https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-129

Defining gaze tracking metrics by observing a growing divide between 2D and 3D tracking

William Andrew Blakey[1,2], Stamos Katsigiannis[1], Navid Hajimirza[2], Naeem Ramzan[1];
[1]School of Computing, Engineering and Physical Sciences, University of the West of Scotland; Paisley, UK;
[2]Lumen Research Ltd; London, UK;

Abstract

This work examines the different terminology used for defining gaze tracking technology and explores the different methodologies used for describing their respective accuracy. Through a comparative study of different gaze tracking technologies, such as infrared and webcam-based, and utilising a variety of accuracy metrics, this work shows how the reported accuracy can be misleading. The lack of intersection points between the gaze vectors of different eyes (also known as convergence points) in definitions has a huge impact on accuracy measures and directly impacts the robustness of any accuracy measuring methodology. Different accuracy metrics and tracking definitions have been collected and tabulated to more formally demonstrate the divide in definitions.

Introduction

At present the gaze tracking community uses a variety of metrics for assessing gaze tracking accuracy. The biggest limitations of having a variety of non-standardised metrics in this field is that comparison between tracking algorithms becomes challenging and therefore it is hard to determine the best performing approaches. There is benefit from dividing what could be considered eye tracking to gaze tracking in 2D or 3D. As a result, understanding the accuracy of eye tracking should be different to that of gaze tracking. Another factor would be appreciating the difference between tracking gazes in a 3D world and in a 2D plane. To establish suitable accuracy metrics, the ease of understanding, the accessibility of data and most importantly what is the subject of the measurement should all be considered. The aim of this work is to use experimental data to demonstrate the difficulties in comparing the performance of different eye/gaze trackers. Due to the limitation uncovered; this paper suggests new definitions for trackers that aim to clear up differences in accuracy metrics.

The accuracy approaches vary dramatically as hardware are inconsistent, accuracy measures can add error or boost results by adding information that isn't present in the tracker and with little regard to other statistical measures or informing readers of potential cause for errors. Kar et al. [1] demonstrated the difficulties in comparing the different approaches by establishing a list of different trackers and the variety of metrics they use. The authors of that work have additionally created tools to help enable researchers in this field to compare and understand their trackers in relation to others [2]. Although this seems incredibly valuable, their proposed re-definitions and clarity in accuracy measures restrict the usage of such a tool as these follow the same limiting factors uncovered in the literature review. It is the opinion of the authors of this paper that the accuracy of errors for gaze tracking needs to be reviewed, and a scope and understanding of the branching field between 2D and 3D gazes that is understood in many papers [3, 4, 5, 6] be redefined with more appropriate accuracy measures for their specific context.

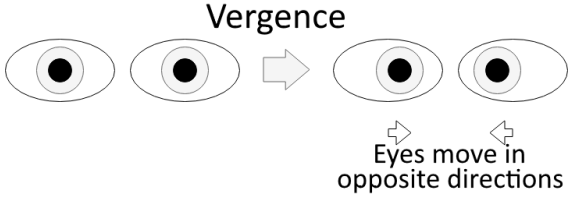


Figure 1: Eye vergence.

Although infrared trackers can be highly accurate, their accuracy is dependent on what type of gaze tracking is being observed. There is a fundamental difference between measuring accuracy in a plane and in 3D space. It is essential that this difference is observed, as incorrect interpretation can lead to misleading conclusions and incorrect accuracy results. Benchmarking data sets allow comparative studies to be conducted using consistent testing data. The EYEDIAP [7] dataset was an important step forward through its understanding of 3D tracking and specifically creating open source comparative data for both 3D and 2D tasks [7], as well as through making a clear distinction between 2D and 3D gaze accuracy in a similar fashion to the processes described in this work. Other datasets include the MPIIGaze dataset [8], which prioritises images for appearance-based methods specifically, with the purpose of estimating gazes in real world contexts, and the Columbia dataset [9], which includes a large number of gaze tracking videos captured using a head rest.

This work examines the different terminology used for defining gaze tracking technology and explores the different methodologies used for describing their respective accuracy through an experimental evaluation using infrared and webcam-based trackers. Based on the experimental results, this work attempts to redefine terms and demonstrate the flawed approaches in perceiving gaze tracking accuracy.

State-of-the-art review

The term "convergence" is one that this paper aims to cover and how it impacts accuracy. The fundamental understanding is based on the definition of eye movements; one of which is called eye vergence [10]. Eye vergence is something that specifically happens when both eyes focus on an object and because both eyes are in slightly different places relative to the object being looked at, they both move slightly differently. Figure 1 describes moving the eyes in the opposite direction for the purpose of focusing on items in a 3D world for binocular vision. Tracking this movement

Table 1: Accuracy metrics from Tobii's specification

Accuracy Metrics	Descriptions	Motive
Monocular Accuracy	Accuracy on one eye for each participant	To illustrate the performance of one eye
Binocular Accuracy	Accuracy as the arithmetic mean of the two eyes	The accuracy as perceived in most user situation

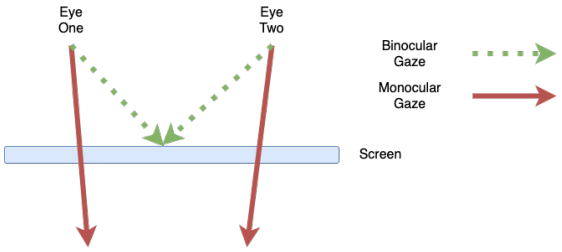


Figure 2: Binocular and monocular gaze.

has most use in 3D gaze applications, as it can establish depth, and it can be said that the gaze of two eyes will converge on real world objects limiting the practical limitations of finding two vectors. From the perspective of eye-gaze vectors and the concept of 3D vision, the vectors from the eyes will intersect at the location of the object the individual is looking at. This intersection of eye vectors is the convergence point. Binocular vision requires two monocular images to converge and this conveys depth [11].

Tobii® is one of the largest gaze tracking companies at the time of writing this article. They have several hardware approaches to gaze tracking: infrared and glasses. Their technology specification [12] for their gaze tracker contains definitions for monocular and binocular gaze, and methods for calculating accuracy, as shown in Table 1 and Figure 2. The technology specification goes on to explain that the accuracy being measured is the gaze angle. The pixel accuracy is then used to calculate the angle from the average intersection point of the two monocular gazes. The difference between what is described as monocular and binocular gaze is the assumed convergence of the vectors. The vectors must intersect for binocular vision and this is the case for "most user situations" that Tobii® refers to [12]. The EYEDIAP dataset, as described in [7], explores different situations for gaze trackers. One experiment has the user look at a floating ball and this experiment is one where the Tobii® definitions do not seem enough. The approach of looking at a random point in 3D space and calculating the Tobii®'s definition of binocular accuracy to that point is challenging.

Figure 3 demonstrates how the average of the gaze vectors' intersection with the screen could cause the "better" red vectors to become worse when the user looks at the orange ball. The predominant reason for this is because Tobii®'s accuracy measures describe binocular accuracy on a screen, which has only been proven true when the user is looking at the screen. Looking at the screen means the point for which eye vectors intersect with each other is also when they intersect with the screen. Another way of describing this would be that the convergence point is on the screen.

Tobii®'s definition for binocular accuracy seems likely to only fit when the tracker is on the screen. An assumption of con-

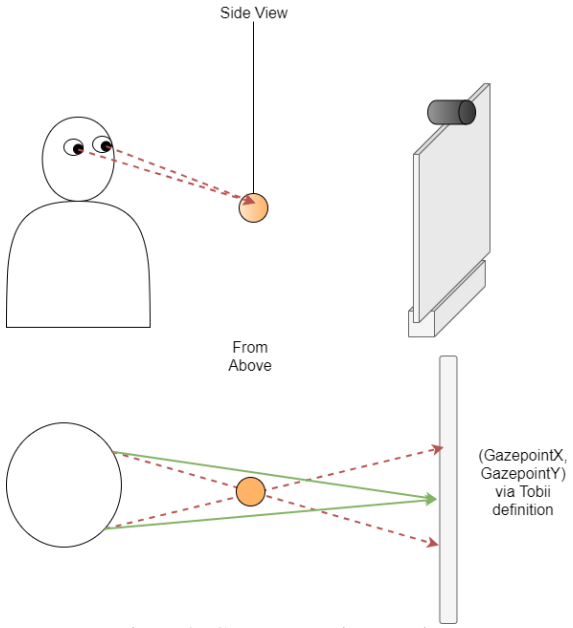


Figure 3: Gaze vectors intersection.

vergence might benefit from better accuracy than the monocular gaze in most cases. When looking around the world, the gaze vectors of an individual's eyes will intersect as described by the definitions for binocular vision in most humans and so another suggestion for binocular error could be the estimated error between the true convergence point and the estimated one. This would require the convergence point to be estimated. In the context of the plane, this could be the average intersection point of the plane with the two gaze vectors and where the user is looking on the screen, which matches the Tobii® definition.

There is a growing need for binocular tracking in applications that fall outside of a 2D plane. Market research has lots of applications, e.g. [13] which shows that there is a huge impact in in-store design and packaging because of attention. The described accuracy methods need to be further explored for gaze tracking glasses as this could impact their use in research. In the next section infrared trackers will be used to explore their potential use in 3D applications. Technology has changed significantly and the use of convergence in definitions of binocular and monocular gaze needs to be explored. Virtual Reality (VR) headsets with built-in eye tracking are being used in market research and to understand where someone is looking requires depth.

The concept of predicting horizontal vergence was reviewed in [14]. That work acknowledges the concept of binocular vision in gaze trackers and questions their accuracy. There is an understanding that when the user focuses on something in front of the screen or behind, then this results in a convergence point that is not on the screen. They explore vergence by using a Pupil Cornea Centre Reflection (PCCR) algorithm where altering the screen colour alters the pupil centre estimation because the change in light from the screen background alters the size of the pupil and therefore alters the estimation. The algorithm used for pupil centre detection can alter the accuracy of the monocular gaze with light change, and this therefore impacts the estimated horizontal convergence. They do so by utilising an SR EyeLink 1000 [15] tracker.

Table 2: Prior art Webcam-based gaze tracking algorithms and their declared accuracy

Approach	Declared Accuracy in Angular Error	Possible Limitations	Prediction and model	Monocular/ Binocular
Webgazer [16]	104 pixels (ideal) 4.16°	Assumed convergence on screen, Mouse aid	Screen gaze location, Ridge Regression	Binocular on screen
Feng Lu et al. [17]	0.62° (with head clamp)	Assumed convergence on screen	Screen gaze location, Adaptive Linear Regression	Binocular on screen
TurkerGaze [18]	1.06°	Assumed convergence on screen	Screen gaze location, Ridge Regression	Binocular on screen
CNN-based [19]	3.65°	Assumed convergence on screen	Screen Gaze Location, CNN	Binocular on screen

As stated previously another huge divide in the field is with the type of predictive model. What is being modelled is a required understanding for prediction. When judging accuracy, should the algorithm be assessed to adhere to prior declared accuracy, or would they be better served judging the accuracy of the prediction. The fundamental aspect that causes this divide relates to whether the model predicts an angle or the end point of the gaze. Table 2 depicts a range of prior art webcam-based gaze tracking algorithms, the hardware they work with, and the accuracy they declared they have achieved along with possible limitations of that accuracy measurement. The algorithms in Table 2 refer to methodologies where the prediction is a screen gaze location. However, what is described and judged as the accuracy is an angular error. The reasons most of these works suggest the use of angular error is to adhere to previous algorithms. This raises the question of whether there is a difference between when the accuracy of the algorithm should be based on what is being predicted, e.g. the gaze, or what applying the gaze to the eye to gain an angular error for comparison. With the concept of binocular and monocular, this problem becomes more apparent.

Understanding limitations through experimental data

There are clear limitations when trying to compare trackers. It is difficult to compare angular error with screen distance error. The very act of measuring tracker accuracy can artificially enhance results. The fundamental reason for this is that eye vectors intersect, through an eye movement called “convergence”. For the purpose of this paper the intersection point of the two eye vectors will be called the convergence point. These limitations are enhanced when screens are included in the result. It is important to understand that removing the need to calculate the depth of the convergence point by assuming convergence on the plane will change the accuracy measure. The following experiment should clarify these points and demonstrate the pitfalls and the difficulties in different accuracy methodologies.

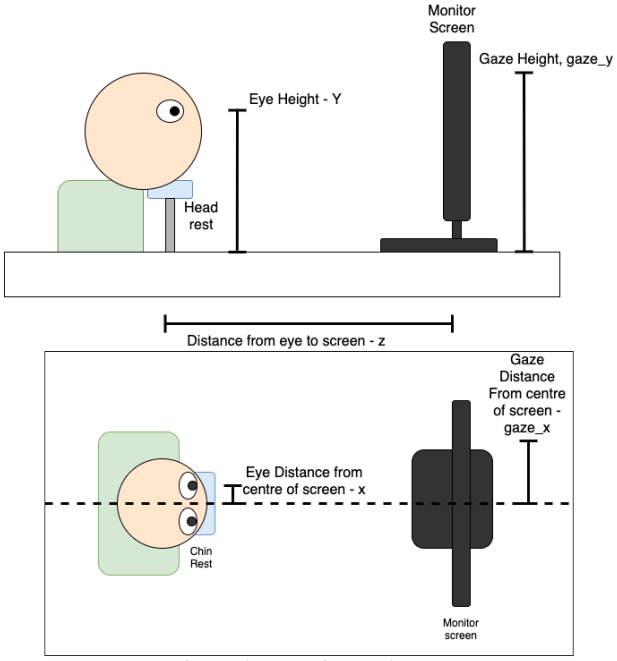


Figure 4: Experimental setup.

Methodology

The accuracy measures will be angular error, distance from the ground truth gaze point on plane and average angular error to the screen gaze point. In order to perform this experiment, a tracker needs to be used where each eye can be assessed individually and a distance measure to the eye is calculated. In this experiment a Tobii Infrared X2 tracker was used and a webcam-based approach provided by Lumen Research [20]. Additionally we have tested the same experiment using a mobile phone based webcam tracker also supplied by Lumen Research [20]. This mobile phone tracker was on an iPhone XR and was placed slightly closer to the participants than the other trackers, due to the need simulate common usage and make it easier for participants to interact with the device as would be common for mobile phones. The phone was placed at a distance of 38.3 cm compared to 53 cm in the other cases. A small number of participants were chosen due to the experiment being demonstrative for the different methodologies. Fifteen subjects participated in this study, 10 male and 5 female, with their age ranging from 20 - 60 years old, and including a diverse background of ethnicity.

Although infrared trackers are capable of being used with free head movement, for the purposes of the experiment and ease of calculating accuracy (by making the distance to the camera more constant) a head rest was used. The participants were kept in isolation and brought into the experiment room individually to avoid bias being introduced. They were then calibrated with the tracker while the head was placed in the head rest. The participants were shown a series of dots and the average gaze intersection per eye was recorded along with estimations for distance to the camera per eye. Ground truth data was created by showing the participant points on the screen to look at. To get the data and calculate the true vectors, the assumption will be that the participant gaze vectors intersect with the point shown.

Figure 4 demonstrates the setup with relation to the screen

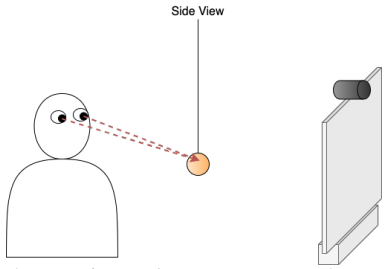


Figure 5: Experimental setup to capture 3D accuracy.

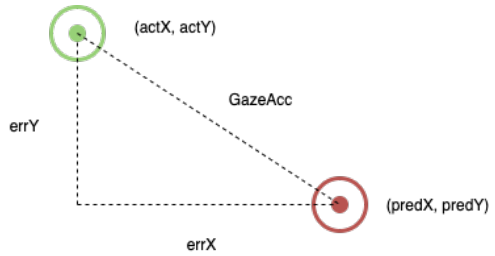


Figure 6: On-screen gaze error.

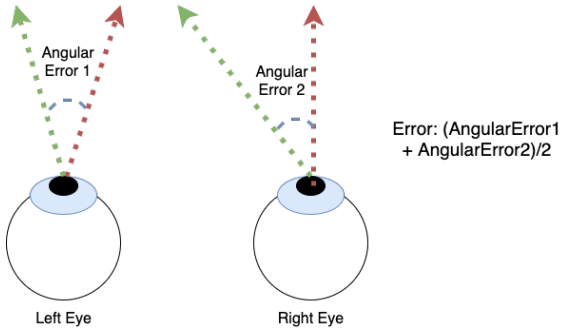


Figure 7: Angular error of eye.

where the targets were shown. This was set up so that each user would be experiencing as similar conditions as possible, with the distance to the screen being kept consistent and the height of the eye being kept the same from participant to participant due to the ability to move the chin rest up and down. The participants were assumed to have the average pupillary distance. Knowing the eye location in 3D and the distance to the targets on the screen enables the angular error for each participant to be calculated. The screen location of the participants were converted to cm by knowing the measurements. The diagram in Figure 5 demonstrates the setup designed to capture 3D accuracy. An orange ball was suspended with a string from a metal arm just off the middle of the setup. In order to keep the ball from moving from participant to participant a weight was attached via a string to the bottom.

Accuracy Measures

In order to demonstrate the accuracy of a variety of tracking methodologies, a variety of different accuracy measurements will be used. The main two forms of metrics for measuring gaze tracking accuracy is angle and distance of the predicted gaze point:

Monocular approaches

These approaches find an error and then calculate the average for both eyes, rather than finding the average gaze point. These approaches can contain both eyes.

On-screen gaze error (cm): Figure 6 and the following equation describe the methodology for calculating the on-screen gaze error:

$$GazeAcc = \sqrt{errX^2 + errY^2} \quad (1)$$

where $errX = |actX - predX|$ and $errY = |actY - predY|$. This measure is the calculated distance of the onscreen predicted gaze point against the ground truth gaze point (the actual location where the person is looking). In order to find the gaze accuracy for both eyes, the difference needs to be calculated first and then averaged to work out the monocular accuracy. This method calculates the accuracy of the gaze point assuming that the gaze has converged on the screen. As can be seen in the diagram in Figure 6, the eye vectors intersect with the screen and the two gaze points are averaged into one gaze point. This averaging into one is the assumed convergence on the screen. The accuracy of this single gaze point is then used to calculate the error. This approach averages and then calculates error on the average point. The average distance in cm is then calculated.

Angular error of eye (monocular accuracy): Angular error of the eye is the angular difference between the true eye direction and the estimated eye direction. This is calculated per eye and is then averaged as shown in Figure 7. This is also referred to as monocular accuracy. The concept of taking the angular error is not simple. In order to acquire ground truth data, the actual eye angle must be calculated, and this requires the eye to be localised in 3D. Infrared trackers calculate distance via the size of the reflection in the eye. This measurement is an estimation and is an aspect of the tracker. Calculating the accuracy by utilising an aspect of the tracker's algorithm seems flawed. The approach needs to be measuring where the eye is in three dimensions by measuring each dimension individually. Using a head rest the head can be kept still, but even then, the head is not perfectly still and this error in slight head movements needs to be calculated and included in the results. There is also error in this approach, e.g. the precision of the distance measurements. The distance to the head mount (z distance) can be calculated to the closest mm and the height of the eye in relation to the desk (y distance) can also be calculated when an individual has their head placed in the rest. In order to calculate how far the eye is in relation to the head rest edge (x distance) an image will be taken, and this will be calculated by understanding the distance to eye centre to the edge of the head rest in the image with respect to the width of the head rest in the image and the width of the head rest in mm. Using these measurements, the eye angle can be calculated with respect to its default direction (looking forward). However, the angular error can be calculated by finding the vectors of the true gaze and the predicted gaze and finding the smallest angular error assuming the same starting point (the centre of the eye in 3D). The predicted gaze vector is calculated by finding the single predicted eye gaze point. This is then calculated for the other eye and the two angles are averaged, as shown in Figure 8.

Binocular approaches

These approaches introduce convergence points and have to include both eyes. The intersection of the two gaze vectors is the convergence point. This point is then used to find the error. In this sense the binocular approach averages the gaze points and then finds the error as opposed to the monocular approach where the averaging is done to the calculated error.

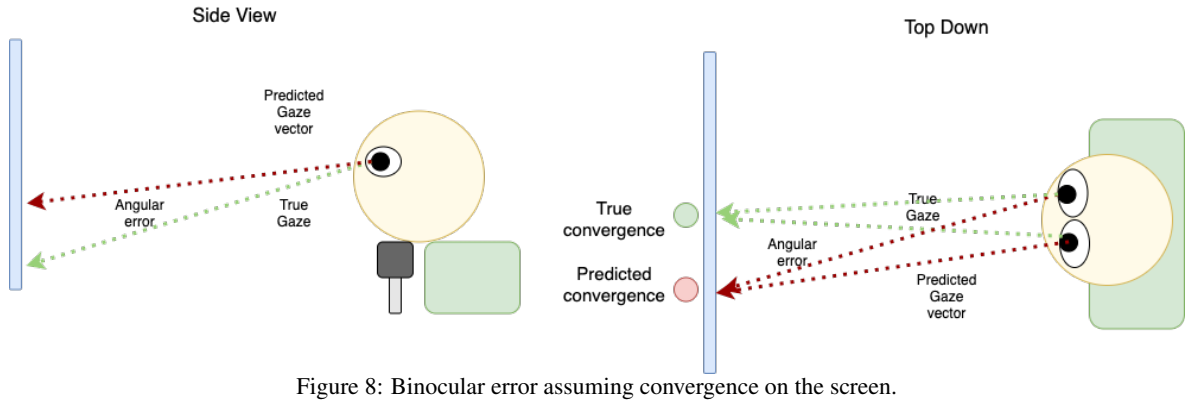


Figure 8: Binocular error assuming convergence on the screen.

On-screen gaze error (cm): Some tracking algorithms work by predicting only one point from the two eye models. In this case this is the convergence point. In the case where each eye is modelled separately the output of the model is two on-screen locations for gazes. In the case where each is modelled individually the convergence is assumed to be on the screen and therefore the actual gaze point is the average location of the two predictions. In this case, the gaze points are averaged and then the distance from the average gaze point to the ground truth (the points where the user is looking) is the error.

Angular error to the assumed convergence point on screen: The gaze point on the screen for each eye is computed and then a single angular error to this single gaze point is computed as described in [21] and as above. Angular error has to be the difference between a predicted value and a ground truth. The only consistent method for computing this error is the average angular error to the average gaze point, which is calculated as above. The reason for this is that the average gaze point is a methodology for getting the convergence points and then the average angular error per eye is the accuracy to the estimated convergence which is described in more detail below.

Angular error to an estimated convergence point (off screen, most likely): In eye tracking, there is not always a convergence point, as the eye tracking vectors may not intersect. In order to consider an eye tracking algorithm as a 3D gaze tracking method, an intersection point for the vectors needs to be established. Binocular vision conveys an understanding of depth to the brain, predicting gaze in three dimensions and requiring this understanding of where the eye is converging. In the example in Figure 9, the predicted vectors do not intersect and so there is no convergence. There are methods to compensate for error in eye tracking. For instance one suggestion could be that the gaze tracker can establish the shortest distance between the two vectors and assume that it is the convergence point and therefore adjust the predicted vectors from the eye. Another suggestion could be that the predicted vectors from each eye match in relation to up and down movements. That is to say that when the head is upright, the vertical movements of the eye are the same. This does not consider roll or yaw of the head and therefore requires head pose understanding to assume that the head is vertical or that the matching vectors are related to the roll and yaw. Additionally, another possible method for convergence estimation is to assume that the convergence points match the closest object to the predicted convergence point. The approach for finding the angular

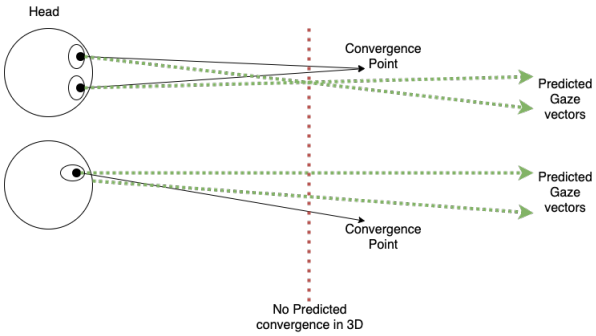


Figure 9: Predicted gazes with no convergence point.

error to the convergence point will be done via finding the convergence point by finding the point on each eye vector where the distance is shortest and moving the vectors to converge at the midpoint of the line of the shortest distance. Once the convergence point is found, the vectors for the left and right eye can be calculated. This can then be compared to the true left and right vectors, assuming the participants were looking at the point shown, and the angular error calculated per eye and then averaged.

Results

Figure 10 shows the mean binocular results and their range, assuming convergence on screen, for the 15 subjects that participated in this study using the three examined trackers. Figure 10a depicts the angular error to the assumed convergence point on the screen, while Figure 10b depicts the gaze error of the onscreen convergence point. It must be noted that none of the participants were dropped due to only presenting the best results. If the same procedure was followed, as was suggested in the Tobii® report [12], where only 20% of the participants were declared in the final measurement, very similar results of 0.4° would be acquired. Additionally all these results were acquired through using a head rest which is required for calculating the angular error. The results are likely to change if the head rest is not used, as the head would move, thus having an additional variable adding error to the estimation.

It is quite clear that from an historical point of view, angular error was a useful measurement when understanding eye movements. To gain an understanding of saccades and fixations, researchers used approaches where they were directly measuring the movement of the eye. Nowadays, the measurements are focused on where the eye is looking, the gaze. What is predicted

Table 3: Angular error for a convergence point assuming and without assuming convergence on the screen

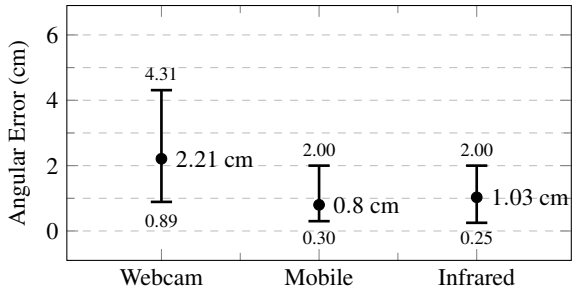
Accuracy metric	Value
Monocular eye error	1.3°
3D Gaze Tracking to estimated convergence	1.3°
2D gaze tracking, assumed convergence on the screen	1.0°

using PCCR or regression approaches is a gaze. Measuring the angular error means converting the predicted gaze to an estimated eye vector which simply adds more elements that can introduce error. The use case of Tobii® trackers and webcam-based approaches focus on the gaze and utilising that gaze to control an OS or for market research. This is a 2D screen location and is predicted on the plane so the accuracy of that location is in metres. Figure 10 also demonstrates the gaze estimation. This is not to say there aren't uses for 3D gaze tracking or eye tracking, such as research or VR and Augmented Reality (AR).

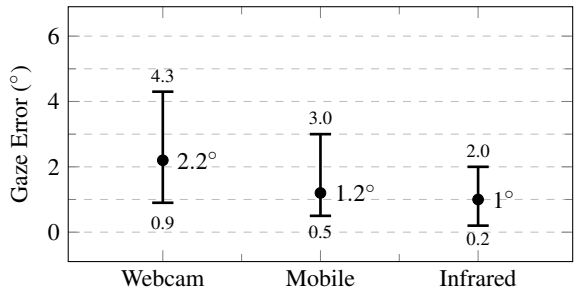
Table 3 demonstrates the angular error for a convergence point assuming convergence on the screen and without, when using the infrared tracker. As can be seen, all the previously described accuracy measures provide different results. This demonstrates the vast difference between methodologies and how a clear split needs to be established for Eye Tracking, 3D gaze tracking and 2D gaze tracking. The problem of calculating the location of the convergence is tough. The results in Table 3 demonstrate how the infrared tracker has a 1.3° error of estimating the eye vector correctly to an estimated convergence and a 1° error when the convergence is assumed to be on the screen, while there is a 1.03 cm (Figure 10a) distance error between the estimated gaze point on the screen to the actual location of the gaze. These are fundamentally different problems and there needs to be a clear divide in how researchers demonstrate the accuracy of their tracker. Suggesting an infrared tracker has a 1° degree error is misleading as it only has this error when the tracker assumes the user is looking at the screen. Given this shows a clear split between 2D and 3D tracking, the application is fundamentally important. If a tracker is most interested in the screen accuracy, then the distance error is a more suitable measure, especially when considering that all trackers predict a gaze point whereas not all trackers predict an angle.

It seems fairly clear that the difference between the monocular accuracy (the average of the two individual eye accuracies) and the binocular accuracy to the estimated convergence are similar. The larger difference is in demonstrating the difference between estimated convergence and screen accuracy. This result specifically implies that the usage of the declared accuracy results for infrared trackers only fits use cases such as market research or controlling an OS. In the examples of AR and VR, the monocular accuracy or the estimated convergence is more fitting. In these cases the user has to understand whether the desired level of accuracy is met.

Most infrared trackers are used on a 2D plane and can be considered as 2D trackers. The acquired results demonstrate how calculating angular error requires additional estimates beyond the actual estimated result to be included. When considering these results for 3D tracking or eye tracking, these estimates for eye



(a) Angular error to the assumed convergence point on the screen



(b) Gaze error of the on-screen convergence point

Figure 10: Mean and range of binocular results, assuming convergence on screen.

location (distance to camera) are required to calculate the convergence point or to calculate the eye direction. They are an aspect of the algorithm and thus judging the algorithm's success based on this is different to tracking gaze on a plane as this assumes convergence on the screen and this distance is known.

The goal of considering degrees of accuracy as a measure of accuracy for gaze tracking is predominantly because it is distance invariant, meaning that however far from the camera the user is, the accuracy would remain consistent. For infrared trackers this is not true. The accuracy will change depending on the distance to the camera because the resolution of the eye image in the camera plane is the main factor that will affect the accuracy when the user gets further from the screen plane. The goal for using degrees is flawed and therefore all degree measurements need to state that the accuracy was calculated at a particular distance.

Considering a gaze on 2D plane and on 3D space require very different approaches. When considering a predicted gaze vector from an eye, the tracker needs to distinguish how far the prediction is from the true vector. Gauging this by forcing a participant to look at a specific point on a screen introduces difficulties as convergence movements need to be considered as there is a fundamental truth that normal eyes will converge on to a point. If the prediction requires the participant to be looking at a point on the screen, then measuring accuracy in degrees assumes that they are converging on the screen and a specific point. Rather than accounting for the actual difference between the perceived vectors, the method assumes convergence to the screen and filters accordingly, which will allow for false measurements of how accurately a tracker is working in 3D, as working out the specific point of convergence is a fundamental aspect of 3D gaze tracking. Considering 3D accuracy by assuming convergence on the screen adds a ground truth that is not present in the prediction and can artificially improve the perceived accuracy, rather than the true

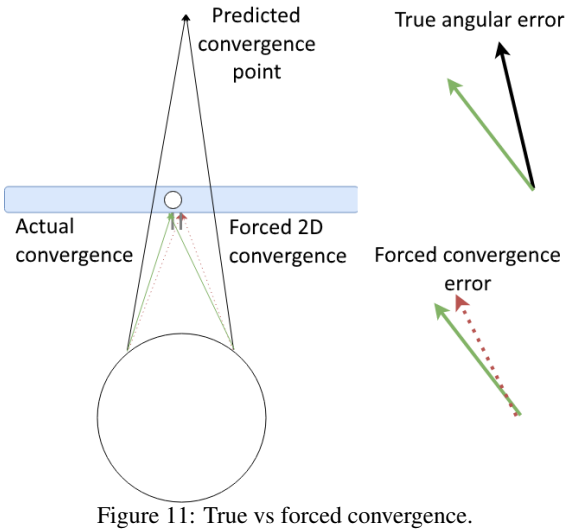


Figure 11: True vs forced convergence.

accuracy of said trackers. As a result, the act of measuring has inflated the accuracy. This demonstrates that, as a 3D tracker, most results are false, and that as a 2D tracker, the results are almost incomparable (Figure 11).

That is not to say that as 2D trackers the predictor and accuracy measure used are not accurate. In fact, they are exactly as accurate as claimed as a 2D gaze. When considering 3D gaze trackers on a 2D plane they are acting fundamentally as a 2D gaze tracker. If the main use of 2D trackers relies on screen accuracy within that plane, then a better measure is in metres to the specific object. Pixels as a measure is flawed fundamentally as TVs, laptops, monitors, phones and tablets can all have different resolutions. For example, 100 pixels on one screen can be 2 cm, which could be 5 cm on another screen. With gaze tracking as an industry growing to sizeable amounts, the way in which accuracy is measured needs to become consistent and a clear divide needs to be established between 2D tracking and 3D.

Definitions

There is a need for more clear definitions as described in an increasing amount of papers and is demonstrated by the above experiment. This section provides suggestions of gaze/eye trackers definitions with the aim of making comparisons easier.

Eye Tracking (very similar to monocular gaze but where the estimation is on the eye rather than the gaze)

Eye tracking: Direction vector and location of an individual eye. Eye tracking is understanding how the eye moves. Historically, invasive approaches were used to better understand eye movements [22]. Currently, models can be created to predict how the eye is moved based on a variety of stimuli, such as infrared light reflection in relation to eye centres (PCCR, pupil cornea centre reflection) [23]. These methods track eye movement regardless of context, head movements and target of gaze and as such, the most important factor is the movement of each eye. When tracking an eye and specifically its movements, the accuracy of such algorithms can be judged as angular error. In tracking eyes movements, an aspect of these algorithms is understanding the directions of the eye. From eye tracking, a gaze can be calculated, but this requires the eye to be tracked in relation to the head and modelled in the world. Most gaze tracking applications are confused as eye tracking, but eye tracking should fundamentally be understanding the eye and its architecture to ascertain direction and location of the eye in an image. Eye tracking accuracy considers each eye individually in terms of accuracy, meaning that for a person this would be averaged over both eyes.

Eye tracking should use angular error per eye as an accuracy measure.

Gaze tracking

Gaze tracking is understanding where the eye is looking. This is different when considering gazes on a plane and gazes in 3D space because there is a different way in which the eye behaves when viewing an object in 3D space to 2D. There must be an understanding of what is being viewed rather than just the direction of the eye. The direction of the eye does give an indicator, but it is also important to understand the distance to the object. It is not always apparent the need for the understanding of convergence, but situations occur where the predicted gaze could fall between several objects and the understanding of distance can be vital in differentiating.

3D Gaze Tracking (always binocular, has to include convergence point)

3D Gaze tracking: Direction vectors and location of both eyes, along with an intersection of the vectors known as the convergence point. Fundamentally 3D gaze tracking and eye tracking are similar apart from the fact that 3D gaze trackers require both eyes and an intersection point for the gaze vectors. What is being looked at could be an object, but does not necessarily need to be, as people can stare into space. The concept of eye convergence needs to be considered. Eye tracking should be measured in degrees as described above. 3D gaze tracking considers a 3D vector where the important details are error in angle and the convergence point very similar to eye tracking but without the convergence estimation.

3D Gaze Tracking should use average angular error (of each eye) to estimated convergence point as an accuracy measure.

2D Gaze Tracking (can be monocular or binocular depending on if averaged)

2D Gaze tracking: The average point of intersection between gaze vectors (of both eyes) and a plane (a screen). One of the most common uses for eye tracking is tracking an individual's gaze on a 2D plane. In contrast to 3D tracking, there can be the assumption that the gaze converges on the screen plane. This added detail is powerful when it comes to tracking accuracy, as demonstrated in the experimental results and therefore needs to be considered separately to above examples in 3D. Most 2D gaze tracking applications don't require an eye tracking step whereas by the very nature of having 3D vectors the eye will be tracked.

2D gaze trackers may not consider an eye model. Additionally, the error that is most important is in the screen plane. Fundamentally, the algorithms estimate the gaze (what is being looked at) rather than eye movements and therefore the error in the plane needs to be the error that is compared from tracker to tracker as this is the only accuracy measure that will be consistent from all tracking algorithms. The limiting factor when considering dis-

tance is the resolution of the camera detecting the eye and how much detail is required in the tracking technique and so considering angular error as one that is invariant with distance. The only method that is a fair assessment of 2D gaze tracking accuracy is in metres and is also the only method that can state how effective different gaze tracking algorithms are. It is also a fair assessment that a rough guide of how close the user was to the screen plane is present with the accuracy measure or an assumption is made that it is for the optimal distance for the algorithm.

2D Gaze Tracking should use screen distance error as an accuracy measure.

Fitting the new definitions to current trackers

For the purposes of 2D gaze it is important to appreciate the different methods that researchers have used to build a predictive model. Generally, there have been two approaches: (i) appearance-based, where the value for the pixels of the eye form the input, and (ii) geometric, where the geometry of the eye is considered (whether that be as a sphere or as a pupil location). There can also be hybrid approaches, as pupil location can be calculated through appearance-based methodologies. The reason it is important to consider the algorithm is because appearance-based approaches may not consider an eye geometrically and is therefore not eye tracking. In that case, there is no angular error. There is only error in the gazes that are estimated. Adhering these methods to angular error will add error, as the eye needs to be detected, tracked in three dimensions with respect to the screen, and then the angle between the screen and the eye needs to be calculated. This is not a trivial problem and is part of the reason why there are no perfectly accurate gaze trackers. Forcing 2D gaze researchers to use these metrics adds error and has limited the ease in which researchers can contribute. Consequently, the screen distance error in metres would be the most appropriate measure.

A depth camera (such as RGBD or infra-red methods) has error when calculating the distance of the user to the screen. If the distance is calculated from a camera, a universal approach of where the distance is measured needs to be considered, specifically for the purpose of gaze tracking. This distance to the eye is the important measurement for the purposes of calculating accuracy. The head (when considering free head tracking) is not stable and so, for methods that don't consider distance, they have to add a depth sensor or allow the head wobble, drastically affecting the error in the accuracy measurement (making the comparison between trackers limited).

As an industry standard, infrared trackers have been considered to have a degree of error from the eye. This measure of degrees of error fundamentally assumes that the eye is the focus of the measurement and it is therefore the eye and not the gaze that is being estimated in the gaze model. When considering an infrared tracker, the error that needs to be observed is the relationship between the eye centre and reflection in the camera plane to the screen to camera in the screen plane. This relationship also considers the distance of the head to the plane by calculating the distance to the eye via the size of the reflection. Ultimately, what is being tracked is a gaze estimate and not an eye estimate. In order to calculate the degree of accuracy for the eye as a measure of accuracy, the gaze that is estimated needs to be applied to the eye. Infrared trackers are fundamentally 2D gaze trackers, rather than eye trackers, and this means that for the estimation of an eye, the gaze needs to be applied to an eye model to gauge angular error. Infrared trackers should use screen distance error as they are fundamentally 2D gaze tracking algorithms.

A 3D tracker applied to a 2D screen should be measured by 2D metrics. If the screen wasn't there, the tracker would not have the same accuracy and as such the tracker only has said accuracy when it is being considered as a 2D tracker. As a 3D tracker, the knowledge that the gaze converges on the screen (something that may not be true) is artificially added. This means that evaluating a 3D tracker on a plane, using information not known to the tracker, will improve the accuracy measures but only in that specific condition.

Conclusion

It is the hope of the authors that by splitting the growing field of gaze/eye tracking technology into eye tracking, 3D gaze tracking and 2D gaze tracking, it will become easier to compare algorithms and better establish the cutting edge in the field. This experiment demonstrates how difficult it can be to measure accuracy consistently, as different algorithms have different priorities. Most gaze tracking algorithms have been developed for different purposes and it is therefore advantageous to appreciate what form of tracking their algorithm is best used for.

Modern uses for gaze tracking technology include improving advertising, gaming, or controlling computers for handicapped people. All of those algorithms are screen based and therefore they require 2D gaze trackers. It is most important to appreciate the accuracy in the screen plane and for this purpose a measure of accuracy in metres will best establish the most cutting-edge algorithms in this field. It is also the case that 2D gaze tracking cannot always be used for 3D gaze tracking or eye tracking unless additional information is collected.

3D gaze tracking and eye tracking both establish vectors. The applications of these algorithms are less obvious, e.g. real world tracking or virtual reality. These focus on eye vectors and so angular error is most important. These algorithms can easily be applied to 2D, as demonstrated by the experimental data. It is the opinion of the authors that when developing a 3D algorithm, its application as a 2D tracker is considered and the experiments should accommodate for this and both 3D and 2D accuracy is shared.

Due to the limitation uncovered, this paper suggests new definitions for trackers that aim to clear up differences in accuracy metrics:

- Eye Tracking: Direction vector and location of an individual eye (or both if averaged after calculation).
- 3D Gaze Tracking: Direction vectors and location of both eyes, along with an intersection of the vectors known as the convergence point.
- 2D Gaze Tracking: The average point of intersection between gaze vectors (of both eyes) and a plane (a screen).

References

[1] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[2] A. Kar and P. Corcoran, "Performance evaluation strategies for eye

gaze estimation systems with quantitative metrics and visualizations," *Sensors (Basel)*, vol. 18, no. 9, Sep 2018.

[3] M. Mansouryar, J. Steil, Y. Sugano, and A. Bulling, "3D gaze estimation from 2d pupil positions on monocular head-mounted eye trackers," in *Proceedings of the 9th ACM Symposium on Eye Tracking Research & Applications (ETRA)*. New York, NY, USA: Association for Computing Machinery, 2016, p. 197–200.

[4] J. Wang, G. Zhang, and J. Shi, "2D gaze estimation based on pupil-glint vector using an artificial neural network," *Applied Sciences*, vol. 6, no. 6, p. 174, Jun 2016.

[5] C. Palmero, J. Selva, M. Bagheri, M. Ca, and S. Escalera, "Recurrent CNN for 3D gaze estimation using appearance and shape cues," in *British Machine Vision Conference 2018*, Sep 2018.

[6] K. A. Funes Mora and J. Odobez, "Person independent 3D gaze estimation from remote RGB-D cameras," in *2013 IEEE International Conference on Image Processing*, Sep. 2013, pp. 2787–2791.

[7] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*. New York, NY, USA: ACM, 2014, p. 255–258.

[8] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4511–4520.

[9] B. Smith, Q. Yin, S. Feiner, and S. Nayar, "Gaze Locking: Passive Eye Contact Detection for Human? Object Interaction," in *ACM Symposium on User Interface Software and Technology (UIST)*, Oct 2013, pp. 271–280.

[10] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, Nov 1998.

[11] H. Ono, R. Angus, and P. Gregor, "Binocular single vision achieved by fusion and suppression," *Perception & Psychophysics*, vol. 21, pp. 513–521, 11 1977.

[12] *Accuracy and precision test method for remote eye trackers*, Tobii Technology, Feb. 2011, Test Specification Version: 2.1.1.

[13] R. Hurley, J. Rice, D. Cottrell, and D. Felty, "The impact of flexographic and digital printing of fruit drinks on consumer attention at the point of sale," *Beverages*, vol. 1, no. 3, p. 149–158, Jul 2015.

[14] I. T. Hooge, R. S. Hessels, and M. Nyström, "Do pupil-based binocular video eye trackers reliably measure vergence?" *Vision Research*, vol. 156, pp. 1 – 9, 2019.

[15] *EyeLink® 1000 Plus Technical Specifications*, SR Research, Feb. 2017, Test Specification Version: 2.1.1.

[16] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays, "Webgazer: Scalable webcam eye tracking using user interactions," in *25th International Joint Conference on Artificial Intelligence (IJCAI)*, Jul 2016.

[17] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Inferring human gaze from appearance via adaptive linear regression," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 153–160.

[18] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," 2015, arXiv:1504.06755.

[19] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 2, pp. 179–187, May 2019.

[20] *Lumen Research webcam eye tracker*, Lumen Research, Nov. 2019, https://www.lumen-research.com/wheretofindus.

[21] A. Kar, S. Bazrafkan, C. C. Ostache, and P. Corcoran, "Eye-gaze systems - An analysis of error sources and potential accuracy in consumer electronics use cases," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2016, pp. 319–320.

[22] A. L. Yarbus, *Eye Movements and Vision*. Boston, MA, USA: Springer, 1967.

[23] A. Nakazawa and C. Nitschke, "Point of gaze estimation through corneal surface reflection in an active illumination environment," in *European Conference on Computer Vision (ECCV)*, 2012.

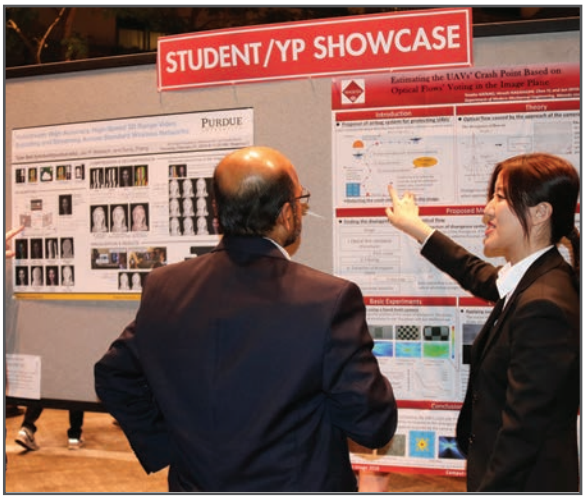
JOIN US AT THE NEXT EI!

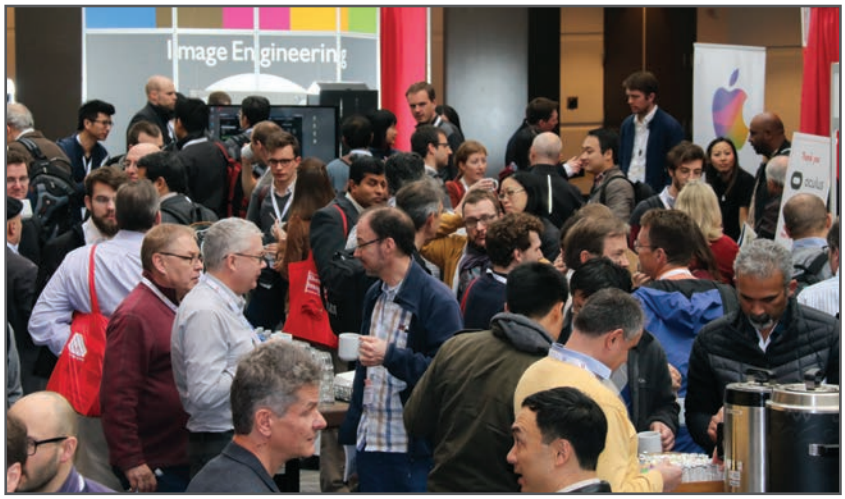
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org