events is higher in a particular order, it provides insight into larger proportions of event merges in the testing sequence. Inter-labeller performance is computed by applying ELC both ways but not for human-classifier evaluation.

In Fig. 8, sequences L1p and L2p show the results of applying ELC to the labels in L1 and L2 respectively. The application of these rules eliminates many minor (mainly temporal) disagreements between sequences and considers only the regions of major disagreement as seen in sequence C2. Event Kappa utilizes the largest overlapping strategy to match events, which results in lower RTO and RTD scores[21]. For instance, event F_2 in L1 gets split into two shorter events F_2' and F_3' by an unexpected event S_2' in L2, the metric tends to match the fixation in L1 with the largest overlapping event (F_3' in this case). This leads to a poor RTO and RTD measures. However, ELC considers the start and end points of F_2 in L1 and matches them with the start of F_2' and the end of F_3' respectively. F_2 is considered as a matched event and the testing sequence is rewarded by increasing the F/F counter in the confusion matrix. Likewise, the testing sequence is scored negatively for the offending event, S_2', by increasing the F/S counter in the confusion matrix. The l_2 distance (functionally equivalent to RTO and RTD measurements) accurately computes the alignment quality. Interchanging L2 as the reference and L1 as the testing sequence, events F_2', S'_2 and F_3' would be considered as unmatched events and l_2 distances would not be calculated.

Overall, ELC provides a faithful indication of timing offsets using the window-based matching strategy. ELC is dependent on a parameter, *i.e.*, the window size. The window size indicates the system tolerance for timing offsets between ground truth and testing events. Since it's easier to identify the start and end points of gaze shifts as compared to other types of gaze events, different window sizes for gaze shift related events ($\pm 25\,ms$) and non gaze shift related events ($\pm 35\,ms$) are used. Researchers may consider using larger window sizes for situations wherein event onset and offsets conditions are relaxed.

Machine Learning for Gaze Event Classification

We trained two standard machine learning models for gaze event classification: a moving window based method and a recurrent neural network (RNN). The input to both classifier models is a sequence of temporally discrete sensor data vectors, i.e., $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_T\}$, where $\mathbf{x}_n \in \mathbb{R}^d$ and n is the current time step. As described in Section 5.2, these data vectors contain information from the IMU and eye tracker. For both models, we merge fixations when stationary and fixations under translation into a single gaze fixation class (see Section 1.1 and Section 3).

Classification models. The moving window model classifies a gaze sample at time n by aggregating information from a window of data vectors adjacent to $\mathbf{x}_n$, i.e., $\mathbf{w}_n = W(\mathbf{x}_{n-s}, \ldots, \mathbf{x}_n, \ldots \mathbf{x}_{n+s})$, where the vector of window features $\mathbf{w}_n \in \mathbb{R}^g$ is computed using a window size of $2s+1$ samples and the function $W(\cdot)$ computes the windowed feature vector. We chose the random forest (RF) classification algorithm since it works well for low-dimensional data, and our framework resembles state-of-the-art gaze event algorithms for controlled 2D environments[21]. RF is an ensemble learning method wherein multiple decision trees are trained on a subset of samples and their feature space[51]. A RF is easy to train and they are robust to noise and over-fitting, which are common problems for decision trees. For gaze classification in 2D controlled environments, Zembyls *et al.* showed that RF performed well with only 16 trees and 10-dimensional features up to a 200 ms window[21]. In our experiments, we use 40 trees, a minimum leaf size of 30, and we use $\sqrt{g}$ randomly selected features per tree where g is the number of features for a given window size. To improve efficiency during the process of training the window-based RF classifier, we removed duplicated $\mathbf{w}$ vectors (samples with equal value up to the second decimal). These duplicates were instead represented by a single sample that was upweighted by the number of duplicates found (e.g. the confidence measure was scaled). No duplicates were removed from the test set.

Rather than using explicit windows, the RNN model operates on the velocity data stream, i.e the absolute, azimuthal and elevation velocity (see Section 2.3). We use two variations of the RNN model. Our one directional

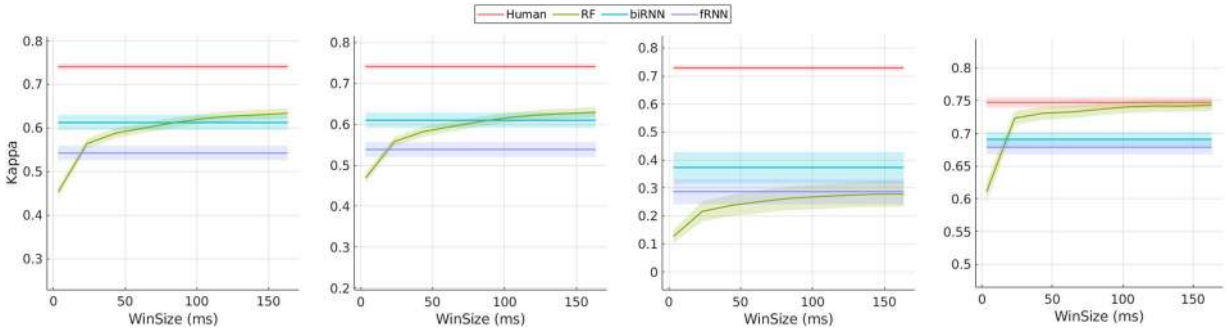

Figure 10. Sample level performance metrics. All performance curves are centered around their mean, $\mu\pm$ standard error. Left - Overall κ score. Inner left - Gaze fixation κ score. Inner right - Gaze pursuit κ score. Right - Saccade κ score. Please note the varying y-limits to accentuate the difference in performance. RNN uses memory to encode temporal patterns, and hence the RNN architectures are represented as horizontal lines as they do not operate in window sizes. We would like to highlight that all window sizes are in the velocity domain. Window sizes in angular domain can be derived by adding 10 *ms* (please refer to Section 2.3).

forward RNN model (fRNN) classifies the gaze at time n using only past and present information, i.e., $F(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. This model would be especially useful for real-time gaze prediction. For offline processing, we also use a bi-directional RNN (biRNN) that has past, present, and future information as input, i.e., $F(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots \mathbf{x}_T)$. Both models are implemented with gated recurrent units (GRUs)[52], which can handle longer-term dependencies than simple RNNs. A similar approach was used by the GazeNet architecture[21], which used an RNN to classify events in a controlled 2D environment. To prevent the over-representation of samples that were labeled by N labellers (where $N > 1$), these samples were weighted by 1//N during the process of training.

The input to our model is a subset of windowed features W. Specifically, the model accepts the absolute, azimuthal and elevation EiH and head velocity as input. Multiple sequences, b, are stacked into a single batch of data. All sequences were padded with zeros to be of the same length as the longest sequence present in the batch, L. This $b \times L \times 6$ dimensional data passes through k fully connected layers which generates a nonlinear representation of EiH and head velocity. Extracted features are fed into a stack of k GRU (see Fig. 9) units with a dropout probability of 10 % which learn to associate temporal patterns with a type of gaze behavior. We use a combination of cross-entropy and generalized Dice[53] loss functions. The network was optimized using ADAM[54] for 175 epochs with a learning rate of 0.001, which we reduced linearly as the training performance improved. We experimented with the number of recurrent and linear layers and found $k = 3$ worked best. All codes related to GW is made publcily available.

Input features. The $\mathbf{x}_n$ features consist of normalized EiH vectors v_e and head vectors v_h concatenated together. For the window-based RF classifier, for each time step n, we extract the following set of handcrafted features from a window of size $2s + 1$ around the n-th sample:

1. **Mean EiH and head angular distance**: $\Delta\theta_e$, $\Delta\theta_h$. Angular distance in degrees between the mean EiH/head vector of s samples before and after the current sample of interest, x_n.
2. **Deviation in EiH and head velocity**: σ_e, σ_h. Standard deviation of magnitude of EiH and head angular velocity.
3. **Confidence**: We supply the confidence measure (see Section 2.1.1) to our classifiers as weights for each sample. High confidence and duplicate samples are assigned larger weights.

We also aggregate velocity measurements from every sample in the window. The velocity measurements are the absolute EiH $|\omega_e|$ and head velocity $|\omega_h|$ (angular velocity extraction has been described in Section 2.3), azimuthal EiH and head velocity $\omega_e^{Az}, \omega_h^{Az}$ and elevation EiH and head velocity $\omega_e^{El}, \omega_h^{El}$. All velocity measurements are expressed in °/*s*. Azimuthal and elevation velocity contain directional information such that a positive sign indicates a clockwise rotation and vice versa. Assimilating features as a time series results in a g dimensional window feature vector, where $g = 4(2s + 1) + 6$. The full window feature vector is given by $\mathbf{w}_n = ([|\omega_e|,\ |\omega_h|,\ \omega_e^{Az},\ \omega_h^{Az},\ \omega_e^{El},\ \omega_h^{El}]_{n-s}^{n+s},\ \Delta\theta_e,\ \Delta\theta_h,\ \sigma_e,\ \sigma_h)_n^T$, where [*] stands for aggregation.

Results. The two classifiers are assessed using *leave-one-out* cross validation by testing on a single person's data (the holdout subject) and training the model on remaining subjects. This process is repeated for subjects 1, 2, 3, 6, 8, 9, 12, 16, 17, 22. For each of these tests, certain steps are taken to prevent overfitting. Training data for the holdout subject is split into five folds with approximately equal frequencies of saccades and fixation gaze events (but an unequal number of frames) per fold. However, due to the low frequency of pursuit events within the dataset, and their unequal distribution across subjects, we did not divide gaze pursuit events between folds. For each holdout subject, four folds were used to train the RNN model while the single fold was used to fit model parameters, which were then saved. Following conversion, the parameters for the best performing model on the single validation fold were saved. The iterative process results in five sets of converged parameters per holdout

	G. Fix F_1	G. Pur F_1	Sac F_1	EER	G. Fix κ	G. Pur κ	Sac κ	Overall κ
RF	0.74	0.26	0.82	0.26	0.46	0.01	0.63	0.32
fRNN	0.74	0.22	0.81	0.30	0.61	0.22	**0.69**	0.47
biRNN	**0.80**	**0.35**	0.83	**0.16**	0.61	**0.27**	0.67	0.47
Human	0.86	0.75	0.89	0.14	0.71	0.54	0.79	0.62

Table 6. Metrics based on various event matching techniques proposed by others. Event based F_1 score proposed by Hooge *et al.*[46]. Event Error Rate (EER) proposed by Zemblys *et al.*[21]. Event and overall κ scores calculated using Zemblys *et al.*[21].

		G.Fix				G.Pur				Sac			
	Overall κ	κ	O_r	l_2	$l_2\,\sigma$	κ	O_r	l_2	$l_2\,\sigma$	κ	O_r	l_2	$l_2\,\sigma$
RF	0.37	0.31	0.93	**12.97**	2.10	0.03	0.88	15.15	3.44	**0.54**	0.75	**12.80**	2.16
fRNN	0.27	0.21	0.90	15.09	3.32	0.03	0.89	15.70	3.44	0.44	0.69	15.01	3.42
biRNN	0.37	**0.34**	0.92	14.93	3.64	**0.14**	0.90	**14.25**	3.84	0.44	0.71	14.73	3.62
Human	0.56	0.54	0.92	13.85	3.64	0.47	0.89	15.23	3.84	0.61	0.73	13.67	3.62

Table 7. Standard metrics derived from the ELC confusion matrix. O_r is the overlap ratio between matched events. $l_2 - \sigma$ distance between matched event start and end times and their standard deviation $l_2 - \sigma$ in *ms*. $l_2 - \sigma$ and $l_2 - \sigma$ are similar to RTO and RTD metrics proposed by Hooge *et al.*[46].

Cohen κ	Overall	G.Fix	G.Pur	Sac
biRNN (only eyes)	0.56	0.55	0.24	0.71
biRNN (only absolute)	0.58	0.57	0.33	0.71
biRNN	0.61	0.61	0.37	0.69

Table 8. Sample based κ score after removing either head movement information or directional information.

subject. The best performing set of parameters on the holdout subject is accepted as the optimal set of weights for that model type, and for subsequent comparison against other model types. One notable exception to this procedure is the RF algorithm, which does not require a validation set. Instead, its parameters were chosen to maximize its performance while maintaining a manageable model footprint ~50 mega bytes.

Classifiers are evaluated using both sample and event level metrics (see Section 4). Classifier output is not evaluated during blinks or for unlabelled data points. As the window size increases, RF gains increasing temporal awareness which results in higher κ performance with diminishing returns. It can be observed in Fig. 10 that RF arrives at an asymptotically improving performance with a window size of 30 *ms* and above. Individual κ scores for each gaze class reveals that all classifiers find it difficult to distinguish gaze pursuits. Overall, sample based metrics convey that RF with a large window size outperforms RNN for detecting saccades but performs poorly on gaze pursuit samples (Table 5).

We report event based metrics and observe that biRNN outperforms RF on all measures. Interestingly, event F_1 and event κ scores computed using Zembyls *et al.* shows an increase in saccade classification performance (see Table 6) for biRNN over RF. However, this increase is not reflected using sample based metrics (Table 5) or event κ computed using ELC (Table 7). Notably, RF outperforms RNN based methods in l_2 scores, indicating a better ability to produce tighter fits around saccades (see Table 7). Overall, gaze pursuit classification baselines fall short on human level performance but the results are consistent with the difficulty in classifying pursuit movements over other gaze movement types in general[19,55,56]. The biRNN model produces higher ratio of detached events (G.Fix: 0.21, G.Pur: 0.20, Sac: 0.05) as compared to RF (G.Fix: 0.09, G.Pur: 0.02, Sac: 0.11) which indicates that a larger number of ground truth events completely overlapped with events of the same category but their transitions did not fall within the matching window. Since it is debatable if these detached events can be considered as matches, we omit them from all measures to avoid inflating scores.

Ablation study. To understand the role of each feature, we systematically removed essential components from the best performing model (biRNN with 3 FC and GRU layers). The input to biRNN comprises of absolute EiH/head velocity, azimuthal and elevation EiH/head velocity. This generates a signal with 6 features. Please note that azimuthal and elevation velocity store relative direction information between the eye and head (−ve sign means an anticlockwise rotation). By comparing different conditions using sample based κ score, we highlight the essential components required for head-free gaze classification in Table 8. For a detailed comparison using all metrics, please refer to Supplementary Table 2.

As expected, the performance of biRNN with EiH information (only absolute eye velocity) did not vary while detecting gaze fixations and saccades, but drops by 35 % (0.37 → 0.24) while detecting pursuit events.

Interestingly, a few pursuit events were still detected despite the lack of head movements. This indicates that head-free eye movements during pursuit behavior show a varied velocity pattern than gaze fixations and can be differentiated without any knowledge of head motion. We also observe that there is a minor loss in performance of 10 % (0.37 → 0.33) when we remove azimuthal and elevation components. This highlights that absolute velocity information alone can provide reasonable certainty for classification.

Discussion

The main purpose of this work was to build the first dataset of labelled gaze movements collected during natural behavior 'in the wild' (outside of the laboratory), to have multiple labellers manually label the gaze events in the dataset, and to showcase the performance of two standard temporal classification techniques, Random Forest and Recurrent Neural Networks, using some common evaluation metrics. To overcome incorrect inter-event timing offsets observed in existing metrics, we introduce the ELC metric. The usefulness of a classifier lies in its ability to generalize in unseen circumstances. Hence, all our baseline performances are evaluated using the *leave-one-out* approach, wherein a classifier is tested on a single subject's data while trained on the rest. Despite the fact that there is variability among human labellers, there is as yet no other choice, so we rely here on their labels as the gold standard. To improve upon existing event metrics and provide a reliable measure of alignment quality, we devised a new event matching technique, ELC, which matches events based on their transition points. ELC provides some control on evaluation strictness by identifying events which belong to the same category but are not temporally aligned due to event fragmentation.

Lower gaze pursuit classification performance by classifiers. The best performing classifier for gaze pursuits is 49 % lower than the average human level performance (sample κ: 0.73 → 0.37) whereas fixation and saccade performance achieves an average of 87 % of human performance (sample κ G.Fix: 0.61 → 0.74, Sac: 0.69 → 0.75). While pursuing moving targets, we observed that participants seamlessly interchanged between fixational and pursuit movements. The distinction between these movements are difficult to observe, especially during low velocity conditions because small angular errors in orientation measurements (a phenomenon common with IMUs) could result in misinterpretation without additional context for consideration (such as scene imagery with overlaid gaze PoR), a modality currently unavailable to our classifiers. Distinction between gaze fixation and pursuit events is further compounded when the head tracks a moving target or makes anticipatory movements but gaze remains stable at a fixation point. This motion elicits a signal similar to VOR, but if we rely purely on visual inspection then these events can easily be confused with pursuit motion. Situations such as these, combined with minor orientation errors, largely contribute to fixation/pursuit confusion seen in Table 5.

Head tracking: A pursuit or fixation?. Previous research has shown that the head *tracks* a moving target (in our case the ball) while the eyes *predict* the ball location using predictive saccades[57]. We find numerous instances of gaze shifts to known targets where head movements precede eye movements in an anticipatory manner[58] to ensure that upcoming eye movements do not deviate too far from the relatively tight distribution seen in Fig. 4. Participants frequently showed tracking behavior with the head and predictive or catch-up motion with the eyes during early phase of the ball trajectory. This behavior is usually followed by gaze pursuits during the next phase, i.e. the ball height is peaked and its projected retinal velocity is low. Following the peak phase, participants made predictive saccades to their hand for successful ball interception. GW also captures instances where the head *catches-up* to the fixation location while maintaining a strict coupling with the ball trajectory. While some may argue that head tracking of a moving object constitutes a pursuit motion, we instructed labellers to mark those sequences as fixations because the signals are identical to a VOR (please refer to Supplementary Fig. 1).

Head and eye tracking can have different coordinate systems. Based on the ablation study, we observed that providing only the absolute velocity information achieved almost the same performance as biRNN-3, our best performing model. Interestingly, it highlights that for a slight drop in performance, future end-end classification frameworks may perform reasonably well if they simply provide unaligned eye and head motion information. While gaze fixations and saccades are distinctly identifiable using only eye-in-head (EiH) information, pursuit movements would be difficult to differentiate with a fixation without head movement information. As a sanity check, we also verified that the presence of a head tracking device improves classification of head-free pursuit movements by up to 35 % as opposed to without head movement information (sample κ: 0.24 → 0.37). It is interesting to note that despite removing head movements, the RNN classifier is still able to identify a few pursuit events which indicates that they demonstrate different EiH velocity statistics as fixations (for more information, please refer to Supplementary Table 2).

Gaze-in-world information for classification. We include head pose as an input modality for the classifiers. While it is possible to classify the gaze-in-world signal, which is the head compensated eye-in-head signal, we wanted to train algorithms which could directly capture eye and head movement dynamics along with classifying it. For instance, we often find gaze pursuit events which are dominated either by head or eye movements, a distinction which would be lost when classifying gaze-in-world information.

General limitations. Given limitations in current technology, it is unavoidable that tracking head position using a low cost IMU will accumulate error over time. All task duration were ~3 minutes long and the error in orientation at the start and end of a recording was found to be 7° on average (see Section 2.1). While this error affects the absolute velocity component by a very small margin (0.04°/*s* on average), it leads to unwanted shifts in the azimuth and elevation velocity component (see Supplementary Fig. 1). Despite the use of a ratcheted head strap, this error accrues, in part, due to slippage of the helmet on the head, which will cause a misalignment of

the helmet-mounted ZED stereo camera and the Pupil Labs eye tracking glasses (see Section 2). Future work might further reduce slippage through using software correction, such as the estimation of rotational slip on a frame-to-frame basis by matching visual features in the stereo camera and Pupil Labs world camera imagery, or through the fusing visual pose estimates with IMU data, as is commonly used in simultaneous localization and mapping.

Limitations of event-based metrics. Although event level error metrics give researchers a better idea of the actual performance of automated classifiers or agreement level between labellers, existing event level metrics suffer from various drawbacks. The majority vote metric by Hoppe *et al*. remains agnostic to the testing sequences' structure. It does not penalize during event fragmentation caused by unexpected short events in the testing sequence[21,48]. Moreover, this metric could be biased by the distribution of samples. Event level F_1 score does not work well in multi-category scenario[21] and gives out unreliable RTO and RTD. EER does not provide any measure of alignment quality and suffers from the *Accuracy Paradox*[21]. Event matching techniques based on the largest overlap ratio, such as the event κ proposed by Zemlys *et al*. do not provide a reliable measure of alignment quality[21]. ELC overcomes these issues by matching events whose transition points fall within a window. A potential drawback of ELC is its dependency on the window size. Although the window size could be carefully chosen for different types of events and transitions, the metric could generate different results due to varying window sizes. For example, if a small window size was chosen, ELC would have a lower tolerance for transition ambiguity between certain event types which could result in higher misclassification scores. Furthermore, ELC is not symmetrical. To alleviate that, we propose that metrics derived using ELC should be averaged when used to evaluate inter-coder performance. While ELC overcomes certain drawbacks from previous evaluation techniques, new event level metrics are needed which accurately reflect performance, is symmetric in nature, provides a reliable measure of temporal alignment quality and is independent of an external threshold.

Conclusion

This work introduces GW, a large-scale dataset for studying eye and head coordination in naturalistic conditions. Participants were asked to perform four tasks without constraining them in any manner and were free to accomplish the tasks in any manner they chose to. Approximately 2 hours and 15 minutes of gaze behavior was manually hand coded by multiple human annotators and used to train gaze classifiers. We benchmark the performance of two machine learning algorithms for classifying these events and found that both achieved near human level performance for detecting gaze fixations and saccades, but they found it difficult to distinguish gaze pursuit behavior without additional contextual information otherwise available to human coders. In an effort to produce intuitive measures for event level similarities between two sequences, we propose the ELC event matching algorithm. We verify that all commercial eye tracking solutions could benefit in classifying head-free gaze pursuit movements by including a low cost IMU. Furthermore, comparable results are observed when head-free gaze movements are classified purely based on absolute velocity information of the eye and head, which indicates that head-free gaze classification is possible without aligning the eye and head coordinate systems.

Permission to share facial imagery. All identifiable people in this manuscript consent to share their information as presented.

Data representation. All statistical figures were generated using Gramm, an open source software for data visualization[59]. All metrics reported are rounded to the second decimal.

Data availability

Compressed data and codes are publicly available at http://www.cis.rit.edu/ http://www.cis.rit.edu/~rsk3900/gaze-in-wild/.

Received: 15 April 2019; Accepted: 23 January 2020;
Published online: 13 February 2020

References

1. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis Mach. Intell.* **20**, 1254–1259, https://doi.org/10.1109/34.730558 (1998).
2. Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M. & Velichkovsky, B. M. Yarbus, eye movements, and vision. *i-Perception*, https://doi.org/10.1068/i0382 (2010).
3. Hayhoe, M. M., McKinney, T., Chajka, K. & Pelz, J. B. Predictive eye movements in natural vision. *Exp. Brain Res.* **217**, 125–136, https://doi.org/10.1007/s00221-011-2979-2 (2012).
4. Hayhoe, M. & Ballard, D. Eye movements in natural behavior. *Trends Cogn. Sci.* **9**, 188–194, https://doi.org/10.1016/j.tics.2005.02.009 (2005).
5. Sprague, W. W., Cooper, E. A., Tošić, I. & Banks, M. S. Stereopsis is adaptive for the natural environment. *Sci. Adv.* **1**, https://doi.org/10.1126/sciadv.1400254 (2015).
6. Kothari, R., Binaee, K., Matthis, J. S., Bailey, R. & Diaz, G. J. Novel apparatus for investigation of eye movements when walking in the presence of 3D projected obstacles. *In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16* **14**, 261–266, https://doi.org/10.1145/2857491.2857540 (2016).
7. Daye, P. M., Blohm, G. & Lefevre, P. Catch-up saccades in head-unrestrained conditions reveal that saccade amplitude is corrected using an internal model of target movement. *J. Vis.* **14**, 12–12, https://doi.org/10.1167/14.1.12 (2014).
8. Barnes, G. R. Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets. *The J. physiology* **287**, 127–47 (1979).
9. Freedman, E. G. Coordination of the eyes and head during visual orienting. *Exp. Brain Res.* **190**, 369–387, https://doi.org/10.1007/s00221-008-1504-8 (2008).
10. Einh¨auser, W. *et al. Human eye-head co-ordination in natural exploration*, vol. 18 (Network, 2007).

11. Matthis, J. S., Yates, J. L. & Hayhoe, M. M. Gaze and the Control of Foot Placement When Walking in Natural Terrain. *Curr. Biol.* **28**, 1224–1233, https://doi.org/10.1016/j.cub.2018.03.008 (2018).
12. Epelboim, J. *et al.* The function of visual search and memory in sequential looking tasks. *Vis. Res.* **35**, 3401–3422, https://doi.org/10.1016/0042-6989(95)00080-X (1995).
13. Fang, Y., Nakashima, R., Matsumiya, K., Kuriki, I. & Shioiri, S. Eye-head coordination for visual cognitive processing. *PloS one* **10**, e0121035 (2015).
14. Allison, R. S., Eizenman, M. & Cheung, B. S. Combined head and eye tracking system for dynamic testing of the vestibular system. *IEEE Transactions on Biomed. Eng.* **43**, 1073–1082, https://doi.org/10.1109/10.541249 (1996).
15. Kinsman, T., Evans, K., Sweeney, G., Keane, T. & Pelz, J. Ego-motion compensation improves fixation detection in wearable eye tracking. *Proc. Symp. on Eye Track. Res. Appl. - ETRA* **12**, 221, https://doi.org/10.1145/2168556.2168599 (2012).
16. Larsson, L., Schwaller, A., Nyström, M. & Stridh, M. Head movement compensation and multi-modal event detection in eyetracking data for unconstrained head movements. *J. Neurosci. Methods* **274**, 13–26, https://doi.org/10.1016/j.jneumeth.2016.09.005 (2016).
17. Tomasi, M., Pundlik, S., Bowers, A. R., Peli, E. & Luo, G. Mobile gaze tracking system for outdoor walking behavioral studies. *J. Vis.* **16**, 27, https://doi.org/10.1167/16.3.27 (2016).
18. Holmqvist, K. *et al. Eye tracking: A comprehensive guide to methods and measures* (OUP Oxford, 2011).
19. Pekkanen, J. & Lappi, O. A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Sci. Reports* **7**, 1–13, https://doi.org/10.1038/s41598-017-17983-x (2017).
20. Zemblys, R., Niehorster, D. C., Komogortsev, O. & Holmqvist, K. Using machine learning to detect events in eye-tracking data. *Behav. Res. Methods* **50**, 160–181, https://doi.org/10.3758/s13428-017-0860-3 (2018).
21. Zemblys, R., Niehorster, D. C. & Holmqvist, K. *gazeNet: End-to-end eye-movement event detection with deep neural networks.* https://doi.org/10.3758/s13428-018-1133-5 (2018).
22. Lappi, O. Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference. *Neurosci. Biobehav. Rev.* **69**, 49–68, https://doi.org/10.1016/j.neubiorev.2016.06.006 (2016).
23. Nystr¨om, M., Hooge, I. T. C., Hessels, R. S., Niehorster, D. C. & Andersson, R. Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Soc. Open Sci.* **5**, 180502, https://doi.org/10.1098/rsos.180502 (2018).
24. Land, M. F. & Tatler, B. W. The human eye movement repertoire. In *Looking and ActingVision and eye movements in natural behaviour*, 13–25, https://doi.org/10.1093/acprof:oso/9780198570943.003.0002 (Oxford University Press, 2009).
25. Barnes, G. R. & Lawson, J. F. Head-free pursuit in the human of a visual target moving in a pseudo-random manner. *The J. physiology* **410**, 137–55, https://doi.org/10.1113/jphysiol.1989.sp017525 (1989).
26. Barnes, G. R. Visual-vestibular interaction in the control of head and eye movement: The role of visual feedback and predictive mechanisms. *Prog. Neurobiol.* **41**, 435–472, https://doi.org/10.1016/0301-0082(93)90026-O (1993).
27. Skavenski, A. A., Hansen, R. M., Steinman, R. M. & Winterson, B. J. Quality of retinal image stabilization during small natural and artificial body rotations in man. *Vis. Res.* **19**, 675–683 (1979).
28. Martinez-Conde, S., Macknik, S. L. & Hubel, D. H. The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* **5**, 229–240, https://doi.org/10.1038/nrn1348 (2004).
29. Angelaki, D. E. Eyes on Target: What Neurons Must do for the Vestibuloocular Reflex During Linear Motion. *J. Neurophysiol.* **92**, 20–35, https://doi.org/10.1152/jn.00047.2004 (2004).
30. Angelaki, D. E. Three-Dimensional Ocular Kinematics During Eccentric Rotations: Evidence for Functional Rather Than Mechanical Constraints. *J. Neurophysiol.*, https://doi.org/10.1152/jn.01137.2002 (2006).
31. Mustari, M. & Ono, S. Optokinetic eye movements. In *Encyclopedia of Neuroscience* (Elsevier Ltd, 2010).
32. Ackerley, R. & Barnes, G. R. The interaction of visual, vestibular and extra-retinal mechanisms in the control of head and gaze during head-free pursuit. *J. Physiol.* **589**, 1627–1642, https://doi.org/10.1113/jphysiol.2010.199471 (2011).
33. Matthis, J. S. & Fajen, B. R. Visual control of foot placement when walking over complex terrain. *J. Exp. Psychol. Hum. Percept. Perform.*, https://doi.org/10.1037/a0033101 (2014).
34. Land, M. F. & Hayhoe, M. In what ways do eye movements contribute to everyday activities? In *Vision Research*, https://doi.org/10.1016/S0042-6989(01)00102-X (2001).
35. Kassner, M., Patera, W. & Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction., https://doi.org/10.1145/2638728.2641695 (2014).
36. Rowberg, J. I2c device library.
37. Ortiz, L. E., Cabrera, V. E. & Goncalves, L. M. G. Depth Data Error Modeling of the ZED 3D Vision Sensor from Stereolabs. *ELCVIA Electron. Lett. on Comput. Vis. Image Analysis* **17**, 1–15, https://doi.org/10.5565/rev/elcvia.1084 (2018).
38. Vercher, J. L. & Gauthier, G. M. Eye-head movement coordination: vestibulo-ocular reflex suppression with head-fixed target fixation. *J. vestibular research: equilibrium & orientation* **1**, 161–70 (1991).
39. Hartley, R. & Zisserman, A. *Multiple View Geometry in Computer Vision*, 2 edn. (Cambridge University Press, New York, NY, USA, 2003).
40. Bahill, A. T., Kallman, J. S. & Lieberman, J. E. Frequency limitations of the two-point central difference differentiation algorithm. *Biol. Cybern.* **45**, 1–4, https://doi.org/10.1007/BF00387207 (1982).
41. Zuber, B. L., Semmlow, J. L. & Stark, L Frequency characteristics of the saccadic eye movement. *Biophys. J.* 1288–1298 (1968).
42. Chaparro, L. F. *Signals and Systems Using MATLAB: Second Edition* (2015).
43. Paris, S. A gentle introduction to bilateral filtering and its applications. In *ACM SIGGRAPH 2007 courses on – SIGGRAPH '07*, https://doi.org/10.1145/1281500.1281604 (2007).
44. Agtzidis, I., Startsev, M. & Dorr, M. In the pursuit of (ground) truth: A hand-labelling tool for eye movements recorded during dynamic scene viewing. *Proc. 2nd Work. on Eye Track. Vis. ETVIS* **2016**, 65–68, https://doi.org/10.1109/ETVIS.2016.7851169 (2017).
45. Cohen, J. A coefficient of agreement for nominal scales. 1288–1298 (1960).
46. Hooge, I. T., Niehorster, D. C., Nyström, M., Andersson, R. & Hessels, R. S. Is human classification by experienced untrained observers a gold standard in fixation detection? *Behav. research methods* 1–18 (2017).
47. Andersson, R., Larsson, L., Holmqvist, K., Stridh, M. & Nyström, M. One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behav. Res. Methods* **49**, 616–637, https://doi.org/10.3758/s13428-016-0738-9 (2017).
48. Hoppe, S. & Bulling, A. End-to-End Eye Movement Detection Using Convolutional Neural Networks. (2016).
49. Powers, D. M. W. The problem with kappa. *Conf. Eur. Chapter Assoc. for Comput. Linguist.* 345–355 (2012).
50. Kisler, T. & Reichel, U. D. A dialect distance metric based on string and temporal alignment. *Elektronische Sprachsignalverarbeitung* 158–165 (2013).
51. Breiman, L. Random Forests. *Mach. Learn.*, https://doi.org/10.1023/A:1010933404324 (1999).
52. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 1–9, https://doi.org/10.1109/IJCNN.2015.7280624 (2014).
53. Sudre, C. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 240–248 (Springer, 2017).
54. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *AIP Conf. Proc.* **1631**, 58–62, https://doi.org/10.1016/j.jneumeth.2005.04.009 (2014).

55. Komogortsev, O. V. & Karpov, A. Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behav. Res. Methods* **45**, 203–215, https://doi.org/10.3758/s13428-012-0234-9 (2013).
56. Santini, T., Fuhl, W., Kübler, T. & Kasneci, E. Bayesian Identification of Fixations, Saccades, and Smooth Pursuits., https://doi.org/10.1145/2857491.2857512 (2015).
57. Mann, D. L., Spratford, W. & Abernethy, B. The Head Tracks and Gaze Predicts: How the World's Best Batters Hit a Ball. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0058289 (2013).
58. Daemi, M. & Crawford, J. D. A kinematic model for 3-d head-free gaze-shifts. *Front. Comput. Neurosci.* **9**, 1–18, https://doi.org/10.3389/fncom.2015.00072 (2015).
59. Morel, P. Gramm: grammar of graphics plotting in matlab. *JOSS* **3**, 568 (2018).

Acknowledgements
We would like to thank the Google Daydream team for funding this research. We would also like to thank the Research Computing group at the Rochester Institute of Technology for satisfying a portion of our compute needs.

Author contributions
All authors contributed to drafting the paper, data interpretation, project development, and approved the final version of the manuscript for submission. R.K. created the hardware setup, data collection, and analysis. Z.Y. developed the evaluation metrics.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59251-5.

Correspondence and requests for materials should be addressed to R.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.