Important Considerations of Data Collection and Curation for Reliable Benchmarking of End-User Eye-Tracking Systems

Iakov Chernyak
K. K. FOVE
Tokyo, Japan
iakov.chernyak@fove-inc.com
jackch@mail.ru

Grigory Chernyak
K. K. FOVE
Tokyo, Japan
gregory.cherniak@fove-inc.com

Jeffrey Keith Spaneas Bland
FOVE Inc.
Torrance, California, USA
jeff.bland@fove-inc.com

Pierre Daniel Philippe Rahier
K. K. FOVE
Tokyo, Japan
pierre.rahier@fove-inc.com

ABSTRACT

In this article we discuss how to build a reliable system to estimate the quality of a VR eye-tracker from an accuracy and robustness point of view. We list up and discuss problems that occur at the data collection, data curation and data processing stages. We address this article to academic eye-tracking researchers and commercial eye-tracker developers with the purpose of raising the problem of standardization of eye-tracking benchmarks, and to make a step towards repeatability of benchmarking results. The main scope of this article is consumer-focused eye-tracking VR headsets, however some parts also apply to AR and remote eye-trackers, and to research environments. As an example, we demonstrate how to use the proposed methodology to build, benchmark and estimate the accuracy of the FOVE0 eye-tracking headset.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Human-centered computing** → **Virtual reality**.

KEYWORDS

data collection, data curation, eye-tracker benchmark, FOVE0

ACM Reference Format:
Iakov Chernyak, Grigory Chernyak, Jeffrey Keith Spaneas Bland, and Pierre Daniel Philippe Rahier. 2021. Important Considerations of Data Collection and Curation for Reliable Benchmarking of End-User Eye-Tracking Systems. In *2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Full Papers), May 25–27, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3448017.3457383

1 INTRODUCTION

There are an increasing number of research papers that contain eye-tracking benchmarks. This has led to a variety of inconsistent benchmark results, even for identical eye-tracking systems. For commercial eye-tracker comparisons, we can see that various papers report different results for eye-tracking accuracy due to different measurement methodologies[Clemotte et al. 2014]. Given the multitude of approaches to estimate accuracy, there is a need for increased standardization in order to make reliable comparisons.

Most papers with eye-tracker accuracy estimation have some description of the data collection and processing, however we found that this description is quite often incomplete, or lacks sufficient data to represent population coverage and robustness, and the results are often difficult to reproduce. There is not much work publicly available which provides sufficient data collection and processing information to make fair eye-tracker assessments.

An extensive summary regarding eye-tracking quality, and the influencing factors therein, was made in [Holmqvist et al. 2012]. It highlights multiple problems inherent to the benchmarking of eye-trackers, and lists various sample filtering approaches to remove invalid data. However, we found that it does not propose a solution for these problems that would help the benchmark results to be reproducible.

The paper [Lohr et al. 2019] extensively describes various ET characteristics, and shows how to build a benchmark using an SMI eye-tracker as an example. However, there is no explanation of how the outlier removal process was designed, how the thresholds were chosen, and the overall impact of the curation remained unclear. We also found that the accuracy of the raw data presented for one of the subjects is different from the total accuracy claimed in the paper after data curation, which raises the question of how the accuracy is distributed across participants.

State-of-the-art works regarding objective eye-tracker quality estimation propose to establish ground truth gaze direction by using another more accurate eye-tracker simultaneously. Many retinal feature tracking devices can provide ground truth with the accuracy of fractions of arcminutes[Sheehy et al. 2015]. Beside high price and complexity, this method is not always mechanically compatible with other eye-tracking devices. The magnetic search coil method is another very precise approach, but requires wearing contact lenses, which makes collection of large amounts of data complicated in practice[Collewijn et al. 1975]. Usage of artificial eyes with a high degree of orientation control might not be an accessible solution for researchers, and it cannot be used to measure robustness, since it does not represent the full diversity of human eyes[Reingold 2014].

This work is licensed under a Creative Commons Attribution International 4.0 License.

ETRA '21 Full Papers, May 25–27, 2021, Virtual Event, Germany

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8344-8/21/05.
https://doi.org/10.1145/3448017.3457383

Numerical eyeball simulation can provide high eye variety, however it partially substitutes human fixation inaccuracy with inaccuracy of the simulation, which can either positively or negatively bias the result of the assessment.

Despite the existence of these advanced methods, most researchers prefer to display stimuli on a screen and use it as ground truth due to the simplicity of this approach. In addition, the recent development of eye-tracking technology for virtual reality (VR) headsets creates new opportunities for eye-tracking evaluation, since these devices are capable of displaying stimulus with high spatial and temporal accuracy.

Presently, eye-tracking is used most often in a laboratory or other controlled setting where an operator is present to ensure that the accuracy of the data collected is as high as possible. However, as eye-tracking is introduced into more consumer devices, an increasing number of people are using eye-trackers in more casual settings. These users, herein referred to as "end-users", usually do not have an operator present, and may not take steps to increase the eye-tracking accuracy, so long as their device is functioning well enough to suit their needs. For this reason, we distinguish two general categories of eye-tracking use cases, as shown in Table 1. Benchmarks designed to estimate eye-tracking characteristics for each category are different in principle. The differences are highlighted in Table 2. The benchmark included in this paper is geared towards end-user eye-tracking, however some of the methods described apply to both categories.

We want to emphasize that with this approach, we also make the result more reproducible, because we minimize the amount of variables and thresholds necessary for benchmarking. The results do not depend on how experienced the operator is, how many calibrations participants completed before actual eye-tracking measurements, or how many "improper" users were excluded.

In this paper we describe various problems that occur at data collection, data curation and data processing stages. We list up factors that can affect accuracy estimation, and discuss how the impact of these factors can be mitigated. Some factors which are difficult to control should be explicitly accompanied with the provided data, giving additional insight to the dataset and processing methods. After that, we demonstrate how to build a benchmark to estimate the accuracy of the FOVE eye-tracker.

2 DATA COLLECTION

Data collection is the most difficult to control part of benchmarking. In this chapter we describe our data collection process and comment on important aspects of collection that have to be considered.

Commonly, eye-tracking video processing can be done in real time during data collection. In that case, the participant calibrates the eye-tracking system and follows the stimulus, while the eye-tracker estimates the gaze direction and compares it with the ground truth. This approach does not require access to the eye-tracker's internal logic, and in some situations it is the only way to estimate performance of the eye-tracking system. In order to allow deeper analysis, video recordings of the eyes during the calibration and fixation segments should be made. The eye images from the camera, paired with the position of stimulus, allow us to compare and assess various eye-tracking methods and calibration approaches across the same data. For example, in our data collections, we begin by asking participants to complete two calibration processes: the smooth pursuit calibration and the single-point calibration. Later, we can playback the eye video to estimate the accuracy of both calibrations against the same data. In addition, it helps us to analyze failures, and identify potential eye-tracking improvements.

In our data collection, the calibration process is "blind". At the time of data collection, we do not know how good the calibration was, or even whether it was successful or not. The reason for the blind calibration process is that, in reality, a typical end-user will recalibrate the eye-tracker only in the case of explicit calibration failure. Arbitrary thresholds or operator judgements that govern whether to recalibrate during the data collection process will affect the overall rate of failure, forcing participants to iterate calibration until it has the best quality, thus creating the potential for manipulation of the final benchmark result. The blindness of the calibration aims to equalize different eye-tracking calibration fail rates, and helps to avoid unintentionally biasing the results towards better accuracy than would be experienced in real-world conditions.

Nevertheless, blind calibration processes can be used only for eye-tracking systems that can continue emitting eye-tracking data regardless of whether or not the calibration succeeded. Many eye-trackers on the market prevent eye-tracking to be used after explicit calibration failure. In that case, if blind calibration is impossible, we propose to accompany the benchmark data with information about the amount of recalibrations.

After calibration is completed, participants have to look at a sequence of 30 fixation points. Participants look to the point, and press a button at the time that they think they are fixated on it. They continue to look for the point for a short period, until the next fixation target is displayed. User-provided timing is shown to have better accuracy when compared to operator-controlled timing[Nyström et al. 2013]. There are multiple ways that recordings are timed. Some researchers skip a small amount of time before starting eye-tracking data acquisition, and others record data prior[Blignaut et al. 2014]. In our case, we record eye-tracking frames immediately after the user clicked, because it more closely approximates the timing of end-user use cases, such as when pressing a button to control a user interface. Participants are permitted to remount or adjust the VR headset any time they feel the need to, since the end-user may also do so.

The field-of-view varies based on the user's facial structure and VR headset mount position, and thus it is impossible to know the boundaries of what is visible for each participant ahead of time. In order to ensure that the entire field-of-view is covered, the area where fixation targets can appear should be wider than the normal headset field-of-view. If the participant cannot find the target, they are asked to press the "skip" button to show another random point instead. The fixation point probability distribution across the field may vary, however it is important to provide information about average angular deviation of the target from the center, because this impacts the final accuracy. The eye-tracking maximum range of the headset is estimated from the eyeball rotation pivot, as opposed to the field-of-view, which is measured from the pupil as shown in Fig 1.

Another aspect that affects the gaze range is the physiological limits of the human oculomotor system. The maximum eyeball

Table 1: Differences between operator-assisted and end-user eye-tracking use cases.

Operator-assisted eye-tracking	End-user eye-tracking
Example applications: Follow simple moving stimulus, watch video, move around in a fixed 3D environment, and later build heatmaps, analyze fixations and saccades.	Example applications: Keyboard typing with gaze, UI operation, aiming at enemies in games, real-time recognition of user focus in virtual or real environments.
Eye-tracking data is recorded, and can be reviewed by the operator.	Eye-tracking data may be interactive and may affect the user in real time.
An operator is present to guide the eye-tracking experiment.	No operator is present.
The operator or user themselves can see the gaze point and confirm gaze quality.	Users may or may not be aware of gaze estimation.
Multiple calibrations can be performed in order to pick the best one to maximize quality.	Users rarely recalibrate, unless the eye-tracker itself reports unsuccessful calibration, or the accuracy is low enough to make the application difficult to use.
The eye-tracking environment is always optimized for the best results. Operators help participants mount the headset for ideal performance.	Eye-tracking may happen in any environment, but application-provided guidance is sometimes available to help users improve the eye-tracking experience.
Eye-tracking participants are preselected in order to improve eye-tracking quality for some study.	There are no restrictions on who can be the end-user.

Table 2: Difference in methodology for benchmarks that target operator-assisted eye-tracking vs end-user eye-tracking.

Operator-assisted eye-tracking	End-user eye-tracking
After data collection is done, a custom approach to pick the human fixation timing can be implemented. Operators may automatically or manually select the most ideal fixations and saccades[Morgante et al. 2011].	Participants complete an automated benchmark, where the participant gazes towards the targets and picks the moment when they believe they are looking at the target.
The resulting quality depends on the operator's experience[Hessels and Hooge 2019].	Operators are optional for data collection, but they should not bias the final score of the benchmark, as this would not be representative of the end-user conditions.
Dataset may exclude people with improper eye shape, eyelashes, or other factors[Tobii Technology 2012].	Dataset includes as wide a variety of people as possible.
Given the above, data curation aims to filter out poor data caused by both eye-tracker and participant error.	Data curation aims to only filter out poor data caused by participant error.

rotation angle is 20 to 60 degrees, with the upper part of the visual field normally being narrower than the rest. The values vary greatly between individuals. Individual oculomotor limits combine with the limits of the headset structure and its mounting position to further restrict the maximum gaze range and make it user-dependent[Lee et al. 2019].

Each time a new fixation point is displayed to the participant, the background brightness is randomly changed to increase variety in pupil dilation. Dark backgrounds should be well represented in the dataset, since dilated pupils have different pupil centers, and cause additional challenges to eye-tracking[Choe et al. 2016].

Since most eye-trackers operate based on pupil position, pupil dilation is a potential source of eye-tracking inaccuracies due to the physiological difference between the pupil center position for different pupil dilations. For each eye, the scale of the error has an average amplitude of 1 degree, but in exceptional cases can reach 5 degrees[Yang et al. 2010]. Even with pupil dilation calibration, this effect can be compensated only partially[Drewes et al. 2012]. Fortunately, the horizontal components of the shifts for the left and right eyes are similar in magnitude and opposite in direction[Anne et al. 1992; Drewes et al. 2014]. When the gaze convergence vector is computed by taking the average of the left and right gaze vectors,

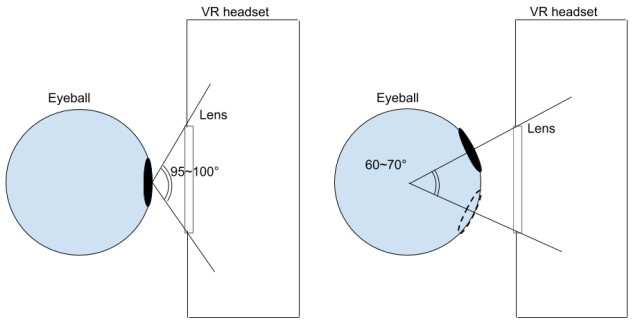

Figure 1: Difference between field-of-view (left) and possible range of gaze (right).

the horizontal components will be partially cancelled out. Thus the gaze convergence can reduce the negative impact of this effect on the resulting gaze direction accuracy. In our data collection, when the background brightness is changing, we do not wait for the pupil to adapt to the new brightness completely. So extreme dilation cases can be underrepresented in our data.

Since the resulting eye-tracking accuracy is very dependent on whether the person wears eyeglasses or not, we split all the data into two categories: with and without eyeglasses. The category without eyeglasses also includes all the participants with contact lenses and colored lenses. We found that the ET accuracy is worse for people who wear contact lenses, since contact lenses alter the cornea surface shape. In principle, if eye-tracking characteristics have a strong correlation with a certain factor, then the datasets should be split into subsets according to this factor. However, the eye-tracker uses the same image processing for people who wear contacts as for those who do not, so we decided to combine them into the same group.

After the data has been divided into categories, every subset of the data has its own resulting eye-tracking characteristics, and this will prevent the results from being dependent on the ratios that these factors are represented in the data. Alternatively, data can be accompanied with a distribution of the participants over this factor, and then this data can be grouped together. This especially makes sense if there is not much data corresponding to a certain factor, such that it will not provide statistically significant results.

Besides eyeglasses / non-eyeglasses / contact lens / colored lens categories, the accuracy might also depend on eye color, skin color, makeup, conditions and surgeries of the subject[Tobii Technology 2017]. Certain conditions (such as nystagmus, strabismus, amblyopia, blindness in one eye) might lead to eye-tracking failure, and for some eye-trackers it might be reasonable to separate them into a different subset in the case of enough statistical significance, or exclude them from the dataset with a corresponding notice in the report otherwise.

3 ESTIMATION OF EYE-TRACKER CHARACTERISTICS

Spatial accuracy is one of the major characteristics of eye-tracking. There are two ways to define it, where the first is the distance between the ground truth point and the centroid of gaze samples[Vehlen et al. 2021], and the second is the mean distance between the ground truth point and the gaze samples[Holmqvist et al. 2012; Lohr et al. 2019].

In the case of the FOVE eye-tracker, the standard deviation of the gaze estimation samples for a single fixation is numerically lower than the accuracy, so averaging multiple gaze samples does not have much impact in general. Furthermore, averaging gaze data over a long period increases the probability of including an undesirable saccade or blink. For accuracy measurement, we use only the first recorded frame from the eye-tracker that corresponds to a given fixation point, without averaging over time.

In contrast to precision, the accuracy depends on how we define the ground truth gaze vector. In the case of a head-mounted display (HMD) that is capable of displaying stimulus on a screen, there are two ways to define it. One is based on the rendering projection matrix, and another is bound to the real world.

The projection matrix-based approach is the most widely used, where the ground truth gaze vector is the coordinate of the 3D target in the VR scene relative to each eye. Eye-tracking aims to match the estimated gaze to the virtual environment rather than to the actual eyeball orientation. This is what the most end-users normally expect from eye-tracking in VR headsets.

However, the vector from the camera towards the 3D target in the VR scene is not always parallel to the actual visual axis of the eyeball. In fact, the digital correction with barrel distortion does not perfectly compensate for distortion that is caused by the HMD lens. The further the pupil is located from the lens's optical axis, the stronger the difference will be[Martschinke et al. 2019]. FOVE0 has no built-in adjustment of the HMD lenses to match the inter-pupillary distance of the user, causing an additional discrepancy between the virtual and real world.

In order to estimate the accuracy of eye-tracking in the real world, we have to estimate the expected eye visual axis for the displayed stimulus. If we know the exact physical coordinate of the highlighted pixel on the HMD screen, and the properties of the lens and gaze origin, we can solve an inverse problem to find the direction of the ray that starts at the gaze origin and is refracted by the lens to hit the target pixel on the screen. The resulting direction is the ground truth gaze.

It differs from the projection matrix-based approach in that this approach is fully physically based. It eliminates any inaccuracy caused by imperfect lens distortion compensation. However, it requires knowledge of the internal structure of the headset, the position of the screen, and the properties of the lens. It also needs the 3D gaze origin relative to the headset, which is part of the eye-tracking output. We believe that this is not a major issue, since normally the HMD optical system is designed in a way that insignificant gaze origin shifts do not affect much the observed image within a reasonable tolerance, and that becomes a secondary order parameter to the ground truth inaccuracy.

We believe the projection matrix-based approach is well suited to 3D VR content, while the real world-based approach suits certain medical applications and AR.

4 DATA CURATION

In order to reduce the impact of human error on the final assessment, the data has to be properly curated. In the case of our data collection, every fixation section has a sequence of random points that appear on the screen, and the participant should look at the currently displayed target and press the "accept" button. If the target is out of their field-of-view, the participant should press the "skip" button. The majority of the participants are first-time VR users, and they are almost entirely first-time eye-tracking users. This has a negative impact on the quality of the collected data, because the participants are more likely to make an error due to unfamiliarity with the technology. Even with proper instructions, some participants accidentally press the "accept" button not when they are looking at the target, but before or afterwards. Sometimes participants press the "accept" button instead of the "skip" button, and in that case they are likely to be looking in a totally different direction.

Involuntary eye movements are another source of human error[McCamy et al. 2013; Yuval-Greenberg et al. 2014]. Microsaccades can have amplitudes up to 0.5 degrees, such that the visual axis may not directly point at the target point, but somewhere nearby. One approach[Nyström et al. 2013] is to identify separate fixations within the recording window, and choose the closest fixation to the target. We strongly oppose this method, because samples can be collected until one is sufficiently close to the ground truth. The extent of the microsaccade range thus becomes the extent to which accuracy can be arbitrarily increased, given a sufficiently long window. Another way[Dalrymple et al. 2018] is to use the longest fixation instead, which is less biased, but very dependent on the fixation detection algorithm that was chosen, and how the participant-specific thresholds are selected[Saez de Urabain et al. 2015]. Algorithms that behave differently based on subtle changes in the data collection procedure, should be avoided to ensure the benchmark reflects the accuracy of the eye-tracker itself, rather than the quality of the data. We do not know of any algorithm that can eliminate the effect of microsaccades without biasing the results. We propose to apply no special algorithms and simply keep in mind that estimated spatial accuracy, even with a perfect eye-tracker, will not converge to zero. In the best case, it will converge to an average human microsaccade range.

Researchers use various data curation processes in order to estimate eye-tracking characteristics and reduce the impact of low quality data. Different approaches to curation lead to different results in estimated characteristics. We describe two major approaches for data curation: "pick the best" and "filter out the worst".

The former underlies many strong curation filters and aims for complete elimination of invalid data and unreliable ET results. The primary goal is to show the maximum accuracy of the eye-tracker in the best-case scenario, where ET has maximum confidence in its results. The main advantage of this approach is the possibility to weaken various criteria one-by-one and see how the eye-tracking accuracy depends on each of them while other conditions remain perfect. For example, [Tobii Technology 2012] uses curation that aims to select good eye-tracking scenarios by filtering out the entire fixation point if the eye-tracker failed on at least 20% of the total frames.

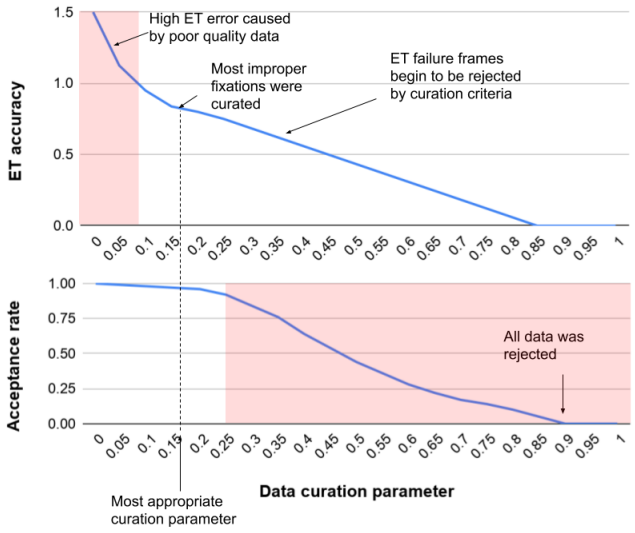

Figure 2: Schematic dependency of the resulting ET accuracy (top figure) and fixation point acceptance rate (bottom figure) for different filtering threshold selections.

Another approach is to "filter out the worst," which aims to filter out bad quality data while avoiding filtering poor results from the eye-tracker. This approach is supposed to be used to represent robustness of accuracy, or the actual accuracy that the average eye-tracking user will experience. To understand the difference with the previous approach, we are not going to filter out the point unless there is not even a single valid gaze sample coming from the eye-tracker. The data curation criteria should be designed in a way to explicitly separate bad data from poor ET output. The resulting accuracy should be stable against small changes to the thresholds used in the curation method. If the curation approach violates this rule, then by adjusting thresholds we can arbitrarily manipulate the resulting eye-tracking accuracy.

In order to estimate the quality of some curation method, we propose to identify the threshold parameter used in the method and investigate how the results depend on it. For simplicity here, we assume that if the parameter is equal to zero, then no filtration is applied, and the larger the value of parameter, the more filtration is applied. The amount of collected data that survived curation, as well as the resulting ET accuracy, can be studied against different values of the selected threshold parameter. We suggest to plot this dependency for every curation method separately while other curations are disabled. This allows us to pick the most appropriate value for the threshold to curate invalid data while minimizing curation of low-quality eye-tracker output.

A typical function for a well designed curation approach is shown in Fig 2. The ET accuracy dependency starts with a steep improvement and is followed by a region with small deviation against the data curation threshold. Beyond that point, the dependency of the fixation point acceptance rate begins to stabilize, and then with higher parameters the curation rejects more and more frames. We claim that the value of the parameter where the accuracy plot

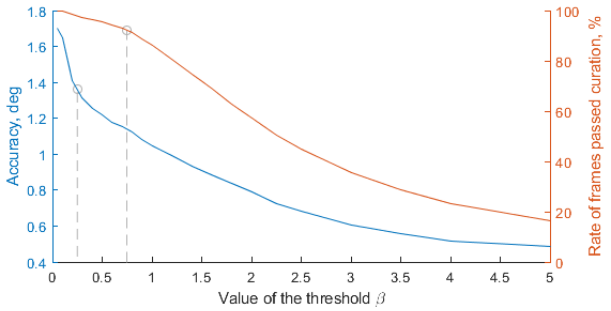

Figure 3: Dependency of the fixation point acceptance rate (red line) and the resulting ET accuracy (blue line) for the curation based on eye difference.

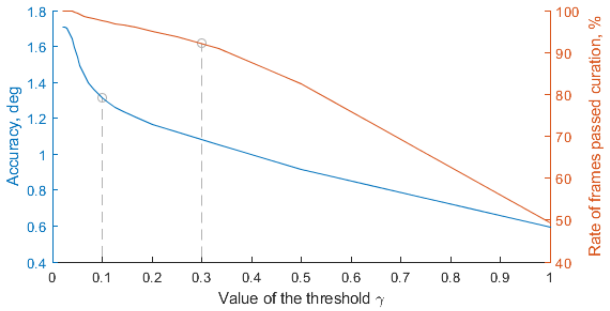

Figure 4: Dependency of the fixation point acceptance rate (red line) and the resulting ET accuracy (blue line) for the curation based on distance from ground truth.

stabilizes, and the acceptance rate has not yet started to drop significantly, is an optimal threshold value. This comes from the assumption that there is not much poor quality data quantitatively, but these data have a strong impact on the final statistics of the eye-tracking accuracy.

Below we show an example of designing curation methods and selecting parameters for them. This curation will be used for the FOVE headset benchmark in the next chapter. Assuming that the eye-tracker output is independent for the left and right eyes, then the probability that failed eye-tracking for the left and right eyes coincidentally gives the same results is very low. We can mark frames as invalid if the gazes for the left and right eyes are close to each other, but both of them are far from the ground truth direction. This usually indicates that eye-tracking worked well, but the person did not look at the target point properly. The condition when we are going to reject fixation points can be represented as:

$$Angle(G_L, G_R) < \beta * Angle(\frac{G_L + G_R}{2}, G_T) - 0.5^\circ \quad (1)$$

Here G_L and G_R are the left and right gazes obtained from eye-tracker. G_T is the ground truth direction. We added an additional empirical constant 0.5 degrees in order to guarantee that no eye-tracking data will be filtered out within $0.5/\beta$ degrees from the ground truth. The resulting rejection rate significantly depends on the value of the threshold β.

The dependency of the fixation point acceptance rate (red line) and the resulting ET accuracy (blue line) is shown in Fig 3. The stationary region of the fixation point acceptance rate line extends until $\beta = 0.75$, while the accuracy line stops rapidly falling at around $\beta = 0.25$. From here we can see that the threshold parameter $\beta = 0.5$ would filter most of the bad quality data, with minimal impact on the amount of filtered data.

This curation method assumes that the data for both the left and right eyes are available at the moment, which is not guaranteed. We can use additional curation that omits any fixation points that have accuracy worse than some threshold. Normally, eye-tracking does not give very large errors in the case that the participant has no eyeglasses, but there are quite a lot of cases when subjects pressed the "accept" button while they were not looking at the target point. This happened often when the point was on the edge of the screen and the subjects could not find it. The condition by which we are going to reject fixation points can be represented as:

$$Angle(G_{Conv}, G_T) > 1/\gamma \quad (2)$$

Where $G_{Conv} = 0.5(G_L + G_R)$ if both eyes gaze available, and $G_{Conv} = G_L$ or G_R if only one eye was tracked.

This is a very dangerous curation where we can guarantee an ET accuracy at least $1/\gamma$ degrees. The dependency for the fixation point acceptance rate and average ET accuracy on γ is shown in Fig 4. The fixation point acceptance rate stationary region extends up to $\gamma = 0.3$, while the accuracy figure decreases its declination from $\gamma = 0.1$. We can see that the most appropriate value for γ is around 0.2, which is equivalent to the rejection of any fixation points that have accuracy worse than 5 degrees.

5 DISCUSSION

We demonstrate how to apply the proposed methodology to build a benchmark and estimate the accuracy of the FOVE eye-tracking headset. We used data from 157 people without eye-glasses, and the majority of these people were first-time eye-tracking and VR users. The raw data for every participant can be downloaded from https://github.com/jbfove/BenchmarkPaperData [Chernyak et al. 2021].

In our benchmark, we had two sequences of fixation points where each contains 30 targets with random uniform distribution in 3D space. Headset remount was compulsory in between these sections. Thus the first set of fixation points reflects the eye-tracking immediately after calibration with a low rate of headset readjustment, while the second set of fixation points will show eye-tracking after headset remounting.

We had 17.8 degrees of average angular deviation of the targets from the forward direction, while the possible maximum gaze range is about 30 to 35 degrees for FOVE0. In our data collection, the minimal distance to the target is 2 meters. At the data curation stage, we further restrict the range of the fixation targets to 25 degrees, as it is the range where we are guaranteed that the targets are visible to both eyes.

After every fixation target shown, the user should look at the stimulus and confirm by clicking. Any fixation point is considered to be successful if at least one frame in the recording had successful eye-tracker output (this filters out 1.56% of the total data). To

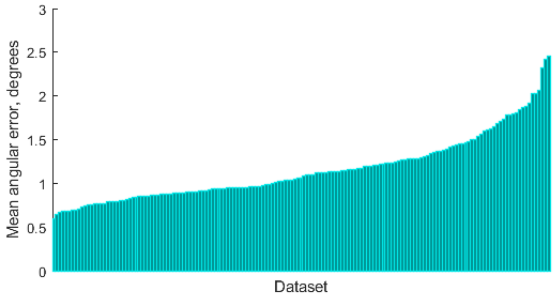

Figure 5: Accuracy per dataset (sorted).

measure accuracy we will use gaze convergence, the average gaze between the left and right eyes. If at a certain frame only one eye had valid gaze output, that eye gaze is used as the gaze convergence.

The benchmark curation was based on the "filter out the worst" principle, so we will avoid filtering points with failed eye-tracking. In order to curate data, we used the two filtration methods described in the previous chapter. The benchmark was divided into 2 sub-datasets: non-eyeglasses and eyeglasses. In this paper we include only data taken from people not using eyeglasses, but we include people with contact lenses and color lenses. We used the projection matrix-based definition of the ground truth, as it is the most reproducible.

As for numerical values, we prefer to use Mean Absolute Error (MAE) over Root-Mean-Square Error (RMSE). RMSE is a good indicator to see how many error spikes the ET algorithm has. In practice, the amount of outliers is more dependent on data quality and data curation methods, rather than on the ET algorithm. Furthermore, there is an ambiguity when calculating RMSE. It can be calculated across all fixation points and across all participants, or it can be the mean value across participants, and the RMSE across fixations of each person.

In order to estimate the statistical significance, we propose to calculate confidence intervals where the average accuracy for each subject is considered as a separate sample. This allows each sample to be statistically independent from each other. The confidence interval for accuracy with probability p becomes:

$$CI(p) = T_p(N-1)std_{ang}/\sqrt{N} \quad (3)$$

Where N is the number of participants in the dataset, $T_p(N-1)$ is the inverse of student's t-cumulative distribution for probability p and degree of freedom $N-1$, std_{ang} is a standard deviation of mean eye-tracking accuracies across participants.

Table 3 shows the statistical data on people who participated in the data collection. The benchmark results may change based on different statistical distributions. For example, we found a noticeable correlation between eye-tracking accuracy and the presence of contact lenses, makeup (and sex as a consequence) and the ethnicity of the participants. After building a linear regression model we got similar results to [Blignaut and Wium 2013], and in our case Asian people on average showed 0.147 degrees worse accuracy compared to non-Asian people (p-value 0.013). Participants with contact lenses experienced an accuracy reduction of 0.203 degrees (p-value 0.018). The most impactful factor was colored contact lenses, with a reduction of 0.362 degrees (p-value 0.003).

Table 3: Distribution of data collection participants

Feature	Value	Amount
Sex	Male	88
	Female	69
Contact lenses	None	131
	Yes	18
	Colored	8
Ethnicity	Asian	88
	Black	4
	Hispanic	11
	Middle Eastern	1
	White	53
Makeup	None	112
	Yes	45
Eye Color	Amber	3
	Blue	21
	Brown	63
	Dark Brown	61
	Green	9
Conditions	None	149
	Anisocoria	1
	Cataract (both eyes)	1
	Cataract (one eye)	2
	Presbyopia	1
	Prescription	2
	Strabismus	1
Surgeries	None	146
	Lasik	9
	Lens Replacement (both eyes)	1
	Lens Replacement (one eye)	1
Birth year	1940s	2
	1950s	1
	1960s	16
	1970s	21
	1980s	49
	1990s	62
	2000s	6

The result of the accuracy benchmark for participants that completed FOVE0 smooth pursuit calibration is shown in Fig 5. Every bar there represents an individual participant's accuracy. We performed an ascending sort of bars with MAE, so we can see the range of how ET accuracy varies from person to person. The overall mean eye-tracking accuracy is 1.136+0.058 degrees (confidence interval 95%).

6 CONCLUSION

In this paper we listed up important factors that have to be considered when building a benchmark to measure the performance of a VR eye-tracker for end-users. A short summary of these factors can be found in Table 4. We proposed a reproducible methodology to estimate eye-tracker accuracy with a minimal amount of dependent

Table 4: List of factors

Factor	What we propose	Why
Participant characteristics	Collect participant data that can affect quality (Age, Iris color, Skin color, Conditions and medical treatment history, make up, Eyeglasses/ contact lenses/ colored contact lenses. Was it the first time the user had used an eye-tracker or VR?	These data can be used to identify the impact of each factor, for example via linear regression. It also makes it possible to subdivide the dataset.
Entire dataset rejection for some set of participants	Avoid dataset rejection, if possible.	Different rejection strategies would lead to different and unreproducible results.
Recalibration	Avoid user recalibration except when the eye-tracking system completely fails to work.	The number of recalibrations can be operator dependent, so the result will not be reproducible. If there was some recalibration due do ET failure, the amount should be described.
Spatial distribution	Use a uniform distribution of fixation points, and describe how the points are selected to ensure the distribution.	The distribution type affects the final overall score, because different areas in the FOV may have different accuracies.
Represent pupil dilation variety	Optionally randomize brightness of background during collection.	It simulates eye-tracking in various environments. Data collected while the pupil is dilated may yield lower accuracy.
Eye-tracker drift	Optionally include an eye-tracker drift section during data collection.	It helps to understand how drift affects the quality of eye-tracking.
Fixation point timings	Allow the participant to control the timing.	It improves accuracy of the benchmark and simulates end-user experience.
Gaze sample timing	Use the frames right at the moment of clicking	It corresponds to an end-user interactive experience.
Curation strategy method	Use "filter out the worst" for general eye-tracker benchmarks.	It seeks to estimate the typical accuracy rather than the base-case accuracy, and minimizes the amount of threshold parameters needed.
Curation	Describe every curation method and the chosen threshold parameters.	This helps other researchers build comparable benchmarks or understand where benchmarks are not comparable.
For VR headsets: Definition of ground truth	Use virtual 3D world ground truth for VR-based eye-trackers.	It's easy to reproduce, and does not require knowledge of internal structure of the HMD.

factors. We want to emphasize that the context of how the final ET accuracy value is obtained is as important as the value itself. We recommend, for the sake of getting roughly comparable and reproducible accuracy numbers, avoiding methodologies that allow unintentional manipulation of the final result. We encourage VR eye-tracker manufacturers as well as researchers to provide more information about their data curation process, and design filters using the "filter out the worst" approach for general benchmarks, since it better measures population coverage, an important characteristic of eye-trackers that is often not included. We also suggest that data on accuracy distribution across participants be included, since it is a simple way to show the robustness towards all varieties of people.

We showed an example of a complete benchmark report that includes a description of data collection, and the procedure of building curation filters for fair estimation of the eye-tracker accuracy. Nevertheless, many decisions that were made in this paper need further investigation. For example, in our data collection, we asked participants to fixate on 30 fixation points twice, before and after headset shift. In this benchmark, these data were combined, however it might be worthwhile to explicitly show how headset drift affects accuracy. In addition, we asked participants to fixate on a total of 60 points, and we have not studied yet how the number of points may affect benchmark quality, since the participants may become increasingly fatigued.

In future research, we plan to investigate further the correlation between eye-tracking accuracy and various characteristics of the participants. We hope to better understand how each factor can negatively impact eye-tracking quality, and thereby develop methods to mitigate such effects.

REFERENCES

Wilson M. Anne, Campbell C. Melanie, and Simonet Pierre. 1992. The Julius F. Neumueller Award in Optics, 1989: Change of Pupil Centration with Change of Illumination and Pupil Size. *Optometry and Vision Science* 69, 2 (1992), 129–136. https://doi.org/10.1097/00006324-199202000-00006

Pieter Blignaut, Kenneth Holmqvist, Marcus Nyström, and Richard Dewhurst. 2014. Improving the accuracy of video-based eye tracking in real time through post-calibration regression. *Current Trends in Eye Tracking Research* (2014), 77—-100. https://doi.org/10.1007/978-3-319-02868-2_5

Pieter Blignaut and Daniël Jacobus Wium. 2013. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods* 46, 1 (2013), 67–80. https://doi.org/10.3758/s13428-013-0343-0

Iakov Chernyak, Grigory Chernyak, Jeffrey K. S. Bland, and Pierre D. P. Rahier. 2021. Benchmark Paper Data. https://github.com/jbfove/BenchmarkPaperData.

Kyoung Whan Choe, Randolph Blake, and Sang-Hun Lee. 2016. Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. *Vision Research* 118 (2016), 48–59. https://doi.org/10.1016/j.visres.2014.12.018

Alejandro Clemotte, Miguel Ángel Velasco, Diego Torricelli, Rafael Raya, and R. Ceres. 2014. Accuracy and Precision of the Tobii X2-30 Eye-tracking under Non Ideal Conditions. In *Proceedings of the 2nd International Congress on Neurotechnology, Electronics and Informatics*. https://doi.org/10.5220/0005094201110116

H. Collewijn, F. van der Mark, and T. C. Jansen. 1975. Precise recording of human eye movements. *Vision Research* 15, 3 (1975). https://doi.org/10.1016/0042-6989(75)90098-X

Kirsten A. Dalrymple, Marie D. Manner, Katherine A. Harmelink, Elayne P. Teska, and Jed T. Elison. 2018. An Examination of Recording Accuracy and Precision From Eye Tracking Data From Toddlerhood to Adulthood. *Frontiers in Psychology* 9 (2018). https://doi.org/10.3389/fpsyg.2018.00803

Jan Drewes, Guillaume S. Masson, and Anna Montagnini. 2012. Shifts in reported gaze position due to changes in pupil size: ground truth and compensation. In *Proceedings of the Symposium on Eye Tracking Research and Applications. ETRA '12*. 209–212. https://doi.org/10.1145/2168556.2168596

Jan Drewes, Weina Zhu, Yingzhou Hu, and Xintian Hu. 2014. Smaller Is Better: Drift in Gaze Measurements due to Pupil Dynamics. *PLoS ONE* 9, 10 (2014). https://doi.org/10.1371/journal.pone.0111197

Roy S. Hessels and Ignace T. C. Hooge. 2019. Eye tracking in developmental cognitive neuroscience–The good, the bad and the ugly. *Developmental cognitive neuroscience* 40 (2019). https://doi.org/10.1016/j.dcn.2019.100710

Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA'12*. 45–52. https://doi.org/10.1145/2168556.2168563

Won June Lee, Ji Hong Kim, Yong Un Shin, Sunjin Hwang, and Han Woong Lim. 2019. Differences in eye movement range based on age and gaze direction. *Eye* 33, 7 (2019), 1145—-1151. https://doi.org/10.1038/s41433-019-0376-4

Dillon J. Lohr, Lee Friedman, and Oleg V. Komogortsev. 2019. Evaluating the Data Quality of Eye Tracking Signals from a Virtual Reality System: Case Study using SMI's Eye-Tracking HTC Vive. arXiv:1912.02083 [cs.HC]

Jonathan Martschinke, Jana Martschinke, Marc Stamminger, and Frank Bauer. 2019. Gaze-Dependent Distortion Correction for Thick Lenses in HMDs. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. https://doi.org/10.1109/VR.2019.8798107

Michael B. McCamy, Ali Najafian Jazi, Jorge Otero-Millan, Stephen L. Macknik, and Susana Martinez-Conde. 2013. The effects of fixation target size and luminance on microsaccades and square-wave jerks. *PeerJ* 1 (2013), e9. https://doi.org/10.7717/peerj.9

James Morgante, Rahman Zolfaghari, and Scott P. Johnson. 2011. A Critical Test of Temporal and Spatial Accuracy of the Tobii T60XL Eye Tracker. *Infancy* 17, 1 (2011), 9–32. https://doi.org/10.1111/j.1532-7078.2011.00089.x

Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior research methods* 45, 1 (2013), 272—-288. https://doi.org/10.3758/s13428-012-0247-4

Eyal M. Reingold. 2014. Eye tracking research and technology: Towards objective measurement of data quality. *Visual Cognition* 22, 3-4 (2014), 635–652. https://doi.org/10.1080/13506285.2013.876481

Irati R. Saez de Urabain, Mark H. Johnson, and Tim J. Smith. 2015. GraFIX: A semiautomatic approach for parsing low- and high-quality eye-tracking data. *Behavior research methods* 47, 1 (2015), 53–72. https://doi.org/10.3758/s13428-014-0456-0

Christy K. Sheehy, Pavan Tiruveedhula, Ramkumar Sabesan, and Austin Roorda. 2015. Active eye-tracking for an adaptive optics scanning laser ophthalmoscope. *Biomedical Optics Express* 6, 7 (2015), 2412. https://doi.org/10.1364/BOE.6.002412

Tobii Technology. 2012. *Accuracy and precision test method for remote eye trackers*. Retrieved November 10, 2020 from https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/use/what-affects-the-performance-of-an-eye-tracker/tobii-test-specifications-accuracy-and-precision-test-method.pdf/?v=2.1.1

Tobii Technology. 2017. *Participant management & recruitment*. Retrieved February 24, 2021 from https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/design/participant-management-and-recruitment/

Antonia Vehlen, Ines Spenthof, Daniel Tönsing, Markus Heinrichs, and Gregor Domes. 2021. Evaluation of an eye tracking setup for studying visual attention in face-to-face conversations. *Sci Rep* 11, 2661 (2021). https://doi.org/10.1038/s41598-021-81987-x

Yabo Yang, Keith Thompson, and Stephen A. Burns. 2010. Pupil Location under Mesopic, Photopic, and Pharmacologically Dilated Conditions. *Invest Ophthalmol Vis Sci.* 43, 7 (2010), 2508—-2512.

Shlomit Yuval-Greenberg, Elisha P. Merriam, and David J. Heeger. 2014. Spontaneous Microsaccades Reflect Shifts in Covert Attention. *Journal of Neuroscience* 34, 41 (2014), 13693–13700. https://doi.org/10.1523/JNEUROSCI.0582-14.2014