NASA/TM—2016–219424

Considerations for the Use of Remote Gaze Tracking to Assess Behavior in Flight Simulators

Donald J. Kalar
San Jose State University Research Foundation

Dorion Liston
San Jose State University Research Foundation

Jeffrey B. Mulligan
NASA Ames Research Center

Brent Beutter
NASA Ames Research Center

Michael Feary
NASA Ames Research Center

October 2016

NASA STI Program…in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question via the Internet to help@sti.nasa.gov

- Phone the NASA STI Help Desk at (757) 864-9658

- Write to:
NASA STI Information Desk
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

NASA/TM—2016–219424

Considerations for the Use of Remote Gaze Tracking to Assess Behavior in Flight Simulators

Donald J. Kalar
San Jose State University Research Foundation

Dorion Liston
San Jose State University Research Foundation

Jeffrey B. Mulligan
NASA Ames Research Center

Brent Beutter
NASA Ames Research Center

Michael Feary
NASA Ames Research Center

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California

October 2016

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program
STI Support Services
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

This report is also available in electronic form at http://www.sti.nasa.gov
or http://ntrs.nasa.gov/

Table of Contents

Acronyms and Definitions

2D	two dimensional
AOI	areas of interest
EEG	electroencephalography
fNIRS	functional near-infrared spectroscopy
Hz	Hertz (cycles per second)
ms	milliseconds
NASA	National Aeronautices and Space Administration
RMS	root-mean-square

Considerations for the Use of Remote Gaze Tracking to Assess Behavior in Flight Simulators

Donald J. Kalar[1], Dorion Liston[1], Jeffrey B. Mulligan[2],
Brent Beutter[2], and Michael Feary[2]

Complex user interfaces (such as those found in an aircraft cockpit) may be designed from first principles, but inevitably must be evaluated with real users. User gaze data can provide valuable information that can help to interpret other actions that change the state of the system. However, care must be taken to ensure that any conclusions drawn from gaze data are well supported. Through a combination of empirical and simulated data, we identify several considerations and potential pitfalls when measuring gaze behavior in high-fidelity simulators. We show that physical layout, behavioral differences, and noise levels can all substantially alter the quality of fit for algorithms that segment gaze measurements into individual fixations. We provide guidelines to help investigators ensure that conclusions drawn from gaze tracking data are not artifactual consequences of data quality or analysis techniques.

1. Introduction

Flight simulators are used to evaluate concepts for novel cockpit interfaces. These interfaces have become increasingly complex as information processing technology has advanced, and it is not always possible to correctly infer the mental state of a pilot on the basis of direct interactions with the system. The task of interpreting user actions can be aided by incorporating gaze tracking data; for example, if an alert is issued but does not trigger a corresponding action, gaze tracking can tell us whether the pilot noticed (fixated) the alert (and chose not to act), or did not notice it at all. When the relevant display elements are widely spaced, the performance requirements of the gaze tracking system are not severe, but it most cockpits the information is densely packed into limited space, imposing stringent limits on acceptable error in a gaze tracking system. In this paper, we examine the nature of errors occurring in practice, and discuss post-processing techniques that can be used to mitigate them.

Eye movement measurements are used across a large range of domains and applications for both basic and applied research [1,2] In the context of human factors investigations, gaze tracking data is used to aid in assessing operator attention, performance, training, efficiency, and interface effectiveness. Models that attempt to measure attention, workload, fatigue, efficiency, or other behavioral constructs generally rely on sophisticated analysis of gaze patterns, using some

[1] San Jose State University Research Foundation, San Jose, CA.
[2] NASA Ames Research Center, Moffett Field, CA.

combination of fixation locations, dwell times, and transition probabilities across the environment. These measures may be reliable in controlled laboratory contexts, but the signals recorded in part-task or more basic experimental paradigms are of much higher data quality than those captured in operational environments, due to the nature of how those measurements are taken. Applying models developed under tightly-controlled experimental environments without careful consideration of how those statistics may be biased by the operational environment will risk developing erroneous conclusions. A grounded understanding of the quality of the data is necessary to provide meaningful high-level summary statistics used to address those questions about behavior and cognition.

Measuring gaze behavior in rich and dynamic environments during the execution of complex tasks provides a unique set of challenges when compared to the more controlled and constrained environments of laboratory experiments. Traditional oculomotor research takes great care to minimize sources of noise, performing experiments in highly controlled environments that physically constrain head movement during an experiment. Simulators are built at great expense to mimic as many aspects of flight as possible (e.g., haptic feedback of physical controls, anthropometric layout of cockpit, acoustic environment, vestibular and somatosensory cues from motion, perceptual cues and optic fields out the window) because it is understood that the fidelity of the simulation environment is important for training and assessment. Gathering gaze data in a complex context is necessarily a compromise. Experimental protocols that attempt to maximize fidelity and ecological validity require that the mechanisms for observing gaze in situ be as non-invasive as possible. In order to do that effectively, instrumentation hardware must meet two goals:

1. The hardware should not be distracting or disruptive while in operation. Effective systems are easily ignored and neither draw nor demand attention during operation.
2. The hardware must not occlude or otherwise interfere with nominal task behavior, or impede or alter how the participant behaves within the environment. If performance or behavioral strategies of the participant are altered due to the presence of the gaze tracking instrumentation, then the data gathered will be selectively biased during some subset of the instrumented behavior, complicating any further interpretations.

Gaze trackers that can be used in simulator environments can either be remote or head-mounted. There are benefits and challenges unique to each of these methods for measuring gaze behavior. Remote trackers (i.e., gaze tracking systems that do not require the participant being tracked to wear specialized equipment) are increasingly popular in experimental contexts where other physiological measurement techniques are being used, such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS), due to the fact that these technologies are difficult to use in conjunction with head-mounted gaze tracking systems. Remote trackers are also advantageous compared to head-mounted systems in situations where the tracked individual makes large or rapid head movements while participating in a study, due to the fact that head-mounted trackers must be recalibrated if they are perturbed from their placement on the head.

There is a growing body of work dedicated to characterizing gaze tracker performance and understanding how to interpret the resulting data. This includes considerations as to how various classes of analysis algorithms compare to one another qualitatively [3], to assessing how they perform using behavioral data as input [4, 5]. It is well documented that a number of different design or methodological decisions can have substantive impact on the final analysis. These include factors such as model parameter choices impacting behavioral summary statistic estimates [4, 6], to hardware differences in tracker sampling rates effecting trace estimates [7].

However, the bulk of this work has been done in the context of gaze tracking in laboratory settings. In simulator environments, the nature of the data gathered are qualitatively different. For example, in [8], they discuss simulating gaze tracker data, using parameters characterizing noise that are much lower than what is currently generally observed in a high-fidelity simulator using a modern remote tracker. Indeed the precision values that [8] use as an exclusion criterion are values that would exclude most if not all gaze data captured in simulator environments using current remote gaze tracking technologies (see Figure 1). This is a fundamental challenge to remote gaze tracking in complex, dynamic contexts: even discriminations as broad as gaze on-the-road versus off-the-road while driving can be very difficult to calculate reliably [9]. Not only is there more noise present when using remote trackers in complex environments, the quality of the signal can vary substantially between subjects, due to a host of physiological and behavioral factors (e.g., posture, gestures). In aggregate, this means that investigators using remote gaze tracking systems as a dependent measure of behavior should exercise caution before applying any of the heuristics documented in the literature as best practices, or applying methods of analyzing gaze data without verifying that the assumptions of the model hold.

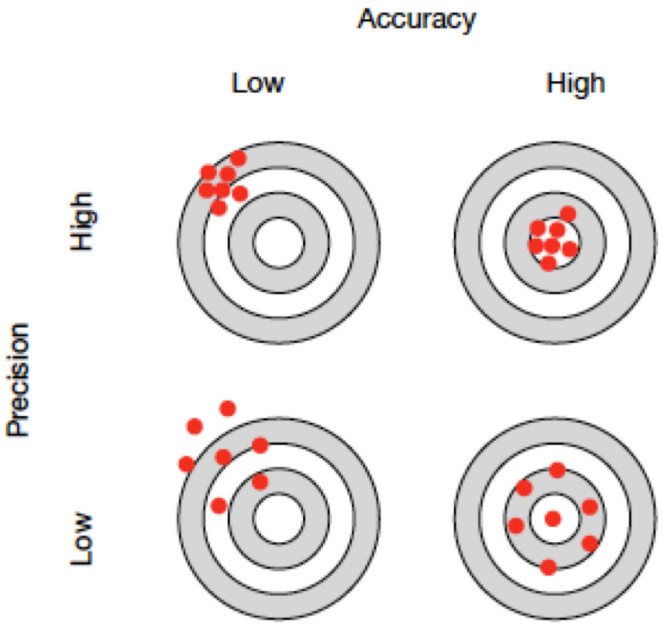

Figure 1. Accuracy and precision. Noise in estimating the point of regard when gaze tracking can take two different forms. Random noise will diffuse the signal, resulting in lowered precision. Systematic bias in the estimate will result in reduced accuracy. Lower precision leads to poorer estimation of the boundaries between fixations and eye movements (saccades). Lower accuracy leads to the potential mis-assignment of fixation locations. Noise in eye-position traces can arise from many sources, including: luminance fluctuations, noise in the camera sensor, camera vibration, and small-amplitude eye movements occurring during fixation (microsaccades, ocular drift, tremor). Accuracy degrades when the assumptions of the underlying tracking model or calibration are violated (e.g., idealized pinhole camera model is incorrect, idealized physiological model is incorrect, a camera shifts position, the geometric world model is incorrect).

In the absence of moving objects that stimulate smooth pursuit eye movements, gaze patterns typically consist of a series of fixations, during which the eye is relatively still, connected by saccades, rapid jumps from one target to another. Small saccades may be made within a region of interest, with the smallest saccades (that occur even when a person tries to hold their eye still) sometimes referred to as “microsaccades.” There is no universally agreed upon definition of exactly what constitutes a microsaccade, but Collewijn & Kowler provide a good historical overview[10]. Terminology aside, the fact remains that saccades occur in sizes ranging from tens of degrees to a few minutes of arc. For many gaze tracking systems (including most suitable for use in flight simulators), the noise level of the measurement is larger than the amplitude of the smallest saccades. Experiments should be designed such that fixation targets that need to be discriminated have separations larger than the noise amplitude.

1.1 Assessing Gaze Tracker Performance via Calibration Recordings

In laboratory research settings, it is customary to take one or more sets of calibration recordings, where the participant fixates on targets with known positions [11]. From those calibration data, researchers can estimate precision and accuracy values. Precision estimates can be used as exclusion criteria for individual trials, or entire participants, if the tracker noise is above some threshold determined in advance. Accuracy estimates can also be used as exclusion criteria, however it is often possible to fit a set of transforms to correct for systematic errors in accuracy (e.g., a global translation, scaling, or skew). These corrections are straightforward to apply when the data generated only involve a single monitor and a fixed head.

We applied a similar methodology of measuring fixations across a number of specified points in a set of studies using a high- and medium-fidelity Boeing flight deck environment, in order to more finely assess the data quality gathered by our remote gaze tracking system. In a level-D high-fidelity Boeing 747 simulator, we placed Plexiglas panels with an evenly spaced grid of fixation points on each of the glass displays in the flight deck. Crews were instructed to fixate on each of the points in turn, and labeled gaze recordings were taken. In a second study using a medium-fidelity Boeing 777 flight deck built on top of X-Plane, fixation points were rendered directly to the same computer monitors used to display the virtual flight deck. A similar protocol of capturing individual fixations at known locations was performed, with a grid of nine fixation points presented to each pilot. Because these fixation targets were generated programmatically, they were also presented in a random order. Random target positions were used to minimize anticipatory eye movements that would introduce noise.

The precision estimates from the observers are given in Tables 1 and 2. There are several important observations in these data. First, the precision values we estimate are much worse (numerically larger) than those encountered in laboratory settings. This should come as no surprise, but it demonstrates that models and measures used in highly-controlled gaze tracking studies may be compromised or incompatible with data gathered in these more challenging configurations. Second, these data show that a great proportion of the variability in precision are driven by individual differences from one observer to the next. The participants shown in Table 2 had calibration scans taken with weeks of time between measurements. Some individuals are simply easier or harder to track than others. And last, these data demonstrate the impact of how the gaze tracker is installed and configured. The 747 level-D simulator was a certified facility, and as such could not be altered in any non-reversible manner. Due to this limitation, we were more constrained with where and how the cameras and illuminators were mounted in the flight deck. By contrast, the medium-fidelity 777 simulator had no such restriction, and we were able to mount the equipment in a way that provided both a better geometry of camera locations, and a more rigid coupling between the cameras and their mounting locations, reducing mechanical noise. At least some of the differences in observed precision between these two studies are due to these constraints, though individual differences between the different pilot populations may also be a contributing factor.

These observed precision values, in all but the worst cases, do not preclude making reasonable estimates about the proportion of time participants spend looking within fairly large bounded areas in their visual field. However, many more sophisticated models of human performance and attention are built upon the notion of being able to identify individual fixations and rapid eye-movements. From these, one can calculate more sophisticated measures of dwell-time distributions, scan path probabilities, or other higher-order statistical models. The ability to extract usable signals for these kinds of models is not well explored when using remote trackers with these noise magnitudes, as most of the research has been focused on more controlled experimental environments. To better understand the scope of these issues, and to identify limitations and challenges of using more sophisticated analyses with these data, we generated synthetic gaze data and attempted to recover the underlying behavioral signals.

Table 1.

Participant	*Accuracy*	*Precision*
C1	0:90°	1:49°
C2	1:78°	2:30°
C3	2:39°	1:42°
C4	2:17°	3:16°
C5	1:34°	2:10°
C6	1:64°	2:39°
C7	3:11°	4:24°
C8	1:68°	2:96°
C9	1:34°	1:45°
C10	0:65°	0:96°
FO1	0:58°	3:33°
FO2	1:66°	4:74°
FO3	0:79°	2:86°
FO4	0:43°	2:89°
FO5	0:38°	2:45°
FO6	1:07°	3:30°
FO7	0:90°	3:92°

Table 1. Observed calibration data for ten captains and seven first officers in a Boeing 747 flight deck simulator. Pilots were measured while fixating on a nine-point grid superimposed on each pilot's primary display. Multiple calibration recordings were taken during a full day of flight simulation, and median values are reported. All values are reported in degrees visual angle, and precision values are one standard deviation from the mean.

Table 2.

Participant	*Session 1 Precision*	*Session 2 Precision*
1	1:14°	0:60°
2	1:72°	1:79°
3	1:01°	1:59°
4	4:41°	3:37°
5	2:07°	2:30°
6	1:56°	2:77°

Table 2. Average precision for six participants across multiple sessions, calculated as one standard deviation of the observed dispersion during each fixation, in degrees visual angle. These sessions were separated by one or more weeks.

2. Methods

2.1 Simulating Gaze Data

Assessing the performance of a gaze tracker or an algorithm for labeling a gaze trace is a fundamentally difficult problem. This is due to the fact that an objective ground-truth describing the observer's behavior is generally unknown. While subjects can be instructed to move their eyes in a well-defined manner (e.g., maintaining fixation on one or more targets in a controlled experimental setting), their compliance will not be perfect. From the prospective of an outside observer, it is unclear whether deviations from prescribed behavior in the data are due to noise in the measurement or noise from the observer. There has been work done in assessing analysis algorithm performance using very simple behavioral inputs, but these approaches require sophisticated models of eye-movements and latencies, and must still make assumptions about observer task compliance [4, 5].

We developed a simple method for generating synthetic gaze traces in order to assess how different methods of analyzing gaze data behave based on the nature of the input signal. By using algorithmically generated traces, we can remove any uncertainty about the underlying truth of the input signal. The goal is to generate traces that have properties similar to what would be generated from a remote gaze tracker in a high-fidelity environment.

In a simulator environment the instrumented operator (driver, pilot, astronaut, etc.) will, as a consequence of the physical environment and their immediate task, repeatedly sample the same physical regions as part of their normal scan pattern. These consecutive fixations may take place in relatively close proximity to one another, as is the case when looking to adjacent controls or displays within a cockpit, or these fixations may be separated by a much larger physical distance, as in the case of alternating between fixating on a cockpit control and out the window [12]. Based on the nature of cockpit designs and the nature of piloting or driving tasks, we expect gaze tracking data measuring behavior within a simulator to have several stereotyped features. These simulated gaze traces attempt to capture some of these expected regularities.

In order to approximate the geometric regularities present within rated by small or large distances. As illustrated in Figure 2, simulated gaze data are generated by randomly fixating and saccading among predefined Areas of Interest (AOIs). The simulated environment takes four parameters: the number of local AOIs within a region, the number of regions, the distance between adjacent AOIs within a region, and the distance between different regions. The simulation begins by fixating at one of the target AOIs for some random duration (sampled uniformly between 100ms and 600ms). At the end of that duration, a Markov process selects between continuing to fixate (resulting in a longer duration fixation), transitioning to an adjacent AOI (small eye movement), or transitioning to an adjacent region (large eye movement). The probability of making a large eye-movement was manipulated experimentally, while the probability of continuing to fixate on a given target versus moving to a new local target was fixed at 50%. Using Figure 2 as a reference, small simulated eye movements between AOIs would mean transitioning to one of the adjacent targets in a different color, while large eye movements would mean transitioning to a point of the same color in an adjacent region. This is of course not required, but it helps to constrain the signal-to-noise ratios present in these simulated traces and makes comparing the performance of analyses across the parameter space more straightforward.

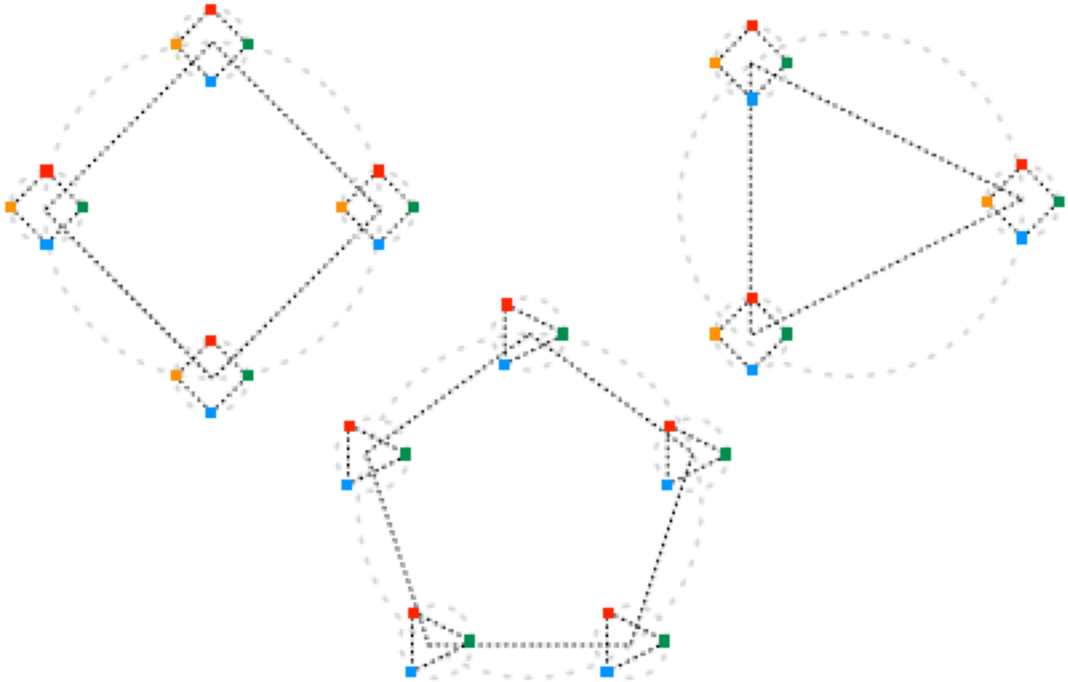

Figure 2. Algorithmically generated fixation locations. Fixation locations are specified by recursively sampling on circles of appropriate radius. This approach generates a mixture of small and large saccades of defined amplitudes.

This set of specified AOI locations and transitions provides most of the simulated signal, but in order to more closely mimic the physiological constraints of eye movements, the saccades between consecutive fixations must also be specified. Saccades are fit between consecutive fixations using estimated main sequence parameters [13]. In doing this, the initial simulated trace has a velocity signature consistent with what is expected by velocity-based methods for saccade detection algorithms. This is a simplified model of actual eye movement behavior, as it does not include any noise or corrective saccades for eye movements (e.g., saccades are never hypo- or hypermetric in these traces).

The simulated trace is sampled at a specified frequency, yielding a time series of *x* and *y* gaze positions. Because the duration of each fixation is specified to the nearest millisecond, saccades are sampled at unique phases along the velocity profile. This means that, much like when measuring actual behavior, equivalent energy saccades may have more or fewer samples during the eye movement, based on when the simulated movement began with respect to the sample rate and phase of the tracker.

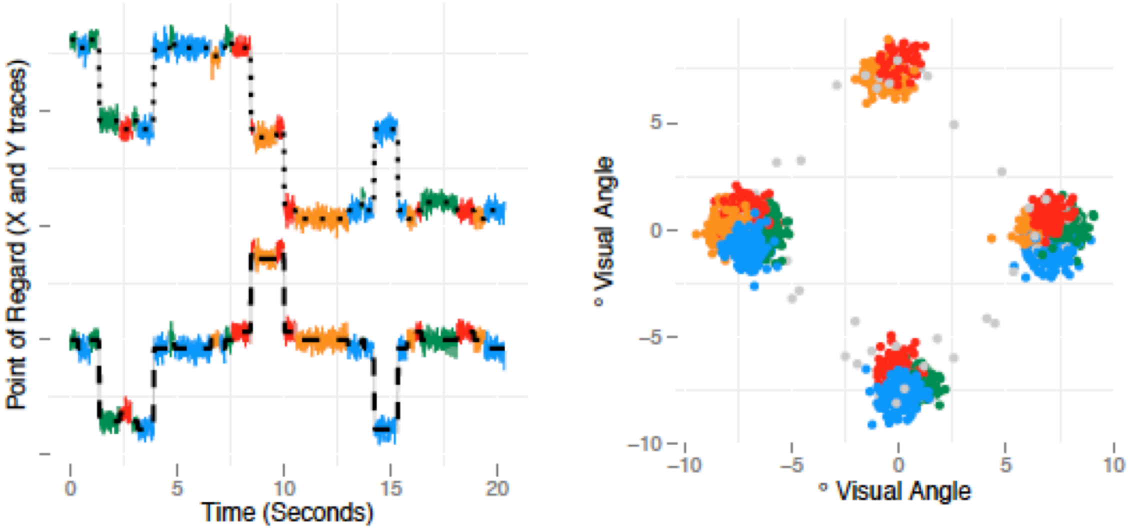

Figure 3. An example synthetic, randomly-generated gaze trace. Left: the time-series for both the x (dotted, above) and y (dashed, below) coordinates of the gaze vector projected onto a plane. Right: the same synthetic trace, projected spatially rather than temporally. The geometry defining valid fixation locations for this example is four local and four global locations, as illustrated in Figure 2.

After the trace is generated, independent Gaussian noise is added to both the x and y positions of the gaze position at every sample. A synthetic gaze trace generated using this procedure is shown in Figure 3. The left-hand portion of the figure shows time courses of the x and y signals, while the right-hand portion shows the samples plotted in visual space. The different colors correspond to the four “local” targets, as depicted in the upper left portion of Figure 2.

The notion that noise within an actual observed trace is purely Gaussian in form is a strongly simplifying assumption. Outside of the fact that this does not speak to issues of accuracy at all (see Figure 1), inspecting actual measured traces from simulator environments demonstrates that the noise present may be far more complex. Some of these complexities and recommendations for addressing real-world noise are discussed in Section 4.

2.2 Algorithms for Fitting and Segmenting

The nature of the eye movement signal leads to two complementary approaches: in the first, we attempt to identify the saccades by looking for spikes in the velocity signal, while in the second we attempt to identify fixations by looking for contiguous samples having similar values. We also describe a novel approach based on independent classification of each sample.

The starting point for analyses of instrument scanning and high-level cognitive processes is the separation of a continuous eye-position trace into fixations, interspersed with rapid saccadic eye movements that reposition the image of the target object onto the fovea. Although the position and orientation of the head and eyes given by commercial gaze-tracking systems could be used for segmentation of the time record into head movements, eye movements, and combined head-and-eye movements using a physiologically-realistic model (e.g., [14]), the point-of-regard traces (i.e., gaze) that give the 3d intersection of the gaze vector with a target object (usually a display) provide a useful starting point for segmentation of the gaze vector into fixations.

There are many different methods for segmenting gaze data into discrete behaviors (for a comparative review, see [3]). We developed three different analysis approaches and applied them to our simulated data. Because the input signals are generated algorithmically, we can compare the results of various segmentation algorithms to the noise-free simulated trace in order to evaluate various aspects of their performance.

2.2.1 Velocity-based Saccade Detection

As a practical matter, displacements in the point of regard (on planar displays) have quasi-saccadic velocity profiles because the relatively slow head component of combined eye-head gaze shifts occurs for movements larger than 20 degrees [15], which also have a substantial saccadic component. Thus, one approach to segmentation of point-of-regard traces into saccades and fixations utilizes traditional saccade-detection algorithms (e.g., [16]).

The current saccade-detection algorithm has three main components which have been described in detail previously [17]. First, a likelihood metric [18] for saccade occurrence is generated by convolving the eye-velocity trace with a saccade-shaped velocity template. Second, a threshold is applied to the likelihood metric to generate flagged regions indicating high likelihood of saccade occurrence within the trace. Last, a non-linear clustering stage prevents false alarms by ensuring that the detected saccades have a minimum duration. While use of the velocity trace offers precise temporal detection of saccade onset (e.g., Figure 7, top), the primary challenge in applying this process to gaze data concerns the relatively high level of positional noise in gaze data (e.g., Figure 9, blue high-noise simulations), and subsequent problems preventing false alarms.

The saccade-likelihood metric is given by the cross-correlation of the eye-velocity trace with a saccade-shaped velocity profile [13], given by:

$$v(t) = \frac{35a}{16\tau_a}\left(1 - \frac{4t^2}{\tau_a^2}\right), \qquad -\frac{\tau_a}{2} \leq t \leq \frac{\tau_a}{2}$$

where a is the saccade amplitude in degrees, and τ_a is the duration in milliseconds of a saccade of amplitude a. The relation between duration and amplitude is given by the saccadic main sequence [19], fit from data given in [15]:

$$\tau_a = 21a^{0.54}$$

Because of this relationship between saccade duration and amplitude, it is impossible to have a single template that is optimal for detecting all saccades within a distribution of amplitudes; instead we choose an intermediate value that provides a good compromise for the amplitudes of interest. For 240 Hz head-fixed 2D eye tracking (e.g., [20]), we assume a minimum detectable saccade size of 1 degree, with a duration of 28 ms, corresponding to 5 samples. Several factors in the real data will inform setting the amplitude of the saccade template for gaze data, including: sampling rate, noise level in the eye position trace, and the distribution of likely saccade sizes in the gaze data. While it may seem desirable to tune the saccade velocity profile to detect small-amplitude saccades (e.g., less than one degree), the sampling rates of commercial gaze trackers (60, 120 Hz) allows for only a small number (1–3) of samples even for a 21 ms one-degree saccade, leading to poor detection performance for saccades of duration equal to the search template [17] and compromised power for detection of larger saccades of longer duration. For the simulations given in this report, we used a one-degree saccade, a saccade threshold of 0.25 degree, and a minimum saccade duration of one sample.

2.2.2 Position-based Time-series Analysis

The velocity-based approach described in the previous section tends to perform poorly when the noise level is high, because high temporal frequencies in the noise are amplified by the differentiation performed to compute velocity; highly-localized noise events can introduce spurious saccades. Thus, in high-noise conditions, a position-based approach often performs better. An example of this is the "dispersion" algorithm [3], in which fixations are "grown" incrementally by adding adjacent samples whose dispersion from the current mean is below a pre-set threshold. Use of a pre-set threshold, however, can cause problems when there is variable, subject-dependent noise (as seen in Table 1). Here we propose an alternative, based on recursive splitting. Instead of a fixed dispersion threshold, fixations are segmented or merged based on statistical significance. A single parameter, the p-value for statistical significance, controls the sensitivity of the process. A low p-value produces very few false alarms, at the expense of possibly missing a few true events; increasing the p-value insures that all true events will be detected, at the expense of a few false alarms.

The basic idea is that, for any segment of data, we search for the best "split point." For any given candidate split point, we approximate the signal by two constant segments, whose values are assigned from the means of the input data on either side of the split point. The residual errors are computed, and the root-mean-square (RMS) error is computed from the residuals. Exhaustive search is performed to find the split point with the lowest RMS error.

Figure 4 shows a fragment of synthetic data with two fixations and a single saccade. (In this example, a single dimension is used, but the extension to two dimensions is straightforward.) The heavy line indicates ground truth (a piecewise-constant signal), while the thin line has been corrupted by Gaussian-distributed white noise. Above the trace, we show a plot of the RMS error for each trial split point, showing a well-defined minimum at the true transition time.

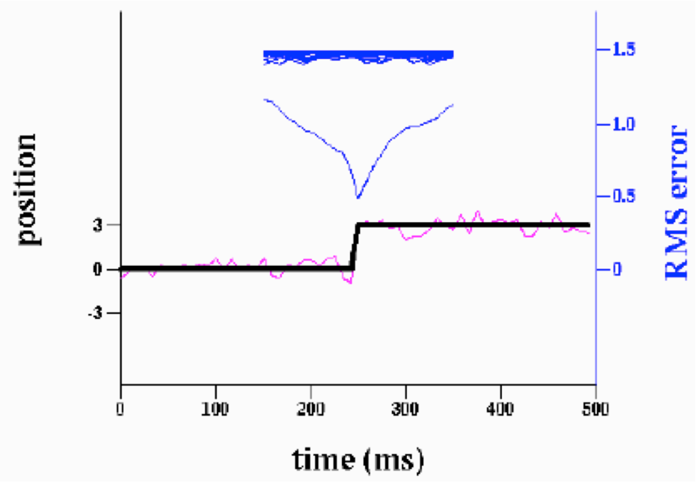

Figure 4. Synthetic data consisting of two fixations linked by a saccade. Ground truth is shown by the heavy black trace, while the thin magenta trace shows the signal corrupted by noise. The RMS error of the fit as a function of the start time of the second fixation is shown above the traces (scale expanded for clarity). Splits that would result in a segment shorter than a minimum fixation duration parameter are not considered. The V-shaped error trace corresponds to that obtained with the original data, while the additional traces above correspond to permutations of the input.

By this procedure we can identify the best point at which to split the input signal. But should we accept the split? Any set of noisy data will always have a best split, but the difference between the two levels may be small compared to the noise level. We accept or reject the split using the statistic for a simple t-test. We compute a the pooled standard deviation as follows:

$$\sigma_{\text{p}} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the lengths of the sub-intervals before and after thesplit ($n_1 + n_2 = \text{N}$), and σ_1^2 and σ_2^2 are the sample variances associated with each sub-interval:

$$\sigma_1^2 = \sum_{i=1}^{n_1} \frac{(\mu_1 - x_i)^2}{n_1}$$

$$\sigma_2^2 = \sum_{\text{i=n1+1}}^{\text{N}} \frac{(\mu_2 - x_i)^2}{n_2}$$

where μ_1 and μ_2 are the sample means associated with the two intervals. Using this quantity, we compute the *t* statistic:

$$t = \frac{v_1 - v_2}{\sigma_p\sqrt{{}^1/_{n_1} + {}^1/_{n_2}}}$$

In two dimensions, we compute independent t statistics for the horizontal and vertical components, and use the one with the larger magnitude. This could be made isotropic by performing a coordinate rotation so that the line joining the two fixation centers is aligned with the coordinate axes.

This t value can be checked for significance using the standard t distribution with $n_1 + n_2 - 2$ degrees of freedom. However, this results in approximately 10 times the number of expected false alarms, when tested with a constant signal corrupted by noise. The t distribution is appropriate only when the split point is fixed and determined ahead of time. In our method, on the other hand, the split point is chosen to produce the best fit. While it may be possible to derive an analytical form of the resulting distribution, we test for significance using a permutation test. After computing the value of the test statistic for the original data, the data are permuted, and the procedure is applied to the scrambled data. The original t value is compared to the distribution of t values resulting from the permuted samples. For example, we might accept at the p=0.05 level of significance if the t statistic computed from the split of the original data is higher than that from each of 19 permutations. In practice, we compute 199 permutations, accepting at the 0.05 level if the test statistic is among the top 10 values, and the 0.01 level if among the top 2.

After a split is accepted, we apply the same procedure to each of the two shorter intervals. Figure 5 shows the progress for a simple example containing 5 fixations. The dendrogram plotted above the traces indicates the order in which the splits are performed: the initial split is placed at the 3rd saccade, and the resulting two sub-intervals are then subjected to the same process, and so on. For any segment, the process terminates when either the best split does not pass the t-test, or the length of the interval is below a parameter representing the shortest reasonable fixation (we use a value of 150 milliseconds).

After this procedure has terminated, we make an additional validation pass. Because of the recursive subdivisions, not all adjacent fixations have been tested with the t-test, only those split on the last pass before validation. We therefore validate the final configuration by merging and re-splitting each pair of adjacent fixations. If a significant split is accepted, then the topology of the dendrogram does not change, although the position of the split may be adjusted. In some cases, however, a pair will fail the significance test, and will be merged, changing the structure of the dendrogram. Occasionally, the validation pass results in an unstable repeating cycle. We therefore terminate the process after 8 validation passes. In rare conditions, a significant split will not be found in spite of the fact that there may be (multiple) obvious saccades. Because of this, when no split is performed but the interval length is longer than 3 times the minimum fixation duration, we introduce a split at the sample corresponding to the maximum magnitude of the velocity, and defer statistical testing to the validation pass.

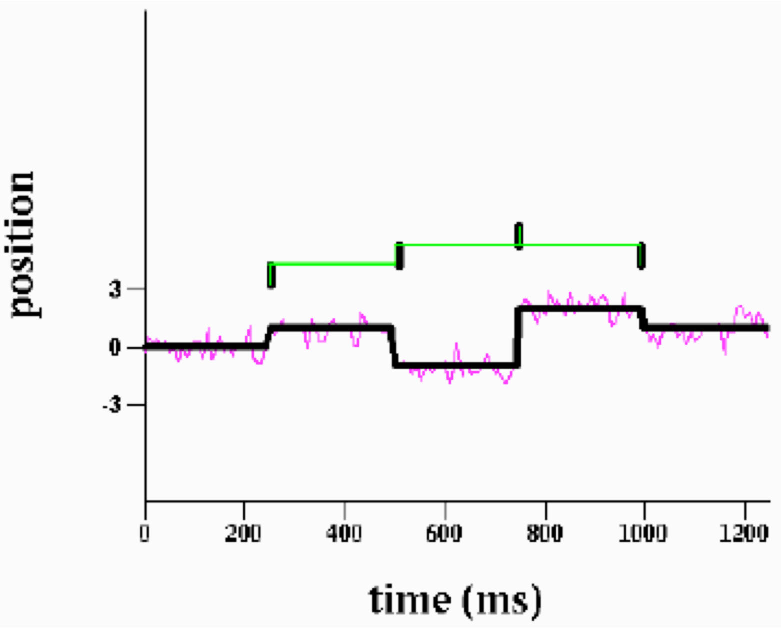

Figure 5. Synthetic data containing five fixations. The dendrogram above the traces indicates the order in which the splits were performed, with the first split (the root of the tree) occurring at the third saccade.

2.2.3 Machine Learning Based Classification of Gaze Data

While velocity- and position-based methods can both segment the data into discrete fixations, there are often problems with metrical accuracy that are not solved by simple calibration mapping functions. Traditionally, one defines AOIs in the environment, and labels observed samples based on the intersection of the estimated gaze vector with these regions [3], requiring precise geometric calibration. An alternative approach that obviates the need for geometric calibration is to sample known gazes to each of the AOIs, and use these samples to build a classifier.

AOIs are specified as planes within the coordinate system of the gaze tracker, and are defined by physical features in the environment. The locations of these AOIs may be static throughout the course of the measurement (e.g., an automobile speedometer or a cockpit primary flight display), or may be dynamic and transient (e.g., some object of interest visible out the window).

Whether static or dynamic in nature, by labeling each gaze sample with a corresponding AOI it becomes straight-forward to calculate descriptive statistics such as total dwell time across the set of AOIs. (In operational environments, it is not uncommon to encounter sections of missing data due to tracker failure, or gazes to unlabeled areas; it may be useful to assign special labels for these events, which did not occur in our simulations.) Fixations can be approximated from AOI-based analyses if it assumed that multiple consecutive fixations do not occur within an AOI, or that collapsing consecutive within-AOI fixations is an acceptable compression of the data and simplification in representing the gaze behavior. Provided either of those assumptions hold, consecutive samples labeled as intersecting the same AOI can be grouped to define a fixation. Algorithms that calculate fixations in this manner may be made more sophisticated by enforcing simple physiological constraints, such as a minimum fixation duration.

In the context of a well-controlled experimental test environment, it is straightforward to use veridical AOIs defined within the environment as the basis for analysis. However, when transitioning from a laboratory test facility to a high-fidelity simulator environment, bias in the measured signals can significantly impact how the data are labeled and subsequently interpreted.

It is common practice in laboratory studies to take calibration measurements, where the participant fixates on known locations on a display. From these measurements, transformations can be calculated to adjust for any systematic bias (i.e., errors in accuracy) present in the subsequent gaze data. This becomes more challenging in the context of a multi-sensor gaze tracker in a rich simulator environment, because accuracy errors can be considerably more complex than what is encountered in laboratory settings.

One straightforward solution to this issue is to sample the observer fixating on each AOI in advance of gathering any behavioral data, and use those observed signals as the basis for subsequent gaze labeling. These calibration samples will estimate whatever geometric distortions are present in the signal coming from the tracker, and can provide a basis for estimating the discriminability of fixations on different AOIs from one another for each unique observer.

To implement this data-driven approach to defining AOIs, we use a k-nearest neighbor machine-learning classifier [21]. The k-nearest neighbor classifier takes a set of labeled data (in this case, the labeled calibration samples where the observer fixated on each AOI in the environment) and uses it to classify new observations (e.g., gaze behavior during a task). The classifier takes a sample, and determines the *k* closest values to that sample, based on Euclidean distance. The unknown observation is then labeled based on the mode of the distribution of labels from those *k* closest samples.

Because each sample is classified independently, there is opportunity for high-frequency noise in the classification that is not physiologically possible. To address this, a simple symmetric filter (in this case, corresponding to a temporal window of 150 ms) is convolved with the labeled trace, calculating the mode label within that window for each sample t. The sample at time t is then assigned the local mode value.

This classifier has a number of desirable properties. It is often the case that the performance of the classifier is robust to choices for the value of k. It is also free from any distributional assumptions, so there is no limitation if the noise is non-isotropic, or non-Gaussian. It is also simple to implement and has acceptable computational performance for large data sets. The original motivation for developing this approach was in response to the anomalies observed in calibration scans taken in simulator environments, where we observed distortions more complicated than what one is likely to encounter in controlled laboratory experiments. Noise and uncertainty from the gaze tracker are then implicitly captured by using those data as the basis to train the classifier.

This approach is only appropriate when a number of constraints are satisfied:

1. The AOIs need to be static within the environment for the duration of the trace. Dynamic AOIs are not well-modeled by this approach (though depending on the specific nature of the displays, one might be able to sample multiple static AOIs in the regions where the Dynamic AOIs appear).

2. No distractor elements in the environment can go untrained. The classifier can only classify based on the best fit of whatever signals were used in training. If there are potential elements in the environment that the observer can fixate on that are unmodeled in the classifier, then those fixations will be mislabeled, leading to false alarms.
3. The AOIs should be small enough and spaced far enough apart that gazes anywhere inside them will be well approximated by the calibration sample.
4. This method assumes that the signals coming from the gaze tracker are stable over time, and that the classifier trained at the beginning of an experimental condition will still adequately describe the data by the end of the observation. This is something all analysis methods assume, but a second set of calibration measurements can be taken at a later time to confirm the test-retest reliability of the gaze tracker output.

2.3 Algorithm Performance Metrics

Several methods for analyzing gaze behavior (e.g., [3]) and assessing the quality of said analyses have been proposed in the literature [4, 5].

Depending on the research question being addressed by gaze data, different metrics will be appropriate. When assessing a new interface or training protocol, investigators using gaze data will calculate some summary statistics within each level of the manipulation, and then compare those statistics for differences. These statistics can describe the behavior ranging from a multinomial distribution of aggregated dwell times during a task (binning each sample independently, effectively ignoring fixations and saccades), to very specific models describing the spatial distributions of fixations, or the transition probabilities that best describe observer scan patterns [11].

As a first-order approximation, we adopt the summary statistics described by [4] to grade algorithm performance. Namely, the number of detected fixations during a trace of some specified duration. In [4], they also discuss summary statistics capturing the average number of saccades, the average saccade amplitude, and the average fixation duration. However, because distinct fixations are always separated a saccades (even if only implicitly), these different measures are effectively equivalent. The results of these analyses are shown in Figures 8 and 9. We also calculate two more general summary measures, one based on the quality of the signal fit with respect to localization, and the other with respect to segmentation.

2.3.1 Distance

The most straightforward question to ask when assessing model fit is to determine how close the modeled gaze position is to the original signal. To capture that, we calculated the root-mean-square error of each model fit with the source trace, as given below:

$$E_{RMS} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{(\hat{X}_t - X_t)^2 + (\hat{Y}_t - Y_t)^2}$$

where $(\hat{X}_t, \hat{Y}_t)$ is the estimated gaze location, (X_t, Y_t) is the ground truth, and T is the number of samples. This finds the average difference in true position versus estimated position calculated for every sample t across the complete trace. This statistic captures the general performance of each algorithm with respect to how accurately the gaze was localized within the simulated environment.

2.3.2 Durations

Outside of errors in localization as described above, algorithms that partition gaze traces into components (fixations, saccades, blinks, etc.) will also generate errors with respect to how the trace is segmented into unique behaviors. These errors will include omission (failing to detect an event), inclusion (false alarming and adding an erroneous event), duration (increasing or decreasing the estimate of how long an event took place), or displacement (translating an event in time).

It is important to capture these kinds of performance errors, because many measures of eye-movement behavior that purport to capture higher-order cognitive processing (e.g., [11]), rely on using measures such as average dwell-time or average fixation duration, average saccade amplitude, geometric regularity in fixation distributions, or changes in scan patterns as estimated by transition probability matrices.

In order to assess algorithm performance when it comes to correctly segmenting these simulated traces, we calculated a mean Jaccard index between each algorithm's output and the corresponding ground truth [22]. The Jaccard index is a set-theoretic measure of similarity, defined as the ratio of the magnitude (count) of the intersection of two sets to the magnitude of their union:

$$J(A,B) = \frac{|\, A \cap B \,|}{|\, A \cup B \,|}$$

The Jaccard index is bounded between zero and one, and can be thought of as a measure of the mutual information that exists between the two sets.

Here we assume that the ground truth, *g(t)*, is a set of labels that uniquely represent each event, with a similar representation for our estimated fit, *f(t).* We define the set G_t to be the set of time points for which g has the same value as *g(t):*

$$G_t = \{s | g(s) = g(t)\}$$

With an analogous definition for F_t, we can then define the mean Jaccard index $\bar{J}$ as the mean of the Jaccard index at each time-point t:

$$\bar{J}(g,f) = \frac{1}{T}\sum_{t=1}^{T} J(G_t, F_t)$$

High Jaccard indices suggest that the boundaries between events are well captured by segmentation algorithm. Low values suggest a poorer correspondence. Low Jaccard index values may be from any of the errors enumerated above (omission, inclusion, duration, or displacement).

3. Results

The algorithmic approach to simulating gaze traces described in Section 2.1 allows for a huge search space of combinations of scan patterns, local and global geometry, and noise. In this simulation, we compared the performance of the three models described in Section 2.2 under two different noise levels (0.5 and 3 degrees standard deviation), two different scan patterns (10% versus 90% probability of scanning locally), and two different geometric configurations (2 or 6 local AOIs, all with 4 global regions).

These parameters were chosen to map onto the kinds of differences we would expect to observe in experimental contexts. The two noise levels represent the data quality extremes that are likely to be encountered in simulator environments, based on prior real-world observations (with the noisier conditions being more common). The two probability values map onto explore versus exploit behavioral scan patterns, which are likely to occur as a function of task. Lastly, the two AOI configurations correspond to the differences between lower and higher density displays.

The performance of each analysis is plotted below, using several different metrics for assessing model performance. Because these data are simulated, each of the model fits can be scored relative to ground-truth. Each model was given the same set of traces to fit. There were 10 traces per condition, and each trace consisted of 20 seconds worth of simulated gaze behavior, sampled at 120Hz.

3.1 Distance

The simulation results comparing the true position of the noise-free simulated gaze trace compared to the model estimate of the gaze position are given in Figure 6, as the average RMSE for the duration of the simulated gaze trace. These results are unsurprising, and demonstrate that the proximity of the estimated position to the true value will be dominated by the magnitude of noise present in the signal. There are differences in fit quality based on scan pattern input (Figure 6, top versus bottom), due to the fact that the condition where most eye-movements are local lead to smaller penalties for mis-estimation.

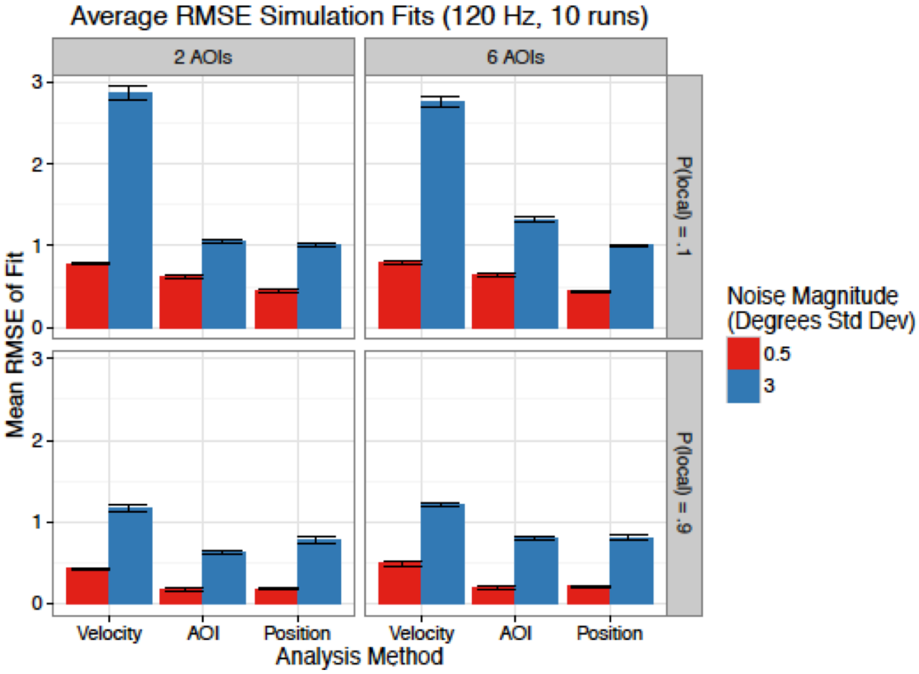

Figure 6. Average RMS (lower values indicate better trace fits). The amount of input noise has a large impact on the deviation of the modeled fit. This is a function of both noise energy and signal energy, where scanning patterns that span large spaces (above) also yield poorer average fits than those behavioral traces that tend towards smaller eye movements (below). Error bars indicate ± 1 SE.

3.2 Durations

The correspondence between each model's partitioning of the simulated trace into independent fixations and the true values is given in Figure 7. Larger values (closer to one) indicate a higher correspondence between the partitioning between gaze events in the original trace and the partitioning fit by each algorithm. In low-noise regimes, the velocity-based methods dominate. This is not surprising, as the velocity model explicitly extracts saccades, and uses those eye-movements as basis for segmentation. The AOI-classifier approach fares the worst, due to the fact that it makes no attempt to identify eye-movements as independent from fixations. The penalty for this limitation is especially pronounced when the behavior is biased towards moving between regions (Figure 7, top), because many of those saccades are large, and are represented in the original trace for several samples in duration.

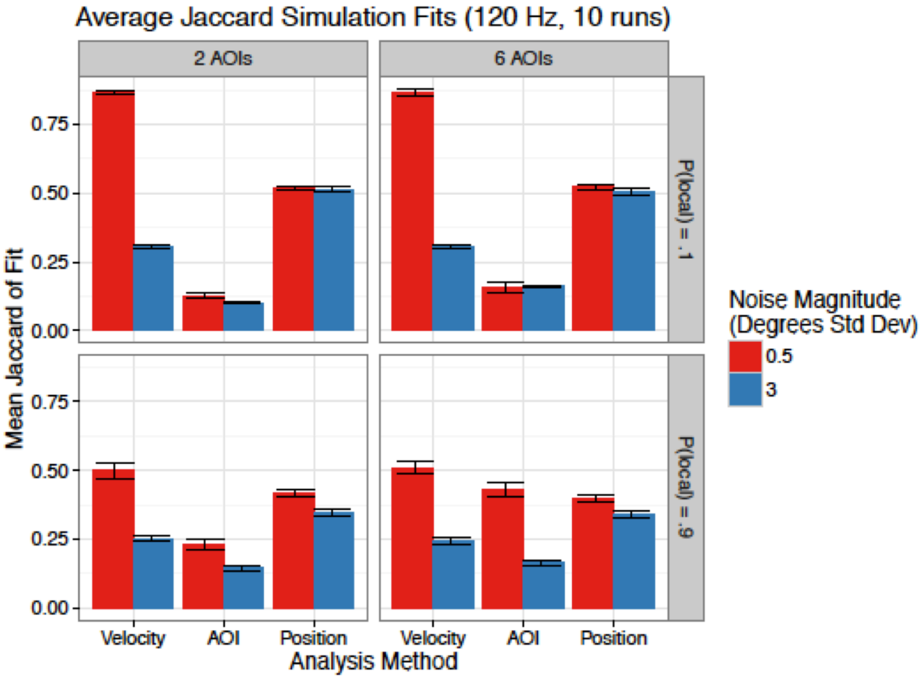

Figure 7. Average Jaccard measure (values close to one indicate more accurate trace partitioning). Behavioral traces with many large amplitude eye movements are most accurately segmented by both the velocity and position-based methods of analysis (above). The classifier AOI approach performs comparatively worse, due to not explicitly modeling eye-movements as unique from fixations.

3.3 Other Summary Metrics

Plotted in Figures 8 and 9 are average log ratios of the model fits compared with the ground truth from the original simulated eye trace. Values below zero represent a systematic under-estimate (misses), while those greater than zero are biased to over-estimate the true parameter (false alarms), as calculated from the initial simulated gaze trace. The specific summary statistic plotted include the number of observed unique fixations in a trace (Figure 9), and the average fixation duration, or dwell time (Figure 8).

These figures plot log-ratios of the true trace parameters compared to the estimated statistics calculated from each algorithm’s fit. All of these values, which reliably deviate from zero, indicate that these methods for segmenting noisy gaze traces are all subject to estimator bias. In the case of estimating the average fixation duration (shown in Figure 8), the trend is generally a negative log-ratio. These negative values indicate a systematic bias towards underestimating the true average fixation duration for these simulated gaze traces. This is due to the fact that all of the algorithms are affected by the noise in the trace (especially as the noise increases, as seen with the velocity and the AOI-classifier methods), and are prone to false alarming and segmenting fixations erroneously. In the low-noise condition, the AOI-classifier algorithm actually tends to over-estimate the fixation duration. This is due to the fact that the AOI-classifier model does not explicitly partition saccade samples from fixations, and in the low noise condition ends up appending those saccade samples to the ends of the true fixation signals. The position-based method also does not explicitly label saccade samples, but does not show this effect, presumably because of erroneous introduction of extra splits in the neighborhood of large saccades.

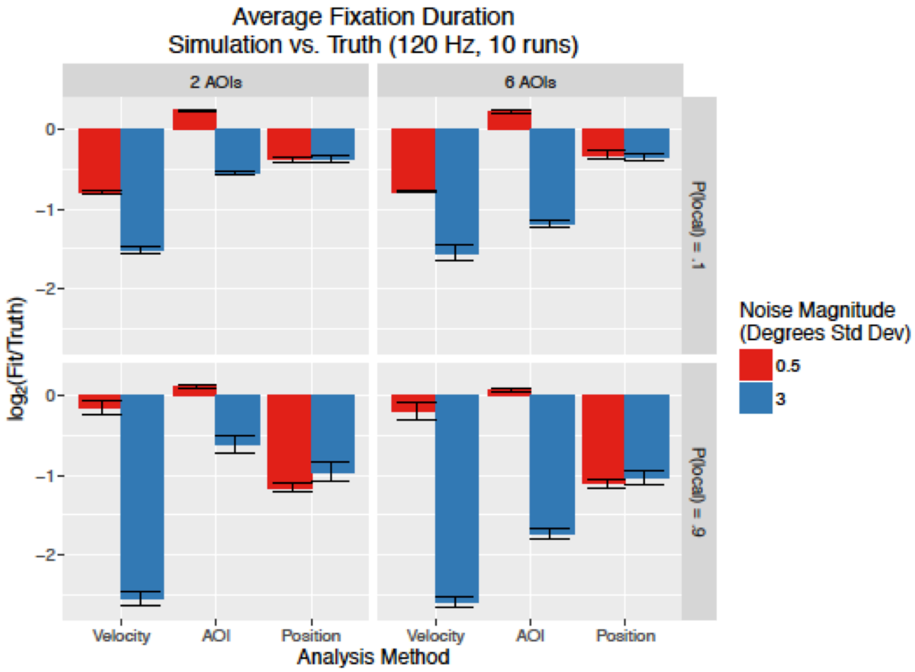

Figure 8. Simulation fit results. Plotted are the log-ratio of the average duration of estimated fixations to true average fixation duration for each simulated trace. These negative values indicate a systematic underestimate of the true average durations.

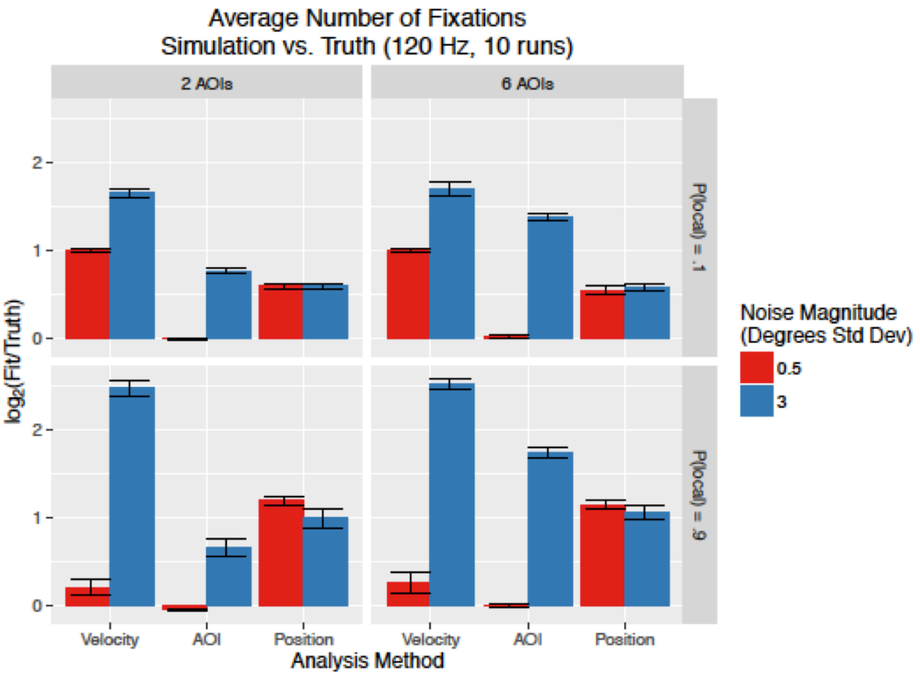

Figure 9. Number of fixations. Plotted are the log-ratio of the number of estimated fixations to true number of fixations for each simulated trace. These positive values indicate a systematic bias towards segmenting the gaze trace into more fixations than existed in the simulated trace.

4. Discussion

There are several important considerations when designing experiments that use gaze behavior as a dependent measure illustrated by the results of this simulation:

1. Some methods of analysis are highly impacted by noise, while others are more robust. These simulations suggest that the position-based method for analysis performs consistently across a range of empirically observed noise levels. Further systematic investigation is needed, but this is desirable given how heterogeneous noise levels can be from one observer to the next.

2. Model performance can be affected by the geometry (number of AOIs) of the simulation when using an AOI-based approach to partition the data. As the local complexity increases, the opportunity for mis-localizing gaze location also increases. This is highly sensitive to the distance between AOIs compared to the noise level in the trace. Any comparisons between alternative physical layouts should be cautious to ensure measured differences (or lack of differences are not driven by artifacts of model sensitivity.

3. Model performance can be affected by the scanning behavior (task) of the subject. Tasks or behaviors that encourage exploring the environment will have more signal due to large amplitude saccades when compared to behaviors that focus on local clusters of information. Care should be taken when comparing different tasks, either between or within a given subject, because these scan path differences can have an impact on the quality of the model fit.

4. These models tend to miss signals in the presence of noise, systematically underestimating how often the eye moves from fixation to fixation. Any theory that uses these descriptive statistics must take these biases into account. Table-top flight simulations or part tasks may identify more sophisticated measures of attention or workload (e.g., [23]), but these results cannot be directly compared to data gathered in high-fidelity environments without accounting for the differences in data quality.

As the simulations above demonstrated, small amplitude saccades and small interval fixations are difficult to detect in behavioral traces with the noise profiles commonly observed in complex environments, such as high-fidelity simulators. Research hypotheses that are based on fixation-based statistics as dependent measures (number of fixations, dwell time, saccade amplitudes, etc.) must take great care to ensure that the effect of interest is detectable given the task behavior, environment geometry, and tracker noise level. This is especially true whenever the null-hypothesis is used as evidence supporting a new technology or intervention technique, by comparing measured gaze behavior to a known baseline. Simulations and power analyses are necessary to ensure that the empirical data are sensitive enough to detect real differences. When possible, studies should use within-subject designs so that the noise levels are matched when making comparisons (taking special care to control for order effects). If within subject designs are not feasible, noise levels of participants should be estimated in advance to attempt a matched- subject design.

Commercial products make the process of gathering gaze data very straightforward, but practitioners must be diligent to ensure that the data are interpreted correctly. This is especially true as the models for interpreting gaze behavior becomes more sophisticated.

5. Recommendations

Based on these findings, we created a list of recommendations that may help to improve the accuracy and precision of the data.

5.1 Calibration

First is the need to collect a series of calibration recordings. This can be accomplished by having subjects fixate a set of known locations in the simulator prior to collecting gaze data, and after the simulation if there is any possibility of projected gaze position shifting during a simulation (which is likely). This will allow you to quantify the accuracy and precision of the eye-position data, which is crucial to all further analysis.

A well-constrained calibration routine provides the best possible conditions for collecting gaze data. First, the subject will usually refrain from making dramatic head movements, even without explicit instruction. Second, the subject can be instructed to hold steady gaze. Third, collection of calibration data at each point can be triggered by the subject, ensuring a minimum of blinks, saccades, and errant fixations. Even with the best calibration protocol, these problems will occur. Note that even if gaze calibration accuracy remains stable before and after collecting data, the accuracy may still have changed during the simulation. AOI sizes must reflect the level of noise in the gaze signal. An AOI diameter of 5–10 times the standard deviation of the eye position signal while fixating is likely discriminable. An AOI diameter that nearly equals the standard deviation of eye position while fixating is not, assuming that the accuracy of the eye-position signal is constant. Keep in mind that the accuracy of the gaze position signal can change between the calibration and during a simulation (e.g., the projected gaze vector may be head-position dependent).

5.2 Cautions for Analysis

Use caution when applying analysis techniques. Conclusions that one can draw are constrained by the level of accuracy and precision in the signals being collected. Constructing a simulated data set with the level of accuracy, gaze position noise, across-observer changes in eye position noise, and other observed artifacts, then running this data set through the analysis code can inform the experimenter about the level of confidence in statistical analyses. A numerical simulation can reveal if a proposed analysis is statistically under-powered for the noise level in the data using the results of a simple calibration.

Different analysis algorithms will provide significantly different results even when using the same data. Some analyses are more robust to across-subject changes in noise level than others. One experimental design approach to mitigate these problems is to use a within-subject design to test an experimental manipulation, and use paired within-subject measurements to test for the presence or absence of an effect. Individual differences in facial structure or mannerisms can have a large impact on data quality, so by performing within-subject comparisons, those differences are controlled.

References

1. Duchowski, A. T.: A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments, & Computers, vol. 34, no. 4, 2002, pp. 455{470. http://dx.doi.org/10.3758/BF03195475. 26

2. Rayner, K.: Eye Movements in Reading and Information Processing: 20 Years of Research. Psychological Bulletin, vol. 124, no. 3, 1998, pp. 372{422.

3. Salvucci, D. D.; and Goldberg, J. H.: Identifying fixations and saccades in eye-tracking protocols. Proceedings of the Eye Tracking Research and Applications Symposium, ACM Press, New York, New York, USA, 2000, pp. 71{78.

4. Komogortsev, O. V.; Jayarathna, S.; Koh, D. H.; and Gowda, S. M.: Qualitative and Quantitative Scoring and Evaluation of the Eye Movement Classification Algorithms. Proceedings of the 2010 Symposium on Eye-Tracking Research &#38; Applications, ETRA '10, ACM, New York, NY, USA, 2010, pp. 65{68. http://doi.acm.org/10.1145/1743666.1743682.

5. Komogortsev, O. V.; Gobert, D. V.; Jayarathna, S.; Koh, D. H.; and Gowda, S. M.: Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors. IEEE Transactions on Biomedical Engineering, vol. 57, no. 11, 2010, pp. 2635{2645.

6. Shic, F.; Scassellati, B.; and Chawarska, K.: The incomplete fixation measure. ACM, New York, New York, USA, Mar. 2008.

7. Andersson, R.; Nystr□om, M.; and Holmqvist, K.: Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. Journal of Eye Movement Research, vol. 3, 2010, pp. 1 { 12.

8. Kenneth Holmqvist, M. N. F. M.: Eye tracker data quality: What it is and how to measure it. Jan. 2012, pp. 1{8.

9. Ahlstrom, C.; Victor, T.; Wege, C.; and Steinmetz, E.: Processing of Eye/Head-Tracking Data in Large-Scale Naturalistic Driving Data Sets. IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 2, 2010, pp. 553{564.

10. Collewijn, H.; and Kowler, E.: The significance of microsaccades for vision and oculomotor control. J. Vis., vol. 8, no. 14, 2008, pp. 20.1{21.

11. Holmqvist, K.; Nystr□om, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; and Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.

12. Dill, E. T.; and Young, S. D.: Analysis of Eye-Tracking Data with Regards to the Complexity of Flight Deck Information Automation and Management—Inattentional Blindness, System State Awareness, and EFB Usage. 15th AIAA Aviation Technology, Integration, and Operations Conference, 2015. 27

13. Garca-Pfierez, M. A.; and Peli, E.: Intrasaccadic Perception. The Journal of Neuroscience, vol. 21, no. 18, Sept. 2001, pp. 7313{7322.

14. Bizzi, E.: Strategies of eye-head coordination. Prog Brain Res, vol. 50, 1979, pp. 795{803.

15. Leigh, R. J.; and Zee, D. S.: The neurology of eye movements. F.A. Davis, 2006.

16. Tole, J. R.; and Young, L. R.: Eye Movements: Cognition and Visual Perception, Erlbaum Associates, Digital Filters for Saccade and Fixation Detection. 1981.

17. Liston, D. B.; Krukowski, A. E.; and Stone, L. S.: Saccade detection during smooth tracking. Displays, vol. 34, 2013, pp. 171{176.

18. Green, D. M.; and Swets, J. A.: Signal detection theory and psychophysics. Wiley, 1966.

19. Bahill, A.; Clark, M. R.; and Stark, L.: The Main Sequence, A Tool for Studying Human Eye Movements. Mathematical Biosciences, vol. 24, no. 3-4, 1975, pp. 191{204.

20. Liston, D. B.; and Stone, L. S.: Oculometric assessment of dynamic visual processing. J. Vis., vol. 14, no. 14, 2014, p. 12.

21. Cover, T.; and Hart, P.: Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, vol. 13, no. 1, January 1967, pp. 21{27.

22. Jaccard, P.: The Distribution of the Flora in the Alpine Zone. New Phytologist , vol. 11, no. 2, 1912, pp. 37{50.

23. Di Nocera, F.; Camilli, M.; and Terenzi, M.: A Random Glance at the Flight Deck: Pilots' Scanning Strategies and the Real-Time Assessment of Mental Workload. Journal of Cognitive Engineering and Decision Making, vol. 1, no. 3, Dec. 2007, pp. 271-285.