Eye-tracking Without Screens: Potential Problems and How To Deal With Them

Stein Bellers

Utrecht University

Master Thesis (27,5 ECTS)

Supervisor: Niilo Valtakari

Second Assessor: Roy Hessels

July 20, 2022

Abstract

Any eye tracking set up has pros and cons and knowing the advantages and disadvantages of using a certain set up and how best to calibrate it is important for researchers choosing a setup for their study. When using an eye tracking setup without screens a lot of relevant questions remain unanswered regarding how to calibrate and use the setup. To answer those questions, a screenless remote eye tracking setup with custom built offline calibration software was used to record the gaze of both eyes of participants while they looked at several calibration and validation points. This gaze data was used to determine the way to calibrate this setup that results in as accurate and precise data as possible and investigate the effect of binocular and monocular data on the accuracy of the results to determine which is better. The constancy and predictability of the parallax error was also studied. The results show that the use of binocular or monocular accuracy led to an improvement in the accuracy of the data, but it cannot be concluded that this difference is significant. Using more calibration points does significantly improve the quality of the data when comparing a 4-point calibration to a 9-point calibration. However, the improvement is so small that in most interaction studies the improvement is not worth the extra time a 9-point calibration takes. The results concerning the distance between calibration points show no difference in data quality when using calibration points placed further apart. The parallax error seems constant enough to manually correct for it with some experience and knowledge about the forward shift of the object from the calibration plane and the location of the stimulus relative to the centre of the screen. It is recommended to test one’s setup extensively before using it in an experiment to determine the best way to calibrate it and use the data and to gain insight into how big of a parallax error one can expect.

Introduction

Eye tracking has grown from being a difficult and expensive technique which was only used in specialised research into something easily accessible, and widely useable in a variety of fields such as science and personal entertainment. Currently, there are a plethora of different types of eye trackers with varying functionalities in a wide price range. So much choice makes it difficult for starting researchers to find exactly what fits their needs leading to researchers not always getting the best possible results from their studies. To ease this start, I will start by explaining the basics of eye tracking, then give an overview of the most common types of eye trackers and some possible applications in different types of studies. Subsequently, I will dive deeper into one specific type of study, Interaction studies, and the different ways of using eye trackers in those leading to a remote screenless setup. The focus will be on how to use this setup, giving advantages and disadvantages. In the end I will give recommendations on how to use this setup derived from experiments and note some extra things to keep in mind. But first, to understand the differences between the different setups and their advantages and disadvantages it is important to understand some of the basics of eye tracking.

Basics of Eye tracking

Eye tracking is a tool that allows researchers to measure and study the gaze behaviour of individuals, meaning where they look. Different eye trackers use different techniques but most modern eye trackers use the p-CR technique which uses both the centre of the pupil and the centre of the corneal reflection to determine where a person is looking (I. Hooge et al., 2016). These gaze coordinates are typically reported per timepoint separately for the x and y axis and, depending on the eye tracker used, for either one (monocular) or both (binocular) eyes. These gaze coordinates are most often reported on a 2d plane. For most purposes, to be able to use the data, these gaze coordinates first must be fitted to points in the scene of interest. This can be done through calibration, where the test subject is made to look at certain points in a certain order. Using the time point on which the subject looked at the point and knowing the exact location of the point, the gaze coordinates can be fitted on the presented stimulus. As all test subjects are different, eye trackers need to be calibrated for every individual test subject. This is necessary because of the effect the differences in features such as differences in the size and shapes of the eyes and whether or not the subject is wearing glasses among other things, can have on the calculated gaze direction and the quality of the data (Feit et al., 2017; Holmqvist et al., 2011; Nyström et al., 2013).

When discussing the quality of eye tracking data two important measures are used, Accuracy and Precision. High accuracy means the average difference between the point where the subject is looking, and the recorded gaze positions is small. Precision is a measure of how similar consecutive measurements are to each other assuming that the true gaze position does not change (Holmqvist et al., 2011). Figure 1 from Holmqvist et al., 2011 shows a clear graphic example of these two measures. Accuracy is calculated using the mean offset of the measured gaze positions to the true gaze position which can be done using data from both eyes and/or only data from the dominant eye. Opinions vary on which calculation method is better and yields higher quality data (Cui & Hondzinski, 2006; I. T. C. Hooge et al., 2019). Precision can also be calculated in two different ways: using the standard deviation of the data sample, which is done using binocular data or by using the root mean square of the data sample which can be done using either binocular or monocular data (Holmqvist et al., 2011).

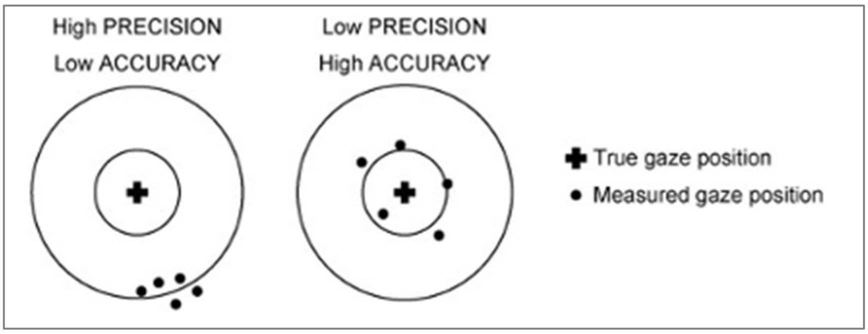

Figure 1. Difference between accuracy and precision (Picture taken from Holmqvist et al., 2011).

How precise and accurate the data should be depends on the type of study. For instance, for a study examining small movements of the eyes, such as micro-saccades, a high accuracy and precision are essential whereas for a study where one only needs to differentiate between a few large areas of interest this is not nearly as necessary. Precision and accuracy vary a lot between eye trackers. Poorer eye trackers can have a precision value of up to 1° while high-end eye trackers typically report a precision better than 0.10°, although precisions down to 0.01° are sometimes reported. Reported measured values for accuracy range from 0.3° to around 2° (Holmqvist et al., 2011).

Now that we've discussed some of the basics about eye tracking, let's discuss some of the most common types of eye trackers that are currently available and their applications.

Types of eye tracking setups

There are two main types of eye trackers to distinguish between: remote eye trackers and wearable eye trackers. Remote eye trackers often resemble a webcam and are placed somewhere in front of the subject to measure their gaze from a slight distance. Wearable eye trackers usually consist of a pair of glasses with eye trackers attached to it to measure the gaze of the wearer and a camera, attached to the glasses or a helmet, to record what is in front of the wearer. With these two types of eye trackers many different eye tracking setups can be created for a plethora of different types of studies.

Traditionally eye-tracking is done with a test subject looking at a screen on which the stimuli is presented with a remote eye tracker in front of them to measure their gaze. The subject sits in one spot with the eye tracker placed in front of them at a fixed distance. The subject can move around a bit but for the eye tracker to detect the eyes the head needs to stay inside a certain range called the headbox. Remote eye tracking can done both with and without screens. When doing it with screens one can calibrate on the same 2d plane (the computer screen) where the stimuli will be shown. This makes it relatively easy to relate the gaze coordinates to the stimuli as one knows the exact coordinates of the stimuli on the screen at any given timepoint. Because of this the data from this type of setup is relatively easy to process and has high accuracy and precision compared to wearable eye trackers where relating the gaze coordinates to stimuli in the scene of interest is much harder as the scene of interest is constantly changing (MacInnes et al., 2018; Valtakari et al., 2021).

This relative ease of calibrating remote eye trackers with screens and processing the data, and the fact the stimuli are presented on a 2d plane makes the technique ideal when one, for instance, wants to determine which areas of a poster draw the attention of the observer or in which order the gaze of customers is drawn to certain products or buttons when looking at a web shop. However, when it comes to studies into interactions between people or into gaze behaviour when someone is walking through a store, for instance, a setup using screens can be limiting. Showing a test subject a video from the point of view of someone walking through a store instead of letting the subject do so themselves is arguably less representative of a real shopping trip.

For a study like that a wearable eye tracker could be a better fit as wearable eye-trackers can be very useful when one wants to do a study where a lot of freedom of movement for the subjects is important (Valtakari et al., 2021). Good examples are studies outside of the lab, for instance in a mall or supermarket where one might want to know which products or billboards draw attention. Wearable eye trackers are generally easy to use and give the test subject more freedom of movement than remote eye trackers which do have a certain range where the head can move and the gaze can still be measured, the headbox, but compared to wearable eye trackers it can still be limiting (MacInnes et al., 2018).

The disadvantage of wearable eye trackers is that as the field of view of the subject, as recorded by the camera, is constantly changing as they walk or look around it becomes harder to relate the gaze data to the corresponding coordinates in the recorded video. When using “static” eye-tracking, where the remote eye tracker and the screen or a camera recording the field of view are in a fixed position, relating the gaze coordinates to the screen or recorded scene of interest is easier (Holmqvist et al., 2011). Also wearable eye-trackers can slip on the wearers head while moving which can also have a considerable negative effect on the data produced (Niehorster et al., 2020). Because of these factors, the data analysis can be a lot more time consuming than when using a remote eye tracker and the data quality is generally lower (Holmqvist et al., 2011). For these reasons, wearable eye trackers are better suited to studies that require cruder distinctions between relatively big areas of interest as opposed studies requiring smaller distinctions like between different facial features.

Eye tracking in interaction studies

As there are many different types of eye tracking setups and many different types of studies to do with them, even within the realm of interaction studies, the unique requirements of each study mean that there isn't one perfect setup for every study. In some cases, it might be perfectly suitable to use a remote setup with screens when doing an interaction study, for instance when you are looking into interaction behavior through a videocall or peoples gaze behavior when looking at pictures or videos of other people performing certain facial expressions or actions. If there is no reason the stimulus must be presented in 3D one should take advantage of the ease of calibration and analysis when using a remote eye tracker with screens and the high data quality.

But when the study requires face to face interaction between two people, an eye tracking setup without screens is more fitting as it is more representative of an everyday interaction and allows for physical interaction. In these cases, a wearable eye tracker or a remote eye tracker without screens can be used.

If the interaction study requires movement of the test subject and only looks at distinguishing between large areas of interest, for instance if the subject is looking at someone's face or body, a wearable setup would be suitable. However, when one wants to distinguish between small areas of interest, like between different facial features, one needs to maximize the accuracy and precision of the data. In those cases, a wearable eye tracker is not going to give accurate and precise enough data. For these types of studies, a remote eye tracking setup without screens, because its more static, is more suitable.

Remote eye tracking without screens is for the most part the same as with screens as the test subject sits in one spot with their eyes inside the headbox and the gaze being recorded from a fixed distance. The difference is that with a remote eye tracking setup without screens, the stimulus can be any 2D or 3D thing presented to the subject in the scene of interest, which is recorded by a camera placed close to, often above, the subject. The recorded gaze data first has to be related to this recorded field of view through calibration, which can be done using any 2D surface or plane with certain points the subject has to look at. The eye tracker can then be used to determine where the subject is looking while looking at another person or certain objects. The calibration can, for instance, be done using a poster or screen showing specific points or using lights that draw the subjects gaze when turned on. Another good example of this calibration process is the calibration done in Falck-Ytter's study in 2015. In this study a toddler's gaze is measured as they look at an adult telling a story. To calibrate, the toddlers attention is sequentially pulled to 5 different points using a board with holes in it and a squeaky toy (Falck-Ytter, 2015). After this calibration, the measurement could take place without any screen between the adult and the toddler, making the situation more representative of, for instance, an interaction at home at the dinner table. This study also shows that, using a static headboxed setup like this, one can use eye tracking on a toddler, which is currently not very practical with other setups (Holmqvist et al., 2011).

Just like with remote eye tracking using screens, data analysis is much simpler than with wearable eye trackers as the gaze data can be easily linked to coordinates in the recorded video since both the eye tracker and camera are in fixed positions (Holmqvist et al., 2011). This also leads to greater precision and accuracy than with wearable eye trackers, making this setup more suitable when one needs to distinguish between subtler differences in fixations like looking at different parts of a face (Valtakari et al., 2021).

Problems with screenless setups

One important aspect to consider when using a remote screenless eye tracking setup is that most eye trackers are designed to be used on a 2d surface like a screen and report their gaze coordinates in 2d. This does not make a difference when you want to use the eye tracker on another 2d surface like a poster or a painting, but it does create problems when the test subject is made to look at a 3d object. This is because the eye tracker must be calibrated on a 2d plane and reports its gaze coordinates on that 2d plane. However, if the stimulus is a 3d object or person it is most likely not (completely) on that same 2d plane. This makes the calibration less effective and can give a Parallax Error, an offset between the measured gaze position and the true gaze position.

This phenomenon is caused by the offset between the eyes of the test subject and the camera recording the field of view and occurs when the stimulus is not on the same plane the eye tracker was calibrated on. Figure 2 explains shows an example of this phenomenon. The “lines of sight” of the camera and the test subject are calibrated to intersect on the 2d plane where the calibration took place, making the measurements on that plane accurate. But when the subject is looking at something on a plane that’s behind the calibration plane the lines of sight cross each other, and the coordinates reported by the eye tracker no longer match where the subject is looking. They will be shifted by an amount depending on how much closer or farther away the plane of the stimulus is to the calibration plane. This shift in accuracy is referred to as the parallax error.

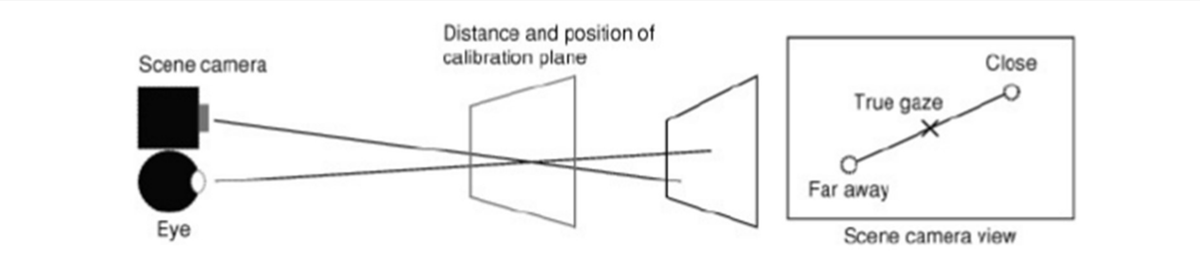

Figure 2. Parallax error. The grey plane is the position of the original calibration screen. The dark plane is the true stimulus. The scene camera view shows a frame from the overlaid scene video. The cross marks the true gaze position, and the two dotted rings mark where the eye-tracker will put the overlaid gaze marker in relation to the true gaze position, depending on the distance to the object looked at. A far away stimulus pulls the error in one direction and the close stimulus pulls it in the other direction. The gaze cursor is only perfectly positioned for stimuli at the same distance as the calibration plane during calibration. This figure assumes the scene camera is mounted above the eye level between both eyes, and the displayed error is true for measurements of the right eye (Picture and text taken from Holmqvist et al., 2011).

This shift in accuracy might not be much of a problem when the areas of interest are very large, for instance if one is only interested in whether someone is looking at someone’s right or left hand. But in a study looking at subtle social cues and face movements an unexpected/unnoticed shift in accuracy could cause the results of the study to be completely misinterpreted.

Even though the parallax error may sound like a huge disadvantage of a screenless remote eye tracking, the benefits still outweigh the drawbacks as the error is supposedly very systematic. It is said to be possible, with some experience and understanding of the phenomenon, to manually correct for it by estimating the error and shifting the data accordingly (Holmqvist et al., 2011).

The present study

Many of the commercially available remote eye trackers come with their own provided software for recording and calibration. However, as most affordable remote eye trackers are designed to be used on screens this software might not allow you to use it in the way you need for your study. So, if you want to use it in combination with a scene camera in a screenless setup the provided calibration and recording software might not accommodate for that. In this case it is possible to create your own calibration and recording software that fits your need. However, this process raises many questions that are not easily answered as not much previous research has been done with screenless remote eye tracking setups in interaction. For instance, many eye trackers allow for recording the gaze of both eyes. But as mentioned in the introduction, opinions vary on whether it is better to calculate the accuracy of a setup using the data from the dominant eye only or using the mean of both eyes (Cui & Hondzinski, 2006; I. T. C. Hooge et al., 2019). And creating one's own method for calibration means deciding how many points to use and where to place them on the calibration surface. Even in "Eye tracking: a comprehensive guide to methods" Holmqvist et al. only go as far as saying that 2,5,9,13 and 16 are common amounts of calibration points and that the calibration points should span the areas where the relevant stimuli are presented (Holmqvist et al., 2011). The experiment done for this thesis attempts to provide insight into these matters and provide tips on what to consider when creating a custom calibration method for a screenless remote eye tracking setup. The parallax error will also be examined to determine its constancy and give insight into dealing with it.

To that end, a screenless head boxed setup with a Tobii Pro Nano eye-tracker with custom built offline calibration software and a webcam was used to record the gaze of both eyes of subjects while they looked at different calibration and validation points. This gaze data was used to determine a way to calibrate this setup that results in as accurate and precise results as possible and investigate the effect of binocular or monocular data on the accuracy of the results to determine which is better. The parallax error was also calculated to determine its constancy and predictability.

The experiment served to answer the following questions:

- What is the effect of using monocular or binocular data on the quality of the data?
- What is the effect of different calibration point configurations on the quality of the data?
    - What is the effect of the distance between calibration points?
    - What is the effect of the amount of calibration points?
- How predictable is the parallax error?

To answer these questions the data was calibrated after the measurements using four different calibration point configurations using a custom-built script. The first two calibrations used four calibration points placed in a large and smaller rectangle respectively. The third and fourth calibration consisted of nine and five points respectively. To determine how accurate and precise the eye tracker data was after the different calibrations five validation points were used which were located at the top, left, centre, right and bottom of the calibration picture. To examine the parallax error, this validation was done at 3 different distances: the same distance as the calibration plane, 25 cm in front of the calibration plane (where a face might be in a test), 50 cm in front of the calibration screen (where hands might be during a test). To determine the effect of using only the data from the dominant eye or the average of both eyes the average accuracy and precision was calculated after the first calibration for both monocular and binocular data. Afterwards the data was compared between the different calibrations and validation distances and between the monocular and binocular data to draw conclusions.

Methods

The setup

The setup used in this experiment consisted of a table with a Tobii Pro Nano eye tracker with a sample rate of 60Hz attached to it with a monitor arm which allowed for adjustment. The table stood in a fixed spot and the eye tracker was adjusted to match the person sitting at the table. All subjects sat in the same spot to ensure a relatively constant distance to the test screen which was placed in 3 different pre-determined positions during the experiment. The first position being used for calibration and the first validation. The second and third positions (25 cm and 50 cm in front of position 1 respectively) were chosen because the planes roughly coincide with where the face and hands of a second person might be when using this setup in an interaction study. The camera behind and above the test subject was also placed in a fixed spot right behind the chair the subject sat on to ensure that the views of the subject and the camera lined up as much as possible. The screen was used to show pictures of points in different configurations. See figure 3 for a picture of the setup. The screen and the points on it were placed in such a way that the centre dot was right in front of the subject.

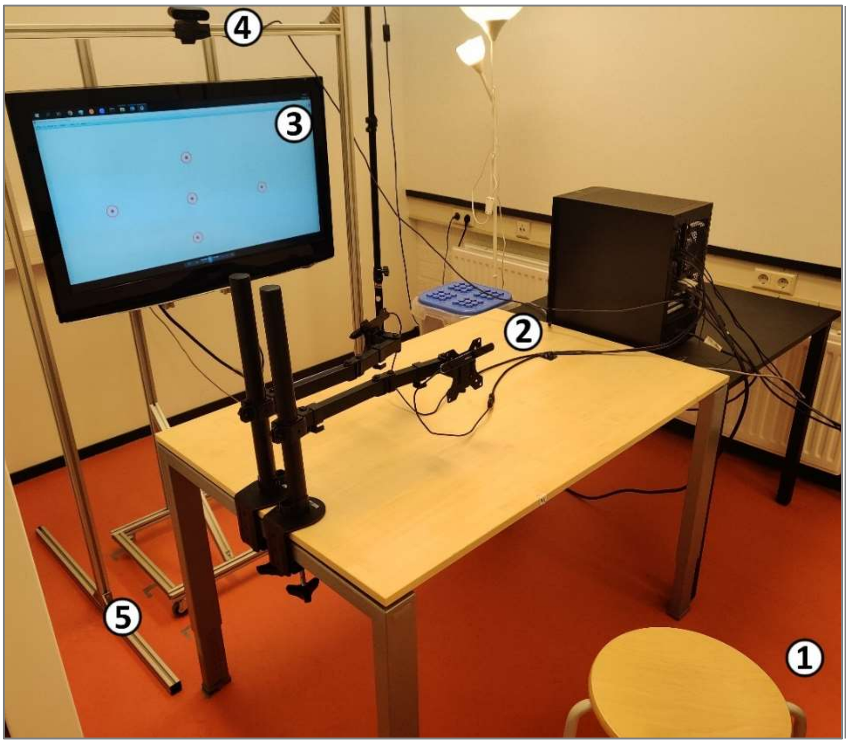

Figure 3. Picture of the setup used in the experiment with the chair(1), eye tracker(2) and screen(3) visible. A webcam(4) such as the one visible above the screen in the picture is placed behind the participant to record the field of view of the test subject. On the floor (5), markings for the 3 pre-determined screen positions are visible.

The experiment

Five participants were recruited to serve as test subjects. The subjects were given an information letter describing the experiment and signed a consent form before the experiment took place. The subjects were placed on the seat. Their dominant eye was determined before testing began. They were then told find a comfortable position and look straight ahead at the centre dot on the screen. The eye tracker was adjusted to detect their eyes when looking at all corners of the screen. The subjects were instructed to try not to move too much from their current position, to not turn their heads when looking at the different points in the experiment and to try only to blink in between looking at the points and as less as possible while looking at them.

The participants were then instructed to look at all 13 calibration points used for the different calibration point configurations. For each point three seconds of gaze data was recorded. To reduce the likelihood of confusion, two or three points were shown at any given time. This process was then repeated for the five points used for validation after which the screen was moved to position 2. The validation process was then repeated after which the screen was moved to position 3 and a final validation was done for a total of three validations, one per screen position. Every participant performed this entire test (calibration and validation) twice for a total of 10 tests.

Data evaluation

The gaze data was transformed using a custom-built offline script. The script calibrates the data using specified calibration points and extracts the recorded data for all the calibration and validation points. The gaze data was calibrated four times per test using the calibration configurations shown in figure 4. Accuracy and Precision measures where then calculated per point. Accuracy was calculated as the distance in pixels between the measured gaze position and the true gaze position. Precision was calculated as the standard deviation in pixels. Using the distance from the eyes of the test subjects to the screen used to present the stimuli a conversion rate from pixels to degrees of the visual field was calculated which was used to convert all results to degrees of the visual field. Accuracy was calculated twice using both data from the dominant eye only and using the mean of both eyes. Precision was calculated using the standard deviation of the recorded gaze position using the mean of the data of both eyes.

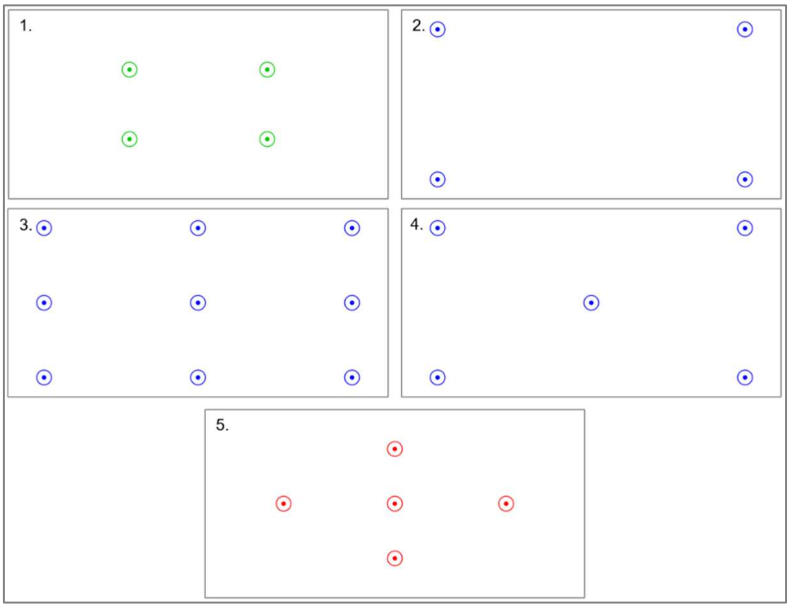

Figure 4. The calibration and validation point configurations used for calibrations 1, 2, 3 and 4 respectively. Picture 5 was used for validation.

The average accuracy and precision over all tests were then used to compare the different calibrations and the monocular and binocular data. See figure 4 for an overview of the different calibration and validation screens. Calibration 1 was compared with calibration 2 to determine the effect of distance between the calibration points. Calibrations 1, 3 and 4 were compared to determine the effect of amount of calibration points. The validation points in figure 4.5 were used to measure and calculate the accuracy and precision of the data after each calibration.

Results and Discussion

Which is better? Monocular or Binocular data

To determine whether binocular or monocular data leads to better accuracy the accuracy was calculated using both and the results were compared. The results can be seen in table 1 and figure 5, shown separately for all 10 tests (2 tests per subject) and as the mean over all tests. In all but two of the tests using data from the mean of both eyes led to a better accuracy to using only the data from the dominant eye. This effect is also seen in the average of all tests: the data saw an average improvement of 0.11° when using the data from both eyes. The data improved most in the X direction compared to the Y direction, which was to be expected given that the position of the eyes varies in the X direction but not the Y direction. Although the average accuracy of the tests did improve, because the Standard Error of the Mean error bars overlap, we cannot assume that this difference is significant. However, as in this dataset it does improve accuracy the binocular accuracy was used for the remainder of the data analysis.

Accuracy measure	**Average accuracy**
Dominant eye only	0.63° (SEM=0.07°)
Mean of both eyes	0.52° (SEM=0.07°)

Table 1. Average accuracy over all five validation points of all tests using either data from the dominant eye only or the mean of both eyes. Given in degrees of the visual field with the Standard Error of the Mean.

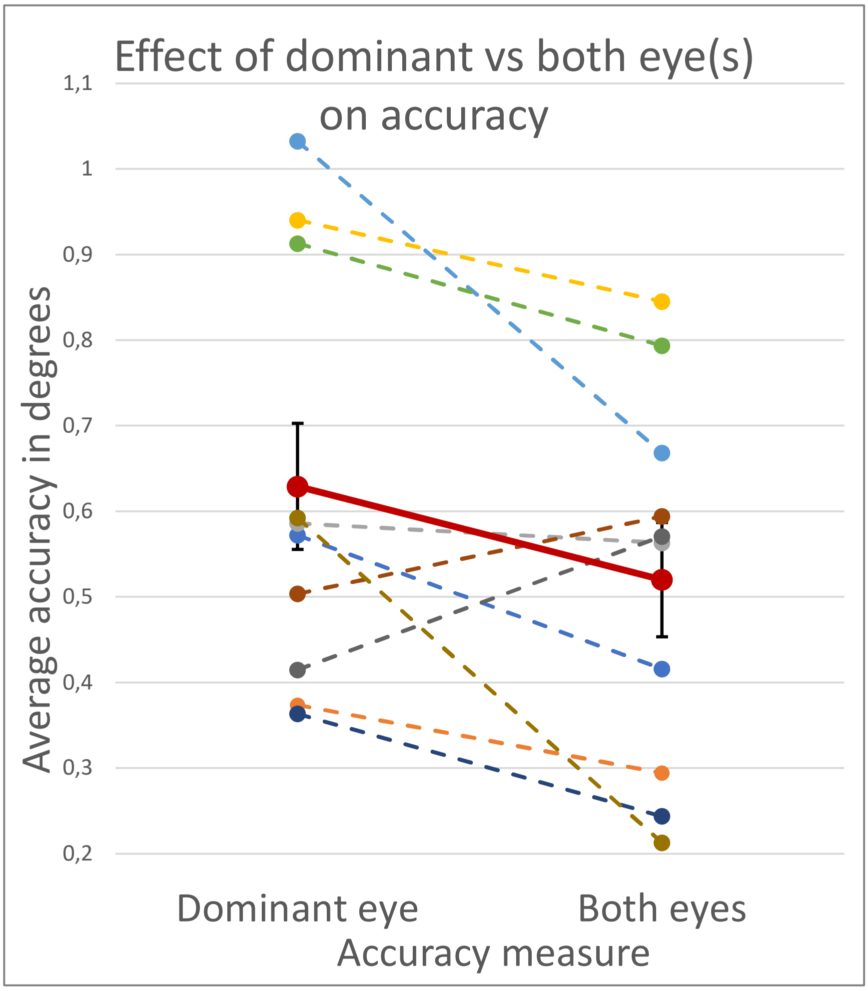

Figure 5. Average accuracy per test after using either dominant eye data only or the mean of both eyes. The dotted lines connect the accuracy measures from the same test. Shown in degrees of the visual field with red showing the average over all 10 tests. The error bars show the Standard Error of the Mean.

The effect of calibration point configuration

To determine the effect of different calibration point configurations the average accuracy and precision was calculated after each of the four calibration methods and compared. This comparison made it possible to see the effect of the changes in configuration on the accuracy and precision of the data. See table 2 and figure 6 for the results.

Calibration	Average accuracy
1 (outer 4 points)	0.52° (SEM=0.07°)
2 (inner 4 points)	0.52° (SEM=0.05°)
3 (9 points)	0.38° (SEM=0.05°)
4 (5 points)	0.46° (SEM=0.05°)

Table 2. Average accuracy over all five validation points of all tests after each of the calibrations. Given in degrees of the visual field with the Standard Error of the Mean.

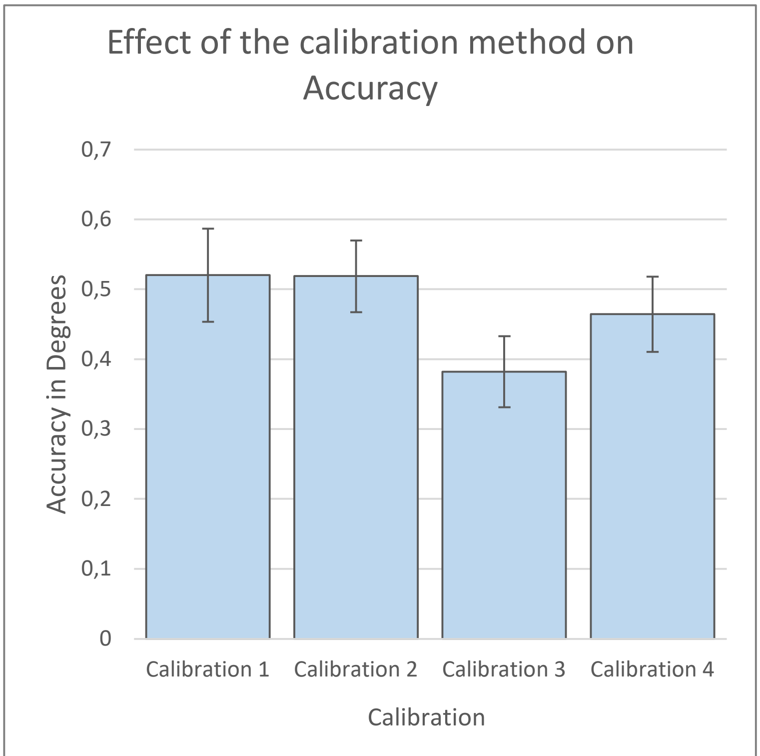

Figure 6. Average accuracy per calibration shown in degrees of the visual field. The error bars show the Standard Error of the Mean.

The placement of the calibration points had no effect on the average accuracy as both calibration 1 and calibration 2, the big and small 4-point calibration respectively, had an average accuracy of 0.52°. Using more points did improve the accuracy as adding a fifth point in the middle of the screen led to a small accuracy improvement as seen by the 0.46°

average accuracy after calibration 4. This improvement is more prominent after calibration 3, the 9-point calibration, with an average accuracy of 0.38°. The 0.06° accuracy difference between the 4-point calibrations, calibration 1 and 2, and the 5-point calibration, calibration 4, cannot be assumed to be significant as the Standard Errors of the Mean overlap. This is not the case for the 0.14° difference between the 4-point calibrations and the 9-point calibration, calibration 3, meaning these differences are significant.

However, in this setup, with the measurements done at roughly 165 cm distance between the eyes of the subject and the screen this improvement from 0.52° to 0.38° translates to an improvement from 1,50 cm to 1,09 cm distance between the true gaze position and the recorded gaze position respectively. So, the improvement is only 0,41 cm which begs the question whether it is worth the extra time required to perform a 9-point calibration as opposed to a 4-point one. As most studies using remote eye tracking on stimuli at this distance do not look at areas of interest small enough for 0,41 cm to make a difference, using a 4-point calibration will in most cases be the better option as it saves time and effort. Especially when the study involves young children for whom performing a 9-point calibration might take too long, causing them to lose their attention.

The effect of the different calibration methods on the precision of the data was also compared. As can be seen in table 3, no relationship between distance between calibration points or amount of calibration points and precision is suggested. This makes sense as the different calibrations affect the “placement” of the gaze data on the video coordinates but not as much the distance between the recorded data points themselves.

Calibration	**Average precision X direction**	**Average precision Y direction**
1 (outer 4 points)	0.13° (SEM=0.02°)	0.17° (SEM=0.03°)
2 (inner 4 points)	0.14° (SEM=0.02°)	0.16° (SEM=0.03°)
3 (9 points)	0.13° (SEM=0.02°)	0.17° (SEM=0.03°)
4 (5 points)	0.13°(SEM=0.02°)	0.17° (SEM=0.03°)

Table 3. Average precision over all five validation points of all tests after each of the calibrations. Given in degrees of the visual field with the Standard Error of the Mean.

The parallax error: predictable or not?

To be able to accurately predict and correct for the parallax error it is important to know how constant and predictable it is. So, to examine how constant the parallax error is the average shift in accuracy when moving the screen from position 1 (the calibration position) to position 2 (25 cm forward from position 1) and the shift when moving from position 2 to position 3 (50 cm forward from position 1) where compared separately for all 5 validation points. See figure 3 for a picture of the setup and figure 4.5 for the picture with validation points that was presented to the test subjects. Average shifts per validation point are shown in figure 7 and visualised in figure 8.

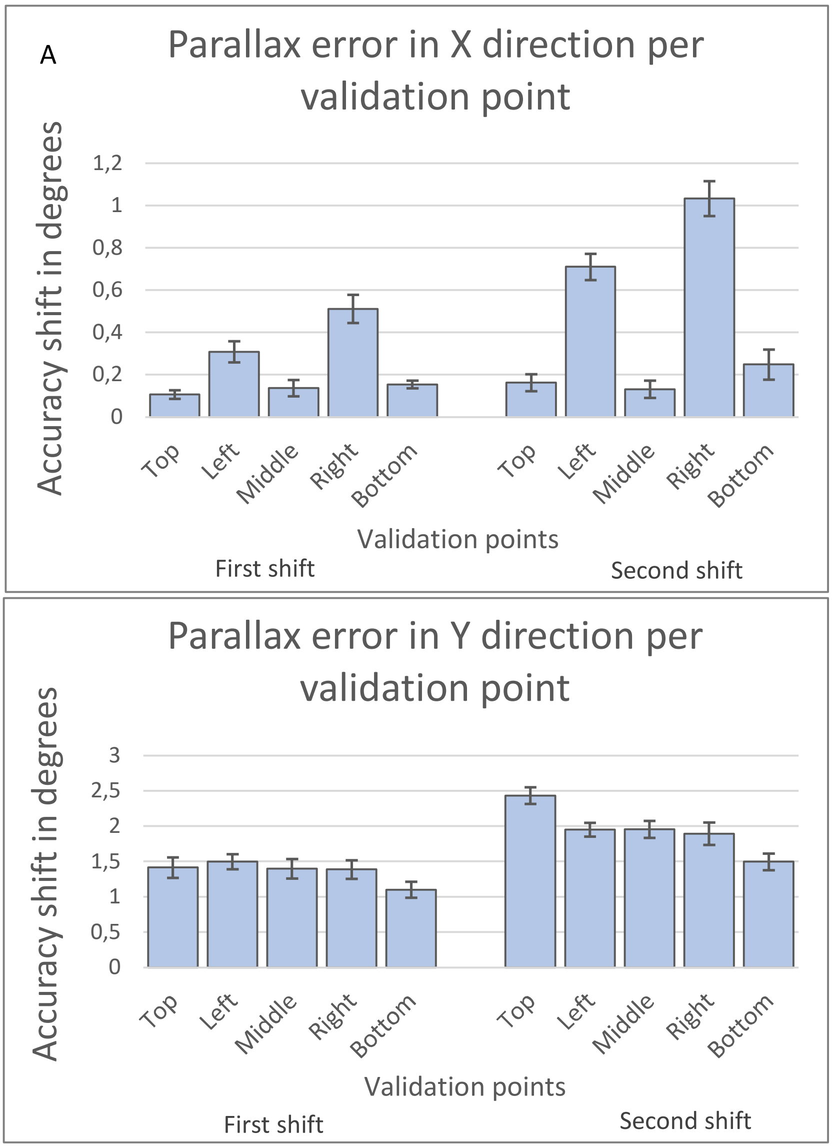

Figure 7. Average parallax error per validation point in degrees of the visual field for A) the X direction and B) the Y direction. The left clusters show the shift in accuracy between the data recorded with the screen in position 1 (the calibration position) and position 2 (25 cm forward). The right clusters show the shift in accuracy between the data recorded in position 2 and position 3 (50 cm forward from the calibration position). The error bars show the Standard Error of the Mean.

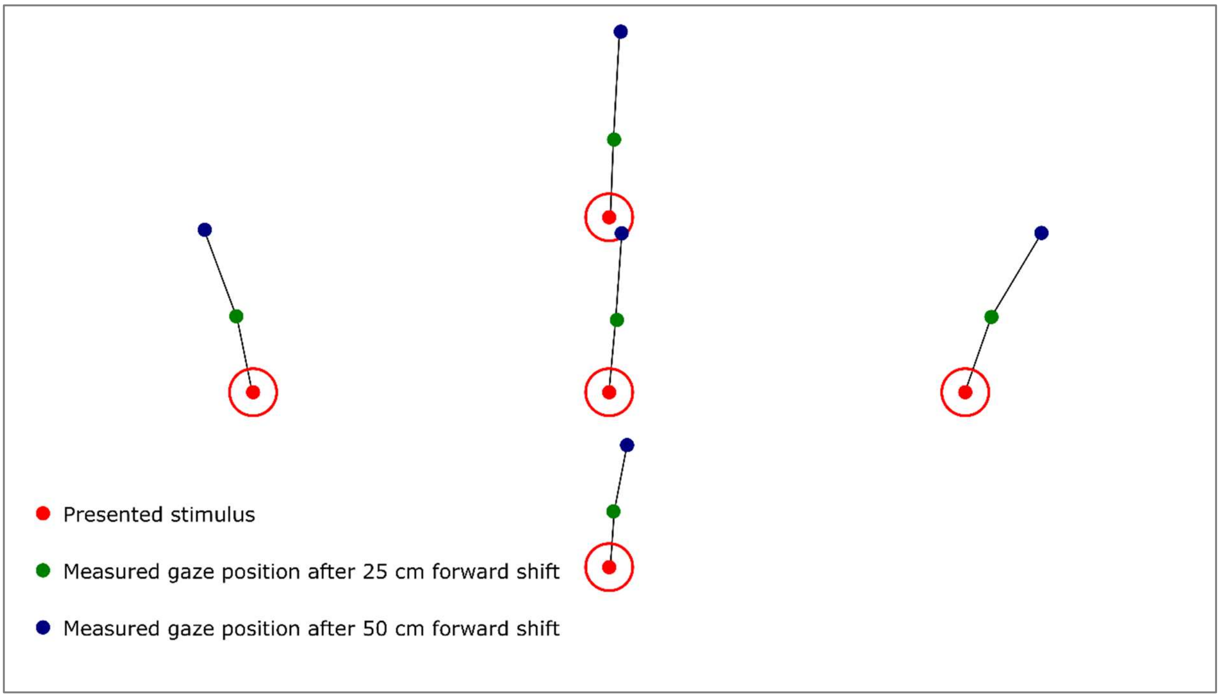

Figure 8. The average parallax error per validation point after the first and second forward shift visualised.

In the results, three things stand out. Firstly, the parallax error is largest in the Y direction, in this case upward. The shift in the X direction is relatively small on all points but does become bigger when the stimulus is presented further away from the middle of the screen in the X direction.

Secondly, the parallax error is larger in the second shift forward, suggesting a logarithmic relationship between the forward shift and the parallax error instead of a linear one. In the Y direction the first shift forward increased the accuracy of the data by 1,36° on average, which is 0,59° degrees less than the 1,95° average accuracy increase after the second shift forward even though both shifts were the same distance. In the X direction the first shift increased the accuracy by 0,24° and the second shift by 0.46° on average meaning a difference of 0.22° which means it almost doubled. This increase in the parallax error when the shift is closer to the subject makes accurately predicting the error harder.

Thirdly, as seen in figure 8, with a forward shift of 50 cm the parallax error becomes big enough that an unknowing researcher interpreting these results might wrongly conclude that the subject is looking at the top validation point when they are in fact looking at the middle point. Given the big parallax error in the Y direction, studies distinguishing between areas of interest presented above each other should be especially mindful of the effect of the error on the data. For instance, seeing as the position of the second shift forward was chosen because it roughly coincides with the plane of the hands of a person in front of the subject one can imagine that the results of a study that has the two subjects interact and present objects to each other might be seriously affected.

However, given the relatively small error in the X direction and the fact that the error shifts the data further outwards, studies that present the areas of interest in the X direction, for instance one object in the left and another in the right hand, will be less affected by the parallax error no matter how far forwards the hands are held.

To determine if the error could be reduced or prevented the effect of the different calibration point configurations and the use of data from the dominant eye or the mean of both eyes was also examined in relation to the parallax error but no effect on the error was found.

Even though the logarithmic relationship between the forward shift and the parallax error and the effect of the stimuli being closer to the sides of the field of view make the error more difficult to predict, manually correcting for it still seems possible with enough knowledge about the distance of the stimuli from the calibration plane and the centre of the screen. To truly understand what one can expect with regards to the parallax error in their experiment, it would be worth the effort to perform a “parallax error measurement” before one’s experiment. One could, for instance, tell the subject during calibration to not only look at points on the calibration surface but also at points roughly placed on the plane/location where the stimuli will be presented. This way an estimate of the parallax error can be calculated in advance, making it easier to correct for.

Future studies

In this study I have tested the effect of different variables on the data of one type of eye tracker, calibrated using one calibration method, a custom-built offline script. It remains to be seen whether these results can be generalized to other eye trackers and methods of calibration, for instance built in manufacturer calibrations. The improvements in data quality seen in this setup may not be the same for a different setup so it warrants extra research into whether these same effects of calibration point configuration and binocular data are seen when using other eye tracking products and calibration methods. As eye tracking setups vary, perhaps it is always best to test the eye tracking setup used in one's study thoroughly in advance to determine how to calibrate and use it to get the highest quality data, keeping in mind that different types of studies require different levels of precision and accuracy.

Conclusions

The results show that the use of binocular or monocular accuracy led to an improvement in the accuracy of the data, but it cannot be concluded that this difference is significant. Using more calibration points does significantly improve the quality of the data when comparing a 4-point calibration to a 9-point calibration. However, the improvement is so small that in most interaction studies the improvement is not worth the extra time a 9-point calibration takes. The results concerning the distance between calibration points show no difference in data quality when using calibration points placed further apart. The parallax error seems constant enough to manually correct for it with some experience and knowledge about the forward shift of the object from the calibration plane and the location of the stimulus relative to the centre of the screen. It is recommended to test one's setup extensively before using it in an experiment to determine the best way to calibrate it and use the data and to gain insight into how big of a parallax error one can expect.

References

Cui, Y., & Hondzinski, J. M. (2006). Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience Letters*, *396*(3), 257–262. https://doi.org/10.1016/j.neulet.2005.11.071

Falck-Ytter, T. (2015). Gaze performance during face-to-face communication: A live eye tracking study of typical children and children with autism. *Research in Autism Spectrum Disorders*, *17*, 78–85. https://doi.org/10.1016/j.rasd.2015.06.007

Feit, A. M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., & Morris, M. R. (2017). Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 1118–1130. https://doi.org/10.1145/3025453.3025599

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye Tracking: A comprehensive guide to methods and measures. In *Oxford University Press* (1st ed.). OUP Oxford. https://doi.org/10.1080/17470218.2015.1098709

Hooge, I., Holmqvist, K., & Nyström, M. (2016). The pupil is faster than the corneal reflection (CR): Are video based pupil-CR eye trackers suitable for studying detailed dynamics of eye movements? *Vision Research*, *128*, 6–18. https://doi.org/10.1016/j.visres.2016.09.002

Hooge, I. T. C., Holleman, G. A., Haukes, N. C., & Hessels, R. S. (2019). Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods*, *51*(6), 2712–2721. https://doi.org/10.3758/s13428-018-1135-3

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). Wearable Eye-tracking for Research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. In *bioRxiv* (p. 299925). https://doi.org/10.1101/299925

Niehorster, D. C., Santini, T., Hessels, R. S., Hooge, I. T. C., Kasneci, E., & Nyström, M. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, *52*(3), 1140–1160. https://doi.org/10.3758/S13428-019-01307-0/FIGURES/12

Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*(1), 272–288. https://doi.org/10.3758/s13428-012-0247-4

Valtakari, N. V, Hooge, I. T. C., Viktorsson, C., Nyström, P., Falck-Ytter, T., & Hessels, R. S. (2021). Eye tracking in human interaction: Possibilities and limitations. *Behavior Research Methods*, *53*(4), 1592–1608. https://doi.org/10.3758/s13428-020-01517-x