Estimating Gaze Duration Error from Eye Tracking Data

John Hawkins
email john.hawkins@playgroundxyz.com

July 3, 2022

1 Abstract

Eye tracking applications produce a series of gaze fixation points that can be attributed to objects within a subject's field of vision. Error is typically measured on the basis of individual gaze fixation point measurements. These applications are often used to infer a gaze duration metric from a series of fixation measurements. There is no direct method for infering the error in a gaze duration measurement from an error in fixation points.

In this work we develop an algorithm for estimating the error bounds on gaze duration measurement through monte carlo simulation based on the content of an eye tracking calibration log file. We provide this algorithm as an open source application that allows for robust estimates of the error distribution of gaze duration measurements. We use this application to conduct experiments on the expected error bounds for different duration measurements across a range of session lengths and area of interest positions within the viewport of mobile devices. The results indicate that error in gaze duration estimation is sensitive to the size of fixation error as well as the proportion of total field of view that the object occupies and the proportion of viewing time that was taken with gaze fixation.

2 Introduction

Eye tracking is an important technology with a wide variety of applications. It permits evaluation of software user interfaces[1], the development of new forms of software interaction, as well as a potential method of biometric identification[2]. Eye tracking has become a core tool for empirical investigations into human behaviour and is widely used as a method of measuring explicit attention to visual stimuli. It has been applied to study psychological phenomena ranging from cognition[3] to mental health[4, 5], and is now routinely used to evaluate advertising [6]. Eye tracking technology has allowed marketing researchers to study many factors that contribute to effective advertising, including brand recall [7], the capture and tranfer of attention [8], the impact of images of faces [9], the attention effects of animation [10], and the relationship with social media posts [11].

Increasingly, eye tracking applications are built using machine learning models trained on data sets containing combinations of facial images and fixation coordinates collected through careful application design. The resulting gaze fixation prediction model can then be applied sequentially to determine how long a subject fixated on a given object. Thus, gaze duration is derived from aggregation of gaze fixation.

Gaze fixation models will typically have an error profile that is dependent on both the nature of the media being presented, the variety of faces and lighting conditions in the training data. In addition, it has been observed that the error profile varies depending on the location of the true fixation point within the subjects field of view. The errors in fixation point measurement for a given eye tracking solution are routinely adjusted through a process called calibration. Improvements in the callibration process are an ongoing focus for the development of algorithms[12, 13] and software tools[14]

When fixation models are used sequentially for the purpose of estimating gaze duration (as is commonly done in point of salience or advertising media studies), then there is the potential for the error to either cancel out (reducing error in gaze duration estimation) or to compound (increasing error in gaze duration). Which of these outcomes occurs will depend on the specifics of the model and the conditions under which it is being used. Influencing factors include the ratio of measured gaze duration to the total viewing time and the error profile of the machine learning fixation coordinate model.

The errors of eye tracking data technology can be decomposed into a range of independent sources that researchers need to consider [15]. Unless they are properly addressed these sources of error can manifest themselves as systematic biases across undesirable dimensions such as the age[16] or ethnicity of subjects[17].

The use of eye fixation models to estimate gaze duration (or dwell time) in Area of Interest studies (AOI) was discussed by Holmqvist et al. [15]. The authors simulated the impact of gaze fixation error by adding noise according to manufacturer specifications to data sets of low margin areas of interest. We extend this idea to calculate probabilistic bounds on the error in gaze duration using calibration data as a source of noise distribution that is specific to both the study participants and the device/environment of the study. The result is an open source application that may be used by a wide variety of reserachers to provide error bounds on any gaze duration measurements.

3 Methodology

We estimate the gaze duration measurement error through a Monte Carlo simulation using the eye tracking model's calibration file of fixation errors. The error profile of the eye tracking model allows us to generate sequences of true fixations and simulate measured fixations with the observed error of the model. These sequences of true and measured fixation are then converted into a distribution of expected gaze duration on a specifc area of interest, for a given measurement.

The algorithm consists of two core steps, first estimating the distrubtion of measurements for a given true duration. Secondly, inverting this into the distribution of true gaze durtions for a given measurement. Note, that the algorithm requires that

the area of interest be defined as a fixed bounding box and that the distributions are estimated for a fixed session length. This session length is the total period of time that the area of interest was in a subject's field of view, and thus represents the maximum possible gaze duration.

3.1 Measurement Distribution

In the first stage of the methodology we provide an estimate of the distribution of measured durations for all possible true gaze durations. These distributions are produced by generating random fixation paths across the available screen dimensions, and adding noise that is drawn from the calibration data to make it consistent with the observed properties of the gaze model. This produces a set of distributions that capture the expected variation in measurement depending on the length of true fixation.

Formally stated this is the probability distribution over measured durations $\hat{d}$ given a known duration d, the session length s and the gaze target location l. We express this distribution as $P(\hat{d}|d, l, s)$. The procedure for making this estimation involves iteratively generating random samples of gaze fixation paths $\mathcal{F}$ consistent with the parameters d, s and l. For each point f in $\mathcal{F}$ we determine a measurement point $\hat{f}$ by drawing samples from the eye tracking calibration set. The samples from the calibration data are used to determine the δ_x and δ_y error in the measurement of gaze fixation point f. We draw these delta values from the calibration file such the the probability of the point being chosen is propotional to its distance from the point f in the generated path. The complete algorithm for estimating $P(\hat{d}|d, l, s)$ is shown in Algorithm 1

Algorithm 1 Estimation of $P(\hat{d}|d, l, s)$

Input: l, s, C
Output: $P(\hat{d}|d, l, s)$

```
N ← samples                          ▷ Configure simulation samples per duration
D ← floor(s/increment)               ▷ Durations are sampled at discrete intervals
P ← Array(D + 1, D + 1)              ▷ The distribution P is a 2 dimensional array
for d ∈ 0toD do                      ▷ Iterate over all possible true gaze durations
    for n ∈ 0toN do                      ▷ Iterature to collect the N samples for d
        p ← generatePath(d, l, s)
        d_e ← generateMeasurement(p, l, C)
        P[d, d_e] += 1/N             ▷ Probability increment for a measuremnt of d_e
    end for
end for
return P
```

The bulk of the work is done by two functions inside the simulation loop. The first function $generatePath(d, l, s)$ takes a true gaze duration (for that simulation iteration) the location of the gaze target and the viewing session length. It then generates a random path of fixation points that is compatible with those parameters. The second function $generateMeasurement(p, l, C)$, takes the path, target location and the eye tracking calibration data and produces a sample of measured attention

time. It does this by using the calibration data to sample fixation errors and apply them to the path data. The noisy path data is then used to determine the estimated gaze duration by looking at the number of fixation points that intersect with the target location. Repeated application of this process for a given value of d produces multiple samples of $\hat{d}$ for that true gaze duration. Repeating the process over different values of d allows us to populate a two dimensional array where the first dimension is the true duration and the second the estimated duration. Note that our probability distribution is discrete because we restrict ourselves to sampling over a fixed set of increments between zero and the session length.

3.2 Gaze Duration Distribution

In real world applications we will have the true session length but rely on the model for the estimate of gaze duration. Meaning that the probability distribution we want is the distribution over true gaze durations for a given measurement, rather than the inverse which is what we have produced in the previous section. We produce an estimate of $P(d|\hat{d}, l, s)$ through an application of Bayes' rule, as shown in Equation 3.1.

$$P(d|\hat{d}, l, s) = \frac{P(\hat{d}, l, s|d)\dot{P}(d)}{P(\hat{d}, l, s)} \tag{3.1}$$

We use a uniform prior for $P(d)$, meaning in the absence of additional information all gaze durations less than the session length are equally likely. As our distributions are discrete estimations of an underlying continuous distribution the value of $P(d)$ is equal to $1/(1 + floor(10 * s))$ We can estimate the value of the denominator $P(\hat{d}, l, s)$ by iterating over all values of d and summing the product of $P(\hat{d}, l, s|d)\dot{P}(d)$. This fully explicated form is shown in Equation 3.2.

$$P(d|\hat{d}, l, s) = \frac{P(\hat{d}, l, s|d)\dot{P}(d)}{\sum_{\delta \in D} P(\hat{d}, l, s|\delta)\dot{P}(\delta)} \tag{3.2}$$

3.3 Implementation

The algorithm described in the preceding sections has been implemented as an open source python package called gazerr. It can be used as a software library, or deployed as a command line application. Source code is available on GitHub () and the package can be installed from PyPi.

The source repository for the gazerr application also contains a series of scripts for generating data sets and running the experiments outlined in the next section. Full details available in the package README file.

3.4 Experiments

We apply the gazerr application to investigate the relationship between measured gaze duration and expected real duration under a variety of changing scenarios. These

scenarios are variations in the underlying error indicated by the eye tracking calibration file and length of the session in which the user could be looking at the area of interest.

For the sake of our simulations we use a fixed sized simulated device with viewport pixel dimensions of 350 by 627. Our area of interest (AOI) is a fixed position medium rectangle ad unit (MREC) that is located toward the top of the screen. We iterate over a range of potential mean error values (determined to be the mean Euclidean distance between the measured fixation and actual fixation). For each of these mean error values we generate a synthetic eye tracking calibration file. We produce two versions of this calibration data, one in which the error is distributed around the true target, and a second biased version in which the error tends toward the upper left of the true position. We include the biased calibration data as it is consistent with our observations of real eye tracking data and we explore the extent to which this bias affects gaze duration measurements.

We feed each of these synthetic calibration files into gazerr to calculate posterior probability distributions over true gaze duration for the set of measured gaze durations. These posterior distributions are converted into expected values of true gaze duration for each measured duration. We repeat the above process for multiple session lengths to investigate the impact of increasing session length on the expected gaze duration for a fixed set of measurements.

4 Results

The first experimental result is the expected true duration against the measured duration for a single eye tracking application with a mean fixation error of 50 pixels. These results are shown in Figure 1.

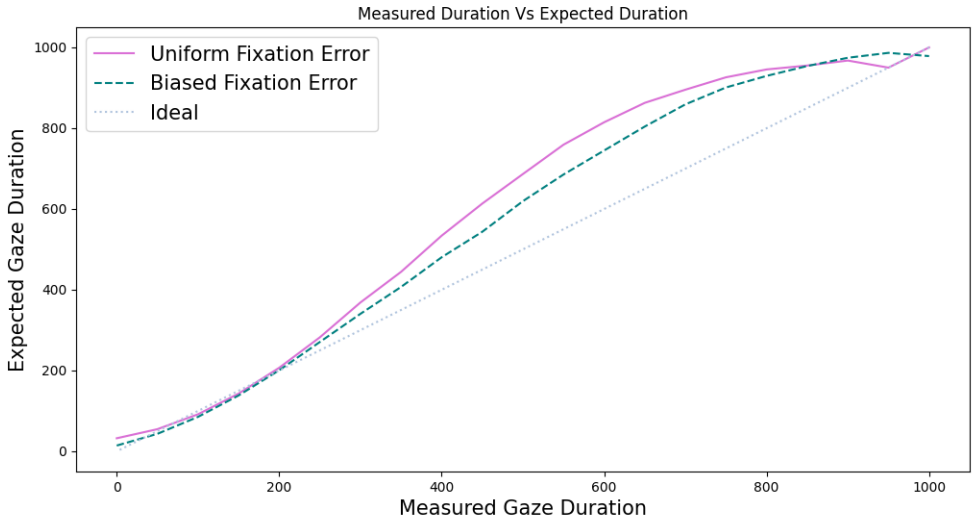

Figure 1: Measured versus expected gaze duration. MREC in 10s Session with 50px Mean Error.

The result shown in Figure 1 is indicative of what we see under multiple variations of the simulation configuration. At lower meaurement values there is a tendency for the measurement to be an accurate estimation, while at higher values it tends toward an under-estimation of the true gaze duration. Surprisingly the results using the biased calibration data seem to mitigate this effect somewhat, but the overall pattern remains consistent.

We next examined the way that exected error in the gaze duration estimate tracks with expected error in the gaze fixation points. This involved calculating the mean absolute error in gaze duration for each of the mean fixation error validation files. We performed this calculation using two methods. The first method, called uniform, assumes that all measured values are equally likely. The second method, called smoothed, assumes that low measurements are much more likely than higher measurement, and that this drop in likelihood follows an approximately exponential distribution.

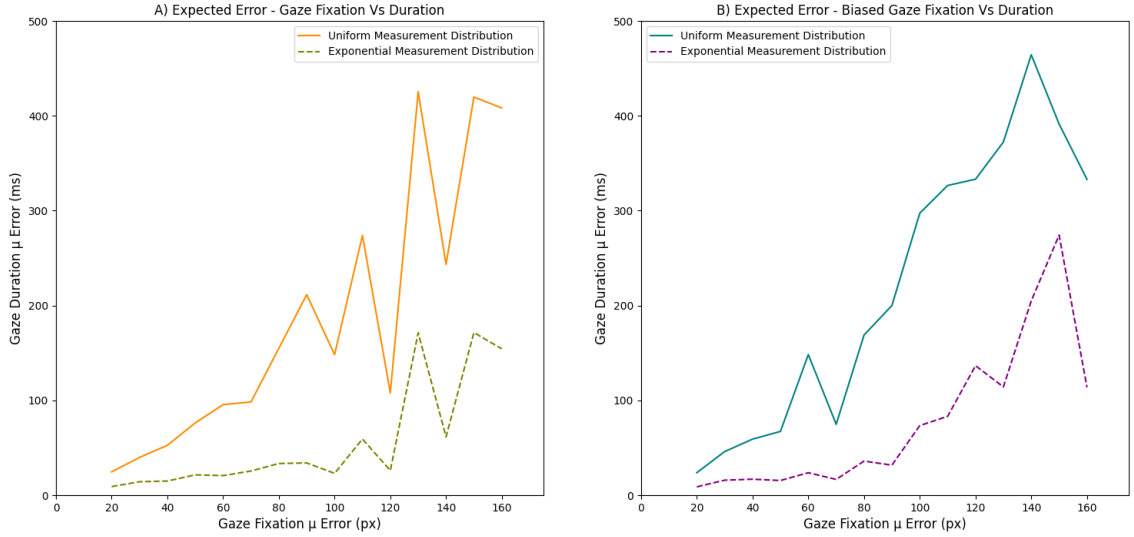

Figure 2: Fixation error versus gaze duration error.

As shown in Figure 2 we see that the error in gaze duration grows with error in fixation points, as would be expected. However, when we use the exponential weighting to calculate the expected error in duration the gaze duration error grows much slower than in the uniform case. This suggests that the observed tendency toward smaller measurements in area of interest studies means lower expected error overall. This pattern is true for the centrally distributed error (A) as well as the biased calibration error (B).

In our final experiment we look at the effect of both the size of the measured gaze duration and the mean error of fixation on the expected error in gaze duration. We display this as a three dimensional expected error surface over these two key dimensions in Figure 3

The left hand plot shows the complete error surface in which you will observe

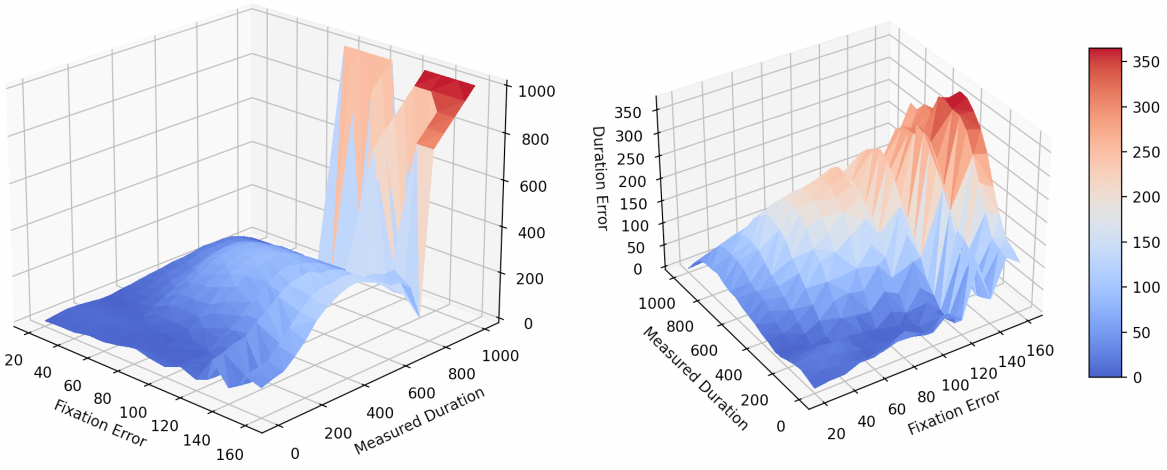

Figure 3: Mean gaze duration error over measured error and fixation MAE.

large spikes in error in the corner correspinding to high MAE for the fixation models and high measurements of gaze duration. This effect is caused by a complete absence of probability density in these regions, meaning that the error is maximal because the expected value is zero. This effect was consistenlt observed for eye tracking models with high fixation error and when the measured duration was high. Meaning that our model estimates that when fixation error is this high, measurements of these values become extremely unlikley to be observed.

We added the right hand plot in Figure 3 to illustrate the error surface in the absence of these deceptive outliers. We see that in general the expected error in a gaze duration is maximal around the mid point of the measurement range. The effect is lessened by lower MAE in the fixation models, but remains consistent for all levels of fixation error.

5 Conclusion

We have demonstrated that the statistial information contained in eye tracking validation data can be utilised for estimating errors in gaze duration estimation. The process involves simulating the distribution of measurements for a given true durastion, and then inverting it into a distribution over true duration for a given measurement.

We have released this approach as the open source application gazerr, and used it to explore the relationship between fixation error and duration error. Our results show that the nature of the error in gaze duration depends on the size of the measurement. Smaller measurements tend toward over prediction, and larger measurements towards under prediction. We plotted the relationship between expected error and saw that observed tendency toward smaller duration measurements (estimated with an psuedo-exponential distribution) delivers a slower growing expected duraiton error with the mean error in fixation points.

Finally, our simualtions revealed that for large error in fixation points certain measurements become some unlikely that our monte carlo simualtions placed none of the probability density in those regions. This strongly suggests that an additional downside to large fixation point error is a reduced range of measurement for area of interest studies.

References

[1] K. Harezlak, J. Rzeszutek, and P. Kasprowski, "The eye tracking methods in user interfaces assessment," in *Intelligent Decision Technologies*, R. Neves-Silva, L. C. Jain, and R. J. Howlett, Eds. Cham: Springer International Publishing, 2015, pp. 325–335.

[2] P. Kasprowski and K. Harezlak, "Biometric identification using gaze and mouse dynamics during game playing," in *Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety*, S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, and D. Kostrzewa, Eds. Cham: Springer International Publishing, 2018, pp. 494–504.

[3] T. Brunyé, T. Drew, D. Weaver, and J. Elmore, "A review of eye tracking for understanding and improving diagnostic interpretation," *Cognitive Research: Principles and Implications*, vol. 4, 12 2019.

[4] A. Duque and C. Vazquez, "Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 46, 09 2014.

[5] A. Rudich-Strassler, N. Hertz-Palmor, and A. Lazarov, "Looks interesting: Attention allocation in depression when using a news website - an eye tracking study," *Journal of Affective Disorders*, vol. 304, 02 2022.

[6] G. Hervet, K. GuĂŠrard, S. Tremblay, and M. Chtourou, "Is banner blindness genuine? eye tracking internet text advertising," *Applied Cognitive Psychology*, vol. 25, pp. 708 – 716, 09 2011.

[7] M. Wedel and R. Pieters, "Eye fixations on advertisements and memory for brands: A model and findings," *Marketing Science*, vol. 19, pp. 297–312, 11 2000.

[8] R. Pieters and M. Wedel, "Attention capture and transfer in advertising: Brand, pictorial, and text-size effects," *Journal of Marketing Journal of Marketing*, vol. 68, pp. 36–50, 05 2004.

[9] S. Djamasbi, M. Siegel, T. Tullis, and R. Dai, "Efficiency, trust, and visual appeal: Usability testing through eye tracking," in *In System Sciences (HICSS), 2010 43rd Hawaii International Conference*, 02 2010, pp. 1 – 10.

[10] K.-C. Hamborg, M. Bruns, F. Ollermann, and K. Kaspar, "The effect of banner animation on fixation behavior and recall performance in search tasks," *Computers in Human Behavior*, vol. 28, pp. 576–582, 03 2012.

[11] A. Barreto, "Do users look at banner ads on facebook?" *Journal of Research in Interactive Marketing*, vol. 7, 05 2013.

[12] Y. Zhang and A. Hornof, "Easy post-hoc spatial recalibration of eye tracking data," in *Eye Tracking Research and Applications Symposium (ETRA)*, 03 2014, pp. 95–98.

[13] A. Hassoumi, V. Peysakhovich, and C. Hurter, "Improving eye-tracking calibration accuracy using symbolic regression," *PLOS ONE*, vol. 14, p. e0213675, 03 2019.

[14] P. Kasprowski and K. Harezlak, "Etcal - a versatile and extendable library for eye tracker calibration," *SoftwareX*, vol. 8, 03 2018.

[15] K. Holmqvist, M. NystrĂśm, and F. Mulvey, "Eye tracker data quality: What it is and how to measure it," *Eye Tracking Research and Applications Symposium (ETRA)*, 03 2012.

[16] K. Dalrymple, M. Manner, K. Harmelink, E. Teska, and J. Elison, "An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood," *Frontiers in Psychology*, vol. 9, p. 803, 05 2018.

[17] P. Blignaut and D. Wium, "Eye-tracking data quality as affected by ethnicity and experimental design," *Behavior research methods*, vol. 46, 04 2013.