

Accurate Eye Tracking from Dense 3D Surface Reconstructions using Single-Shot Deflectometry

Corresponding Author: Professor Florian Willomitzer

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors propose a novel eye-tracking method based on single-shot stereoscopic deflectometry of a crossed sinusoidal pattern for human eyes. This method is designed for the tracking of human eye movement but no qualitative eye trace (eye movements over time) nor quantitative evaluation on human eye movements (even few subjects) is given.

The apparent closeness of the stereo rig and of the illumination screen, large at this distance - a large portion of subject field of view is occluded by this structured illumination source - makes it difficult to realise if this eyetracking method is indeed feasible in a normal or controlled environment such as a lab with a chin and headrest. It would have been interesting to add a more realistic distant structured source of illumination (e.g. of same size and same distance as most eyetrackers) to evaluate how a reduced amount of points maintain good 'precision'; which might indeed be the case. Hence a measurement of the variations of RMS S2S precision and 2-steps 9 or 13 points accuracy on few human eyes as a function of the reduction of the size-reflection points could bring insights on a potential feasibility of structured illumination source and PMD for eyetracking. Not mentioned the wavelength used here, were tests using a NIR source would also bring more insights on feasibility.

Secondly and most importantly, the main claim being that acquiring 43000 reflection surface points instead of 1-8 (typically 2) glints would lead to an increased precision, if not accuracy, it is hard to conceive how this method can compare to the Literature without using the same methodologies (2 step calibration-validation on 13 points for accuracy measurement) nor the same metrics: 2 step accuracy and RMS S2S precision. Authors chose to evaluate the accuracy of their system on model eye which is a necessary and instructive step but incomplete.

Specifically regarding the way accuracy is usually measured in a 2-steps 9 or 13 points process, it is true this method is far from ideal. However:

- the head is made as still as possible using chin and head rest, bite bar when possible
- there is indeed constant eye movements, called fixational eye movements, that prevents to consider the eye is not moving, -but these eye positions during fixation on a calibration target are small (typ SD 0.15 0.25°) not normally distributed, (see Cherici Poletti 2012) BUT not far from normal distribution at the considered scale of measurement here. So the assumption that movement of the eye during fixation could be averaged during a short amount of time still hold true. (especially on small targets, see McCamy 2013, Thaler 2013).

Regarding precision, the measurement here on model eye are interesting, but similarly should have been extended on human eye to be comparable to existing methods. The Root mean square sample to sample (RMS-S2S) deviation is a way to measure the precision of a tracking method suggested by Holmqvist and Nyström in 2011. Normative values of RMS-S2S exist and would then allow to locate the potential of this method among the numerous existing ones.

Hence conceptually, it is difficult to consider how one can claim a significant improvement over the SotA without using any of the same methods to quantify it.

The 3 contributions i) stereo deflectometry ii) correspondence via crossed sinusoidal illumination pattern iii) gaze estimation and quick surface normals estimates are interesting and well-developed.

It remain unclear to me 2 Important things:

as the authors noted, the estimated axis is the optical axis and not the visual one, and need to account for the angle Kappa, so in practice any measurement of the gaze on screen or in 3D world would actually need a calibration step. So please bring this accuracy measurement step into the study as it is needed at any point to retrieve the gaze.

A hidden or undescribed necessary pre-segmentation of the reflection pattern must have been done so that the Camera 1 only use the part of the reflection from the Cornea (Fig7a and 6c) or Camera 2 from the sclera (Fig7b and 6b) to get the corresponding reconstructions.

The normals obtained from the limbus appear as outliers, but I doubt that almost half of the points (cornea or sclera) are also rejected as outliers: please describe the preprocessing if needed in a real time implementation of an actual human eye tracker.

Finally, there must be a dead range, an asymmetrical tracking capability in the visual field where this method (where cameras are place to the side of an eye and not above or below) would show its limits. Would you consider positioning the camera along a vertical axis below the primary gaze position of a subject ? Please report these possibility in the discussion.

In conclusion, I would largely encourage the authors to pursue and finalize their study that carry potential benefits to the research community.

Providing 2-step accuracy evaluation , RMS S2S precision on a limited number and basic vertical and horizontal traces of a LIMITED number of subjects would be of highest interest.

Yours sincerely.

note: other comments and note and typos enclosed

Reviewer #2

(Remarks to the Author)

In general, this paper proposes a new method for eye tracking that is fast, accurate, and precise based on advancements in single-shot phase measuring deflectometry. The paper is well-structured and make a significant contribution to the field of eye tracking. However, it is uncertain whether the novelty of this paper is sufficient for acceptance in the Nature Communication journal.

1. The main contribution of this paper lies in utilizing phase measuring deflectometry to estimate gaze direction. However, when evaluating the accuracy of the proposed gaze estimation method, the authors employed their own error metrics, which differ from those used in previous literature. The authors acknowledge the difficulty of comparing their method with others due to the variance in error metrics and evaluation procedures. How can we ensure that this method is superior to others in the literature if direct comparisons cannot be made? Moreover, even though it is reported that the error of current eye tracking methods varies within a range of 0.5o to 1o, this error range was also evaluated using different error metrics. If the authors cannot clearly demonstrate the improvement compared to state-of-the-art methods, it becomes difficult to convince readers of the novelty of this method.

2. In the introduction, the authors mentioned: "For those methods, it can be intuitively understood that more acquired 3D surface information about the eye generally leads to higher gaze reconstruction accuracy [14]". This statement reflects an intuition about the potential relationship between the amount of 3D surface information and gaze reconstruction accuracy. In this research, the authors collected more than 40,000 3D surface points. It would be valuable and necessary to validate this assertion by varying the number of 3D information points and calculating gaze accuracy accordingly.

3. The authors should provide further explanation regarding the shape reconstruction part: How to overcome the depth-normal-ambiguity problem by utilizing only the overlapping region between two cameras? Additionally, the authors should include more detailed explanations and figures to describe the proposed surface reconstruction method.

4. In this paper, the authors also assume that the cornea and sclera can be approximated by two spheres with different radii. What if the authors were to use more accurate models, such as the prolate shape for the cornea, where the radius of curvature increases from the center to the periphery [1]? Would this lead to a greater reduction in gaze error or not?

5. The authors use the screen in the off-state as a mirror to calibrate display position, which might introduce errors due to the non-perfect planarity of the screen. How could this affect the calibration accuracy?

6. The authors mentioned a high-precision rotation stage in the experiment of gaze estimation. The authors should specify the model and precision level of that rotation stage.

Reference:

[1] Melanie C. Corbett, David P.S. O'Brart, David C. Saunders, Emanuel S. Rosen, The Topography of the Normal Cornea, European journal of Implant and Refractive Surgery, Volume 6, Issue 5, 1994, Pages 286-297, ISSN 0955-3681, [https://doi.org/10.1016/S0955-3681\(13\)80229-8](https://doi.org/10.1016/S0955-3681(13)80229-8)

Reviewer #3

(Remarks to the Author)

[Paper summary]

This paper presents an interesting solution for the head-mounted eye tracking task with dense 3D surface reconstruction. Although the main techniques such as single-shot deflectometry have been proposed in previous works, it has not to be used for the eye tracking task yet. The proposed method has great potential to boost the performance of eye tracking into a new standard. Unfortunately, the author does not provide the gaze estimation performance with real humans and the proposed method also has some strong limitations that have not been properly discussed.

[Paper strength]

- Applying the single-shot phase-measuring-deflectometry (PMD) for eye tracking is a novel and interesting idea. The method can reconstruct the dense 3D structure of the cornea and sclera to find the optical axis. It is a promising method for accurate gaze estimation. Considering the fast eye movement, the author applies the single-shot techniques for the reconstruction, which make sense to me.

- The proposed method mainly relies on the previous method [37] for single-shot deflectometry, and [36] for the stereo deflectometry. Therefore, the technical novelty of the proposed method is limited. However, I think applying these techniques to eye tracking systems is a novelty. Taken together, I think it is one of the paper's strengths.

- The paper is well-written. The method itself is well-motivated in the introduction that we do need a new method for the accurate gaze estimation for the head-mounted eye tracker. The method is clearly demonstrated with illustrative examples.

- The author also explained well about the optical and visual axis for eye gaze estimation.

- The proposed camera calibration method is interesting. The author turning off the phone screen to make it like a mirror is an interesting idea. I agree with the author that it could provide good localization of the phone (which is the mirror) for the display-camera extrinsic calibration.

- The author provides three experiments to show the advantages of the proposed method with surface reconstruction quality, gaze estimation with eye model, and human eye in-vivo. The first experiment on surface reconstruction results show the proposed system works well.

[Paper weakness]

- From my understanding, the main drawback of the proposed method is that it requires the projection from the display on both the cornea and sclera, which is very difficult to satisfy when the person looking away from 0 degrees. The author only evaluates the small gaze angle $[-4, 4]$ degrees in the experiments section 3.2. I suspect the proposed method will not work for large gaze angles. It will strongly constrain the proposed method to be used for the eye tracking system. Although this requirement can be fulfilled with more cameras from different angles, this limitation is not been discussed in the paper.

- Another drawback of this method is that it requires highly accurate camera calibration. Conventional eye tracking also needs accurate camera calibration. However, these conventional methods are much more robust than the proposed method.

- Fig. 1 is a bit confusing. Especially, I hope the author makes it more clear what is shown in the Fig. 1 (c). Is this the image for one image or from the overlay of two cameras? It would be better if the author showed the eye images from both cameras to clarify it.

- For the camera calibration method in section 2.3, there is a risk that the calibration pattern on the "off-state" is not clearly visible for cameras. The "off-state" phone is a black mirror that may not have the clear and shape reflection of the display. The author could show the image examples captured by the camera to show how good the reflected calibration pattern is.

- The author did not perform the gaze estimation experiments with real human users. From my point of view, it is a pity that the author did not conduct the experiment, or compare it with conventional methods. For the first paragraph in section 3.2, I agree with the author for the first argument that using a third-party eye tracker could have ground-truth errors. However, I do not agree with the author that they cannot use the screen to calculate the ground-truth gaze direction. The experiment in section 3.2 with the eye model cannot reflect the real gaze estimation accuracy. For example, the eyeball does not have the eyelid occlusion.

If the author insists on not conducting the gaze estimation evaluation with real human eyes, at least, they can use the conventional method for the comparison. By checking the example in Fig. 6, I think the conventional pupil and iris detection methods may also achieve comparable accuracy with the proposed method. In this case, the author needs to provide good lighting conditions for these conventional methods for a fair comparison.

I understand the current system is just a prototype that cannot be integrated into an XR helmet. Nonetheless, I feel very pity that the author does not conduct gaze estimation experiments with real human eyes for measurement, especially compared with conventional methods.

[Questions]

- The author motivates the single-shot solution (section 2.1.2) with the delay of the sequential image capture solution (section 2.1.1). I am curious to know if we lose any reconstruction accuracy with the single-shot solution compared to the sequential image capture solution. Assume we have a camera with a high frame per second, can we just use the sequential image capture solution?

- To save the computation, the author picked a few sparse "anchor points" for the calculation (section 2.1.3). Does it mean

that the author did not use all of the 40,000 surface points as claimed in the introduction section? The author needs to clarify how many surface points they utilized.

- The author uses the normal vector to calculate the cornea center and sclera center as described in section 2.2. How many normal vectors they usually can obtain?

[Justification for rating]

I appreciate the effort from the author to introduce the single-shot deflectometry for the eye-tracking task. I think it is a promising method to reconstruct the dense 3D eyeball surface. The method could greatly improve the gaze estimation performance upon success. The paper is well-written and the method is interesting. I hope the author could (1) further discuss the limitation of the method in terms of gaze angle coverage, even better if they could build a system to overcome the limitation, and (2) conduct the comparison with conventional eye tracking methods even better if they could conduct experiments with real human users quantitatively.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I would like to thank the authors for the additional work and experiment conducted for this study.

Most of my concerns have been addressed, so I would like to endorse this work for publication, with some changes or statement that need to be addressed. Some bold statements make the paper appear weaker than it actually is.

1) In a response to my concerns regarding the large occlusions induced by field of view of the deflectometry display, you replied that "the setup did not occlude the subject's field of view for this experiment" which is not true, just by looking at Fig 7a. Yes, it does not occlude the 30x15° viewed display, but considering a majority of applications, and notably VR headsets that you consider, this might be an issue. That deflectometry display is probably larger than 30x15° in the lower right field.

"This means that the proposed technique and setup are indeed feasible in a normal and controlled environment (we admit that problems might occur for very large gaze angles, but similar restrictions apply for standard glint tracking where all glints need to be "located" on the cornea)". Well Tobii as previous researchers have developed an eyetracking glass where the IR led are embedded in the frame around the eye; similar restrictions does not apply and alternative method to track the eye in a large FoV do exists.

You did mentioned however "fixed known back-lit patterns that could cover the interior of VR glasses, or exploiting the known content of the main (VR) display", which is good and important to mention if you consider an actual usage of such specific deflectometry tracking requiring a large *illumination* field.

2) I am actually a bit concerned by the final relatively low precision and accuracy obtained whether with a rotated model eye or 3 human subjects.

-On a model eye, in table 2, the precision is close to the accuracy, and even sometime higher. An error or 0.1° at 4° is actually quite large for a static model eye.

-On human, the obtained accuracy [0.5 to 1°] is higher than what is typically obtained with an Eyelink across 9 fixation targets, typically [0.25 to 0.5°]. The value obtained are closer to what is found by a mid-range mid-frequency commercial eye tracker, such as find within VR displays.

I understand that no polynomial fitting is used but a 2 steps methods using an initial surface reconstruction step and a refinement stage. I would presume that further improvements might be obtained by working on such method, probably by incorporating a fit with the actual pixel position correspondance on screen.

So stating that "Although early stage, our first experiments already reach accuracy and precision values which are on par with the best current state-of-the-art methods. " is a largely incorrect statement.

3) In a long paragraph, the authors try to justify that i) their system provide performance similar performance than a model eye by saying that one specific location it is close to the model eye (which is not true, a factor of x1-x4 depending on subjects)

ii) then contradict themselves stating that "there is no obvious degradation in performance when comparing measurements at the center stimuli positions to those at the outer positions"

Which is also wrong both in term of mean accuracy (over the 3 subjects, per location) or mean precision: accuracy and precision appear to be best in central location.

4) Why is precision reported in the 4 corners of the 9 dot calibration chart, but not the accuracy ? The fact that it is missing lead us to think that it was of low accuracy, which might not be the case but is concerning. Please add these 4 values if possible, or add an explanation.

5) I acknowledge the measure of the number of glints on accuracy. The degradation for a randomly sampled subset of points is actually quite large, but appear to plateau. (a figure would have been better in my opinion). A thinning of the 2D distribution

would have provided an even worse effect on accuracy. Could you please discuss on the effects that a smaller (noise) OR thinner amount of points (effect of deflectometry screen distance) would have on precision and accuracy and what it means in terms of application and restriction to VR/AR usage?

Thank you

Reviewer #2

(Remarks to the Author)

The authors have made significant revisions to the manuscript, focusing on two main concerns: measurement and comparison with real eyes in vivo using similar evaluation metrics as in previous methods, and the shape reconstruction method for gaze estimation from deflectometry measurements. The authors have addressed all the comments from the preview review, and I believe the manuscript is ready for publication. However, it would be beneficial if the authors could include a discussion on phone screen flatness, as mentioned in comment R2.5, in the revised manuscript.

Reviewer #3

(Remarks to the Author)

I thank the author for their professional revision and rebuttal. Most of my concerns are addressed properly.

For my first question R3.1 related to the novelty, I thank the authors for a more detailed explanation and for rewriting the respective section of the registration algorithms. I also recognize that the novelty of this paper is that the complete pipeline goes from setup design, system calibration, measurement surface evaluation, and gaze evaluation.

For the discussion related to the R3.2, I think I did not make myself clear in my original comments. My main concern about the method limitation is that the method requires capturing parts of both the cornea and sclera "at the same time". I agree with the authors that this limitation can indeed be solved with more cameras or positions on the screen. I am glad that the author discusses these limitations in section 4 now.

For point R.3.3, I agree with the author that the requirement for the calibration of active methods can be similar to the method proposed in this paper. I am also happy to see that the author discussed it in section 4. Therefore, I am happy to suggest this paper be accepted as it is.

For point R.3.4, I thank the author for providing the example images in Fig. 4, and pointing out that they can pick a large exposure time to obtain a good signal.

Finally, I appreciate the effort from the author in providing experiments on real human users. Although this paper is a proof-of-concept, it is still good to have this user study to show the effectiveness of the proposed method.

I am also satisfied with the answers provided by the authors for the rest of my questions.

It is great that the authors changed the error metrics according to the suggestion from R1. I did not notice the problem of the error metric in the first-round review.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Author's response to the Reviewer's comments

We thank the Editor and all three reviewers for the time they took to review our manuscript and for their thoughtful, smart, and stimulating comments. Our manuscript underwent a major revision. We have performed all additional experiments that have been suggested by the reviewers, added all text passages, explanations, and figures that the reviewers suggested (in particular for the further clarification of the evaluation algorithm), and addressed all other comments of the reviewers in this document or the main manuscript file. In particular, we have:

- Performed an additional quantitative experiment with real human subjects (see Rev. comments R1.1-R1.3, R2.1, R3.5)
- Extended the description of our algorithmic evaluation pipeline (see R1.4, R2.3).
- Performed an additional quantitative experiment which proves that the accuracy and precision of our method indeed improve with the number of measured surface points (R1.2, R2.2)

The submitted “clean” version of our revised manuscript is accompanied by a “red-lined version” that marks everything we have added during the major revision.

In the following, we address the comments reviewer by reviewer, point by point. All mentioned text passages, equations, figures, etc. are references wrt. the revised version of the manuscript.

Small errors, mistakes, and typos in our initial manuscript we have discovered during the revision (errors have just been corrected and not marked in the red-lined version):

- 1) Typos, such as “calbiration”
- 2) We have discovered a small error in our previous code. Correcting this error changed the real-world experiments on the model eye (Tab1) from

Rotation position a	-4°	-2°	0°	2°	4°
Mean relative error ϵ_0	0.14°	0.25°	0°	0.13°	0.12°
Precision σ_a	0.08°	0.06°	0.04°	0.07°	0.02°

to

Rotation position a	-4°	-2°	0°	2°	4°
Mean relative error ϵ_0	0.11°	0.12°	0°	0.03°	0.10°
Precision σ_a	0.08°	0.06°	0.04°	0.07°	0.02°

The results are largely similar, but the most prominent change is the smaller mean relative error value for -2° , which changed from 0.25° to 0.12° . It is also evident that the new error values now show greater similarity, which lets us consider the previous value of 0.25° as an outlier caused by the slight code error. The precision values did not change.

- 3) Section 3.1.: The $7\mu\text{m}$ value of the standard deviation of the bearing ball center was a typo. It was initially $70\mu\text{m}$. After the code error correction described above, the evaluated radius and standard deviation of the bearing ball changed from 11.92mm and $70\mu\text{m}$ to 12.02mm and $62\mu\text{m}$.
- 4) Not an error, but still an unmarked change: We switched the previous images from “qualitative measurements on the human eye” with new images that we took from our new experiments from “quantitative measurements on human eyes”. This concerns Fig.1c and Fig.7b.

Our answers to the reviewers' comments, point by point:

We were very happy to read that the reviewers think that our paper is novel (R3), and well-structured and consider it as a significant contribution to eye-tracking (R2), and think that the presented contributions are interesting and well-developed (R1)

We are convinced that the major revision has greatly improved our manuscript, and we thank the reviewers again for their contribution to this improvement.

Reviewer 1 (R1):

We appreciate R1's detailed and thoughtful feedback on our work. We appreciate that R1 recognizes the novelty of our work. We address R1's comments point by point in the following.

R1.1) "The apparent closeness of the stereo rig and of the illumination screen, large at this distance - a large portion of subject field of view is occluded by this structured illumination source - makes it difficult to realize if this eyetracking method is indeed feasible in a normal or controlled environment such as a lab with a chin and headrest. It would have been interesting to add a more realistic distant structured source of illumination (e.g. of same size and same distance as most eyetrackers) to evaluate how a reduced amount of points maintain good 'precision'; which might indeed be the case. Hence a measurement of the variations of RMS S2S precision and 2-steps 9 or 13 points accuracy on few human eyes as a function of the reduction of the size-reflection points could bring insights on a potential feasibility of structured illumination source and PMD for eyetracking. Not to mentioned the wavelength used here, were tests using a NIR source would also bring more insights on feasibility."

Answer: We thank R1 for this very constructive comment. Following the suggestion of R1 (and all other reviewers), we performed quantitative experiments with 3 human subjects to demonstrate the feasibility of our novel approach in a realistic eye-tracking scenario (we appreciate R1 acknowledging that a limited number of subjects is sufficient). The experiments were performed in a lab environment, using a chin rest (see Fig. 7 in the revised manuscript). For our performed "2-steps 9 points accuracy" experiment, all subjects observed 9 stimuli on a computer monitor that was placed at a distance of around 600mm. The observed angular range was 30° in azimuth and 15° in elevation. As Deflectometry display and cameras need to be set up at the side of the eye to capture parts of both the cornea and sclera, the setup did not occlude the subject's field of view for this experiment. This means that the proposed technique and setup are indeed feasible in a normal and controlled environment (we admit that problems might occur for very large gaze angles, but similar restrictions apply for standard glint tracking where all glints need to be "located" on the cornea). The results of our experiments are described in sec. 3.3 and briefly summarized here. Depending on subject and point location, we obtained accuracies between 0.46° and 0.97° with the mean accuracy values of the three subjects being 0.76°, 0.65°, and 0.73°. The precision values have been evaluated between 0.07°

and 0.58° with the mean precision values of the three subjects being 0.28° , 0.35° , and 0.30° (see Fig. 7). To evaluate how precision and accuracy of our method degrade under a reduced amount of captured points (“glints”), we have performed an additional experiment where we artificially thin out the measured points and evaluate the gaze from a reduced amount. Although this more general experiment does not directly follow R1s suggestion to increase the display distance, we comment in **sec. 3.3** that a reduced amount of points can be caused by several factors, including occlusions, smaller displays, or increased display distance.

Regarding NIR: As this work presents an early (first-ever) proof-of-principle of our technique and our setup, we refrained from using NIR illumination and detection, mainly due to the additional technical effort which is not an inherent limitation of our method itself. For example, we would have had to use a switchable high resolution IR display and an IR sensitive camera (coverglass removal). However, using NIR is an interesting direction for a potential next-stage prototype: For example, as soon as the pattern engineering is finalized, a fixed-pattern display can be utilized which can be more easily designed in IR. IR illumination even has the potential to further improve our method: IR bandpass filters can be applied to the cameras, which block out most of the undesired “background light”, increasing the signal-to-background ratio in our camera images.

R1.2)” Secondly and most importantly, the main claim being that acquiring 43000 reflection surface points instead of 1-8 (typically 2) glints would lead to an increased precision, if not accuracy, it is hard to conceive how this method can compare to the Literature without using the same methodologies (2 step calibration-validation on 13 points for accuracy measurement) nor the same metrics: 2 step accuracy and RMS S2S precision. Authors chose to evaluate the accuracy of their system on model eye which is a necessary and instructive step but incomplete.”

Answer: We thank the R1 again for bringing this up, and for acknowledging that an evaluation on a model eye is also useful and necessary. As discussed in R1.1 (above), we have now implemented real eye experiments evaluation including the kappa calibration and validation on a few (three) subjects. Moreover, we have performed an experiment where we randomly thin out points in our measurement, showing that more points lead to higher accuracy and precision for our proposed technique.

R1.3)”Specifically regarding the way accuracy is usually measured in a 2-steps 9 or 13 points process, it is true this method is far from ideal. However:-the head is made as still as possible using chin and head rest, bite bar when possible there is indeed constant eye movements, called fixational eye movements, that prevents to consider the eye is not moving, but these eye positions during fixation on a calibration target are small (typ SD 0.15 0.25°) not normally distributed, (see Chericci Poletti 2012) BUT not far from normal distribution at the considered scale of measurement here. So the assumption that movement of the eye during fixation could be averaged during a short amount of time still hold true. (especially on small targets, see McCamy 2013, Thaler 2013). Regarding precision, the measurement here on model eye are interesting, but similarly should have been extended on human eye to be comparable to existing methods. The Root mean square sample to sample (RMS-S2S) deviation is a way to measure the precision of a tracking method suggested by Holmqvist and Nyström in 2011. Normative values

of RMS-S2S exist and would then allow to locate the potential of this method among the numerous existing ones.”

Answer: We greatly thank the reviewer for these additional comments and for sharing those literatures. It was very helpful for us and we have included them in our paper references. We have evaluated accuracy and precision of our real human subject experiments for each measured point for each human subject (see Fig. 7). Although the average accuracy and precision values are higher than for our model eye experiments, best accuracy and precision value of 0.46° and 0.07° , respectively (both from subject 3 for stimulus middle-left), display performance similar to our model eye measurement (although the absolute error was evaluated instead of the relative error), meaning that effects like eye jitter or head movement are indeed largely neglectable for these “best” measurements. We further discuss this topic in **sec. 3.3** of our revised manuscript.

R1.4)” A hidden or undescribed necessary pre-segmentation of the reflection pattern must have been done so that the Camera 1 only use the part of the reflection from the Cornea (Fig7a and 6c) or Camera 2 from the sclera (Fig7b and 6b) to get the corresponding reconstructions. The normals obtained from the limbus appear as outliers, but I doubt that almost half of the points (cornea or sclera) are also rejected as outliers: please describe the preprocessing if needed in a real time implementation of an actual human eye tracker.”

Answer: Thank you. This comment (and other comments from R2 and R3) made us aware that we need to describe our algorithmic reconstruction pipeline in greater detail. We added several describing paragraphs in **Sec 2.1.3**, and briefly answered the specific question about segmentation below:

We segment the cornea and sclera in our measurement by detecting the abrupt change in pattern frequency at the limbus region via 2D wavelet transform. The known geometry of our setup allows us to assume, that all reflections right to the detected limbus region are located on the cornea, while reflections left of the limbus region are located on the sclera (applies for the measurements on real human eyes, for the model eye, left and right is flipped as the device was placed on the other side of the eye). In the previous version of our manuscript, we additionally cropped the detected cornea and sclera region to a rectangle. This step was abandoned during the revision phase, and we now use the “full” pattern information from the cornea and sclera.

R1.5)” Finally, there must be a dead range, an asymmetrical tracking capability in the visual field where this method (where cameras are place to the side of an eye and not above or below) would show its limits. Would you consider positioning the camera along a vertical axis below the primary gaze position of a subject ? Please report these possibility in the discussion.”

Answer: There is indeed a “dead range” or “null space” of our method. This range is reached when the display reflection does not simultaneously cover the cornea and sclera anymore, as

we need information from the cornea and sclera to unambiguously reconstruct the gaze. We have included this point in the discussion in **sec 4**.

In principle, there is no preferred placement position of our device, meaning that a camera could theoretically indeed be placed along a vertical axis below the primary gaze position of a subject. Of course, it needs to be taken care that the display or camera does not block the FoV of the subject. However, R1 might allude to the fact that the lower eyelid occludes most of the sclera part below the cornea. This would indeed present a challenge to our approach. In this case, the display would need to be placed at a position that still allows to retrieve reflection information from the sclera, such as at the upper side part (we emphasize that the position of camera and display influence the location of the reflection, meaning the placement of the display gives us additional flexibility if the camera position should be fixed).

An interesting idea we plan to explore in the future is to use the display or a VR headset itself for eye tracking (see discussion in **sec 4**), meaning that the display would now be placed in front of the eye. This would again influence the setup geometry and camera placement, and placement of the camera along a vertical axis below the primary gaze position might become more feasible.

Reviewer 2 (R2):

We are grateful for the comprehensive and considerate comments from R2 on our manuscript. We are excited that R2 thinks that our paper is well-structured and considers it as a significant contribution to eye-tracking. We address R2's comments point by point in the following.

R2.1) "The main contribution of this paper lies in utilizing phase measuring deflectometry to estimate gaze direction. However, when evaluating the accuracy of the proposed gaze estimation method, the authors employed their own error metrics, which differ from those used in previous literature. The authors acknowledge the difficulty of comparing their method with others due to the variance in error metrics and evaluation procedures. How can we ensure that this method is superior to others in the literature if direct comparisons cannot be made? Moreover, even though it is reported that the error of current eye tracking methods varies within a range of 0.5° to 1°, this error range was also evaluated using different error metrics. If the authors cannot clearly demonstrate the improvement compared to state-of-the-art methods, it becomes difficult to convince readers of the novelty of this method."

Answer: We also thank R2 for this very important comment and refer to our answers to R1.1 - R1.3 for the sake of brevity, which address a similar concern. In short: We have now performed and implemented quantitative experiments with real human subjects. For the evaluation of these experiments, we have used the experimental protocol and error metrics suggested by R1, which is used by many papers in the field. We hope that this adequately addresses R2's point about making our results more comparable to others. Please also refer to our revised manuscript **Sec 3.3**.

R2.2) “In the introduction, the authors mentioned: “For those methods, it can be intuitively understood that more acquired 3D surface information about the eye generally leads to higher gaze reconstruction accuracy [14]”. This statement reflects an intuition about the potential relationship between the amount of 3D surface information and gaze reconstruction accuracy. In this research, the authors collected more than 40,000 3D surface points. It would be valuable and necessary to validate this assertion by varying the number of 3D information points and calculating gaze accuracy accordingly.”

Answer: Thank you also for this comment and the great suggestion that triggered us to include an additional experiment: As also described in our answers to R1.1 and R1.2, we have now quantitatively proven that the accuracy and precision of our method indeed improve with the number of acquired surface points (See sec 3.3. and specifically Tab. 3). In the experiment, we randomly thin out the acquired point clouds and calculate accuracy and precision for different numbers of surface points. A clear trend of degradation with decreasing point numbers is visible.

R2.3) “The authors should provide further explanation regarding the shape reconstruction part: How to overcome the depth-normal-ambiguity problem by utilizing only the overlapping region between two cameras? Additionally, the authors should include more detailed explanations and figures to describe the proposed surface reconstruction method.”

Answer: Thank you for also making us aware that our initial description of algorithmic steps is not sufficient to completely understand our procedure. We have now significantly expanded our description of our reconstruction algorithm and also added new figures that describe our reconstruction pipeline. We hope that this revision largely addresses R2’s question.

A short high-level description (more in **Sec2.2**): The overlap region is the only region where an unambiguous evaluation of depth value and normal is possible without making any assumptions about the shape (basic idea behind Stereo Deflectometry). If a good initial estimate of the surface shape exists, overlap and stereo deflectometry is not required. In our reconstruction, we use the retrieved unique shape and normal map in the overlap region (from stereo deflectometry), to “build” a good initial estimate of the eye surface, using the 2-sphere assumption. Eventually, we expand the normal evaluation to the regions outside the overlap region, using our “good” initial surface estimate. After completing the reconstruction and an additional refinement step, we have obtained the complete normal and shape information from the deflectometry measurement.

We are not sure if R2’s question also was about the depth-normal ambiguity problem of deflectometry itself. Hence, we include a brief explanation below and also explain how Stereo Deflectometry solves this problem:

The depth-normal ambiguity problem happens in single-camera deflectometry. Even if the corresponding points between camera and display are known, only the camera ray (reflection ray) given the camera calibration can be determined. The display emission ray (incident ray) remains unknown because the display pixel emits light in all directions. Therefore, all the points

along the camera ray can be potential surface points. When we connect them to the corresponding display pixel, we may have different surface normals. Adding the second camera can help provide one more set of camera-display correspondence. When both cameras are measuring the same surface point, their calculated surface normal must be identical. This means that only one of the possible normals along each camera ray “fits”, and the ambiguity problem is solved. This is called Stereo deflectometry invented in 2004. We refer to the important references in our paper.

R2.4) “In this paper, the authors also assume that the cornea and sclera can be approximated by two spheres with different radii. What if the authors were to use more accurate models, such as the prolate shape for the cornea, where the radius of curvature increases from the center to the periphery [1]? Would this lead to a greater reduction in gaze error or not?”

We thank R2 for this very good question and suggestion. As we now describe in the new more detailed explanation of our evaluation algorithm, we only assume the 2-sphere model in the first step of our reconstruction (“initial surface reconstruction”, see **sec. 2.2** and **Fig.3 a and b**). We mitigate the 2-sphere model in the subsequent “refinement step”, and only assume that the eye is rotational symmetric around its optical axis. This final assumption is already close to R2’s suggested alternative eye model.

On top of that, R2’s suggestion also has great value for potential future work. In the future, we might replace the initial 2-sphere assumption in the first “initial surface reconstruction” step with a more sophisticated model. This might lead to a more realistic initial surface reconstruction, which can improve the performance after the refinement step. We have included a related discussion in section **sec.4**.

R2.5) “The authors use the screen in the off-state as a mirror to calibrate display position, which might introduce errors due to the non-perfect planarity of the screen. How could this affect the calibration accuracy?”

Answer: Yes, the non-perfect planarity of the screen will indeed affect the calibration accuracy. If the calibration screen is not perfectly planar (e.g., slightly curved) the camera rays that are reflected at this curved parts will land on slightly different Deflectometry display pixels, which introduces a residual curvature due to this calibration error. This curvature affects the evaluation of the depth and the normal value.

The next easier calibration procedure would be the “classical” method that we briefly describe in the manuscript: Using a planar mirror (can be manufactured with very good flatness!) and apply markers on the edges. For this procedure, the errors mostly come from marker displacement or mislocalization, which leads to the calculation of a wrong tilt for the mirror. This introduces a shift in all pixels in the camera image. Moreover, the edge regions of the FoV can not be calibrated.

During the development phase of our calibration methods, we found that our current method (with the phone screen) gives by far the best results, even if imperfections of the screen might

cause slight errors. A typical reprojection error of our calibration is below 0.5 pixels, which is a fairly good value for this kind of setup geometry. This indicates that a normal phone screen is indeed flat enough for our purpose.

R2.6)“The authors mentioned a high-precision rotation stage in the experiment of gaze estimation. The authors should specify the model and precision level of that rotation stage.”

Answer: Thank you for bringing this up. The model of rotation stage is ‘Thorlabs TTR001’ with rotation resolution at 0.03°. We also included this information in the revised manuscript.

Reviewer 3 (R3):

We are glad that R3 recognizes the novelty of our work and thinks our paper is well written. We appreciate R3’s insightful comments and feedback. We address R3’s comments point by point in the following.

R3.1) The proposed method mainly relies on the previous method [37] for single-shot deflectometry, and [36] for the stereo deflectometry. Therefore, the technical novelty of the proposed method is limited. However, I think applying these techniques to eye tracking systems is a novelty. Taken together, I think it is one of the paper’s strengths.

Answer: Thank you for recognizing the novelty of applying Deflectometry to eye-tracking. We fully agree with R3 that Deflectometry (including stereo-deflectometry and single-shot deflectometry) is not new and has been demonstrated by the optical metrology community before. Applying these techniques to eye-tracking systems is indeed a novelty.

Thanks to the numerous comments of reviewers about our reconstruction, we rewrote the respective section and described our registration algorithm in greater detail. This allowed us to emphasize, that our reconstruction algorithm is distinctively different from the “standard” deflectometry reconstruction algorithms, and hence can also be considered as a novelty. We refer to our descriptions in R1.4, R2.3, and **sec. 2.2**.

We should also mention that we would not define the “technical novelty” of our method only by the sheer measurement method Deflectometry (which is only a part of the whole technique), rather than by the complete pipeline, that goes from setup design, system calibration, measurement, surface evaluation and gaze evaluation.

A last comment: Although Deflectometry is a very powerful measurement principle, it is not well known outside the optical metrology community. Bringing Deflectometry to a new field and drawing the attention of the diverse readership of Nature Communications can help to disseminate this powerful technique to other communities so that everybody can profit.

R.3.2) From my understanding, the main drawback of the proposed method is that it requires the projection from the display on both the cornea and sclera, which is very difficult to satisfy when the person looking away from 0 degrees. The author only evaluates the small gaze angle [-4, 4] degrees in the experiments section 3.2. I suspect the proposed method will not work for large gaze angles. It will strongly constrain the proposed method to be used for the eye tracking system. Although this requirement can be fulfilled with more cameras from different angles, this limitation is not been discussed in the paper.

Answer: Thank you for this very good comment. It made us aware that the reader could think that our method only works in a very limited range [-4°,4°]. As described in R1.1, R2.1, this and other comments triggered us to perform an additional quantitative experiment on 3 human subjects. For these experiments we picked a much larger range of 30° azimuth and 15° elevation to show that our method also works for larger gaze angles.

Where is the limit? R3 stated correctly that display reflections have to be seen on cornea and sclera for our method to work. This “coverage” of the display reflection can be increased by a larger or closer display, multiple cameras, etc. We have added a few sentences in **sec. 4** to make the reader more aware of these tradeoffs.

R.3.3) Another drawback of this method is that it requires highly accurate camera calibration. Conventional eye tracking also needs accurate camera calibration. However, these conventional methods are much more robust than the proposed method.

Answer: We agree that “passive” methods (i.e., methods where only one or more cameras observe the eye, potentially under IR flood illumination) have less calibration effort and might be more robust, as only the cameras need to be calibrated. We have added a few sentences in **sec.4** to discuss this topic.

We are not sure if this also holds true for other “active” methods, e.g., glint tracking. In order to obtain the very best performance, we assume that the position of the glint point light sources would need to be calibrated as well, which is of similar effort as calibrating the positions of pixels on a display. We also emphasize that the position of all components in our system remains fixed once calibrated, meaning that the system does not have to be recalibrated frequently.

R.3.4) For the camera calibration method in section 2.3, there is a risk that the calibration pattern on the “off-state” is not clearly visible for cameras. The “off-state” phone is a black mirror that may not have the clear and shape reflection of the display. The author could show the image examples captured by the camera to show how good the reflected calibration pattern is.

Answer: Thank you for bringing this up. We have added captured images in the revised manuscript (see Fig. 4, upper right corner of (c) and (d)). Also: Even if we use a “black mirror”

during calibration (as R3 correctly notices), we can pick a large exposure time to obtain a good signal, as the calibration is not time-sensitive (the measurement of a moving eye is very time-sensitive).

R.3.5) The author did not perform the gaze estimation experiments with real human users. From my point of view, it is a pity that the author did not conduct the experiment, or compare it with conventional methods. For the first paragraph in section 3.2, I agree with the author for the first argument that using a third-party eye tracker could have ground-truth errors. However, I do not agree with the author that they cannot use the screen to calculate the ground-truth gaze direction. The experiment in section 3.2 with the eye model cannot reflect the real gaze estimation accuracy. For example, the eyeball does not have the eyelid occlusion.

If the author insists on not conducting the gaze estimation evaluation with real human eyes, at least, they can use the conventional method for the comparison. By checking the example in Fig. 6, I think the conventional pupil and iris detection methods may also achieve comparable accuracy with the proposed method. In this case, the author needs to provide good lighting conditions for these conventional methods for a fair comparison.

I understand the current system is just a prototype that cannot be integrated into an XR helmet. Nonetheless, I feel very pity that the author does not conduct gaze estimation experiments with real human eyes for measurement, especially compared with conventional methods.

Answer: We also greatly thank R3 for encouraging us to perform a quantitative experiment on real human eyes, but also for acknowledging again the problem of third party errors. As described above, we have now performed a quantitative experiment on real human eyes, by instructing subjects resting their heads in the chinrest and staring at targets on the computer screen. We included the experiment in the revised Sec. 3.3. We refer to our answers to R1.1-1.3, R2.1, as well as Sec. 3.3 for details.

In short: We have performed a “2-step 9 points” experiment with human subjects suggested by R1. Depending on subject and point location, we obtained accuracies between 0.46° and 0.97° with the mean accuracy values of the three subjects being 0.76° , 0.65° , and 0.73° . The precision values have been evaluated between 0.07° and 0.58° with the mean precision values of the three subjects being 0.28° , 0.35° , and 0.30° (see Fig. 7). This means that the best accuracy and precision values are indeed comparable with our results from the model eye, meaning that the respective measurements might have been free from third party errors like head movement, etc.

R.3.6) The author motivates the single-shot solution (section 2.1.2) with the delay of the sequential image capture solution (section 2.1.1). I am curious to know if we lose any reconstruction accuracy with the single-shot solution compared to the sequential image capture solution. Assume we have a camera with a high frame per second, can we just use the sequential image capture solution?

Answer: Very good and sharp question! Single-shot fringe evaluation procedures like Fourier Transform Profilometry [1] or single-shot Deflectometry [2] typically perform poorly at high

object frequencies, e.g., at surface discontinuities (steps). The performance on smooth surface parts is comparable to multi-shot procedures. The use of single-shot is justified in our case, as the eye has a mostly smooth surface, although the poor performance at high surface frequencies is one of the reasons (not the only one) why we do not consider the data from the limbus region.

Single-shot vs fast multi-shot: The exposure time for fast multi-shot would need to be decreased drastically. Assuming the same light source brightness, this would lead to lower SNR which also negatively affects the accuracy. Moreover, the standard multi-shot fringe procedures (e.g., the 4 phase shift algorithm) are very sensitive to even small movements between the phase shifts. These movements often lead to large phase errors and large surface reconstruction errors respectively.

As we mention in **sec.4**, another big benefit of single-shot is that we can use a fixed pattern (no switchable display necessary!). Theoretically, we can use a fixed backlit display (also IR possible), or we could even cover the inner surfaces of a VR headset with a fixed pattern (or course, this would result in additional engineering and calibration effort).

[1] M. Takeda and K. Mutoh, “Fourier transform profilometry for the automatic measurement of 3-d object shapes,” Appl. optics 22, 3977–3982 (1983).

[2] H. Liang, T. Sauer, and C. Faber, “Using wavelet transform to evaluate single-shot phase measuring deflectometry data,” in Applications of Digital Image Processing XLIII, vol. 11510 (SPIE, 2020), pp. 404–410.

R.3.7) To save the computation, the author picked a few sparse “anchor points” for the calculation (section 2.1.3). Does it mean that the author did not use all of the 40,000 surface points as claimed in the introduction section? The author needs to clarify how many surface points they utilized.

Answer: Thank you for this important comment about potential confusion. As we now specify in greater detail in the reworked description of our evaluation pipeline (see sec. 2.2), the anchor points are only used to provide a first guess of surface and normals in the overlap region. For the shown results, we roughly use 500 points in the overlap region as anchor points. The initial guess is used to fit a first 2-sphere model, which is later optimized using all (>40,000) captured data points. Eventually, the 2-sphere model is mitigated and the model is further refined, using again all captured data points.

We have now specified this point in our reworked section and hope that everything is now clearer to the reader.

R.3.8) The author uses the normal vector to calculate the cornea center and sclera center as described in section 2.2. How many normal vectors they usually can obtain?

Answer: For each acquired surface point, we can obtain its normal. For a typical measurement with our setup, we have > 40k evaluated normals.

Author's response to the Reviewer's comments

We thank the editor and all three reviewers again for the time they took to review our manuscripts and for their thoughtful, smart, and stimulating comments. We are glad to see that all reviewers acknowledged our first revision and that all reviewers showed their endorsement for publishing our paper. We have read the second round of reviews with great excitement and happily included or answered all additional reviewer comments. In the following, we address the comments reviewer by reviewer, point by point. All mentioned text passages, equations, figures, etc. are references wrt. the new revised version of the manuscript (2nd revision, submitted together with this document). Passages in the manuscript that have been changed/modified wrt the first revision are marked in blue.

Reviewer 1 (R1):

We appreciate R1's detailed and thoughtful feedback on our work. We are excited to hear that R1 endorses our work for publication. We are very happy about R1's additional questions and comments as they helped us again to significantly improve the clarity of the paper. We address R1's additional comments point by point in the following.

R1.1) In a response to my concerns regarding the large occlusions induced by field of view of the deflectometry display, you replied that "the setup did not occlude the subject's field of view for this experiment" which is not true, just by looking at Fig 7a. Yes, it does not occlude the 30x15° viewed display, but considering a majority of applications, and notably VR headsets that you consider, this might be an issue. That deflectometry display is probably larger than 30x15° in the lower right field.

Answer: R1 is correct. We have not fully understood R1's previous comment about this and we are grateful for this additional clarification. For large FoVs that are not limited to the measured 30°x15° range, our setup will indeed occlude parts of the FoV, especially in the current configuration (without the potential future improvements mentioned below). We have edited the text in the discussion accordingly: We note that typical VR/AR applications require FoVs that are even larger than our tested 30°x15° range, and we discuss the fact that our current setup concept could lead to occlusion problems, especially for larger or closer displays.

R1.2) "This means that the proposed technique and setup are indeed feasible in a normal and controlled environment (we admit that problems might occur for very large gaze angles, but similar restrictions apply for standard glint tracking where all glints need to be "located" on the cornea)". Well Tobii as previous researchers have developed an eyetracking glass where the IR led are embedded in the frame around the eye; similar restrictions does not apply and alternative method to track the eye in a large FoV do exists.

You did mentioned however "fixed known back-lit patterns that could cover the interior of VR glasses, or exploiting the known content of the main (VR) display", which is good and important to mention if you consider an actual usage of such specific deflectometry tracking requiring a large *illumination* field.

Answer: We thank R1 also for this valuable clarification and agree with R1's comment about standard glint tracking: With the "smart" arrangement of glints (e.g., by embedding them in the frame) one can indeed ensure that some glints are always visible on the cornea. Inspired by this comment, we slightly extended our discussion about this topic in section 4.

R1.3) I am actually a bit concerned by the final relatively low precision and accuracy obtained whether with a rotated model eye or 3 human subjects.

-On a model eye, in table 2, the precision is close to the accuracy, and even sometime higher. An error of 0.1° at 4° is actually quite large for a static model eye.

-On human, the obtained accuracy [0.5 to 1°] is higher than what is typically obtained with an Eyelink across 9 fixation targets, typically [0.25 to 0.5°]. The value obtained are closer to what is found by a mid-range mid-frequency commercial eye tracker, such as find within VR displays.

I understand that no polynomial fitting is used but a 2 steps methods using an initial surface reconstruction step and a refinement stage. I would presume that further improvements might be obtained by working on such method, probably by incorporating a fit with the actual pixel position correspondence on screen.

So stating that "Although early stage, our first experiments already reach accuracy and precision values which are on par with the best current state-of-the-art methods. " is a largely incorrect statement.

Answer: Thank you also for this comment, which also unearthed an important "historical remnant" from the initial version of our paper: In our initial version (no real human eye experiment in this version), we calculated the model eye gaze error at a specific rotation position as the mean of all relative gaze errors at this position (wrt the 0° position). [Sidenote 1: We have also discovered an error in the mean value calculation at -2° , but this is now obsolete anyway because we changed our metric now → See next sentences]. In the real human eye experiment (1st revision), we calculated the gaze accuracy as RMSE w.r.t the ground truth gaze. To make the two experiments more comparable, we have now changed the error metric in the model eye experiment and evaluated the RMS of all relative gaze errors (w.r.t the 0° position). The values have changed slightly compared to the previous evaluation using the mean, and all values are now larger than the respective precision values (although only slightly larger for the -2° position). [Sidenote 2: Although the metrics are now more comparable between model eye and real eye experiments, we emphasize that they are still not identical. The model eye experiment evaluates relative errors (no ground truth for this experiment), while the human eye experiment evaluates absolute errors].

R1's observation that the precision on the model eye is close to the RMS of all relative gaze errors expresses that the statistical variations have a large contribution to the final results. Besides the statistical variation of our method itself, the large precision values might have also been caused by the design of our experiment: Consistent with standard practices in metrology, we performed the model eye experiments in a "take out, put in" fashion. This means that we have always actively rotated the model eye from one position to another position before we took one single measurement at this position, and repeated this step 20 times for each position. The remaining inaccuracies in the rotation stage and the fact that we rotated to the respective position "by hand" might have contributed to the large statistical variation. If we would have taken 20 measurements at each position while not rotating the eye "in and out," we assume that the

statistical variations would have been lower, as only the statistical variations of our evaluation process (e.g., due to random noise, variations in surface evaluation, etc.) would have contributed.

Real human eye:

We thank R1 again for reminding us about the Eyelink, which shows indeed very good performance across 9 fixation targets. We have included a few related Eyelink references in our bibliography (see [24,58]) and adjusted our statements accordingly so that our paper does not appear weaker than it is (as R1 points out correctly). We also thank R1 for pointing us towards the polynomial fitting approach. A modification that is tailored to our data might be indeed a promising approach that could help to further improve our results in future work. We have included a few remarks about this in our discussion section.

R1.4) In a long paragraph, the authors try to justify that i) their system provide performance similar performance than a model eye by saying that one specific location it is close to the model eye (which is not true, a factor of x1-x4 depending on subjects)

ii) then contradict themselves stating that "there is no obvious degradation in performance when comparing measurements at the center stimuli positions to those at the outer positions"

Which is also wrong both in term of mean accuracy (over the 3 subjects, per location) or mean precision: accuracy and precision appear to be best in central location.

Answer: We thank R1 for making us aware of possible contradictions and wrong statements in our article. We have removed the sentences stating that our human eye performance is close to the model eye. We agree with R1 that these sentences are misleading and not necessary (the numbers speak for themselves).

We are not sure if we correctly understand R1's statement about the degrading performance of the outer stimuli. We have calculated the mean accuracy and mean precision for each stimulus over all three subjects and measurements as follows:

calib	0.787°	calib
0.697°	0.684°	0.750°
calib	0.667°	calib

Mean Accuracy

0.353°	0.267°	0.340°
0.330°	0.273°	0.273°
0.340°	0.340°	0.293°

Mean Precision

R1 is correct that the middle value is always one of the best, but it is actually never THE best. The mean accuracy varies between 0.667° (middle, bottom) and 0.787° (middle, top) and the mean precision varies between 0.267° (middle, top) and 0.353° (left, top). This means that the maximal range (difference) of values is 0.12° for the mean accuracy and 0.086° for the mean

precision. Given that these differences are fairly small, and given the fact that the best value is not located in the center, we think it is justified to say that the performance does not substantially degrade for the outer stimuli. However, we made sure to emphasize again that this is only shown for our 30°x15° FoV on 3 subjects, meaning this statement might not be true anymore for larger FoVs.

If R1 is still not happy with this statement (e.g., because we have overlooked an important aspect in our argumentation above), we are happy to change it in the final version of our paper.

R1.5) Why is precision reported in the 4 corners of the 9 dot calibration chart, but not the accuracy ? The fact that it is missing lead us to think that it was of low accuracy, which might not be the case but is concerning. Please add these 4 values if possible, or add an explanation.

Answer: Thank you for the comment which makes it obvious to us that this point requires further clarification: We used a 3 by 3 stimulus pattern in our experiment and acquired data at all 9 stimulus points. For each subject, we obtain a bundle of 9 vectors, that represent the evaluated optical axis of the eye. The data from the 4 corners are used to calibrate the kappa angle, i.e., to obtain the transformation matrix from the estimated optical axis to the real visual axis. The remaining 5 points are used to evaluate the accuracy. We do not report accuracy values at the 4 corner points, as these 4 stimuli have been used to fit the calibration transformation. Hence, the RMSE at these points is lower than on the other 5 points that have not been used for calibration. In short: adding the 4 corner points would actually seemingly improve the results for our method, which is not a fair representation (to our opinion), as we used these points for fitting. Precision: As the precision only considers the statistical variations at each stimulus (free from a systematic offset), it is still valid to calculate a precision value in the 4 corners. The fitting has only influence on the systematic offset, not on the statistical variations.

We have slightly reshaped the respective text passages and italicized the text passage “For each stimulus *that was not used for calibration*, we calculate the RMSE....”. We hope that this makes our procedure now more clear to the reader. If not, we are again happy to make further changes.

We note that we are aware that the eye tracking literature also reports a multitude of alternative evaluation procedures. For example, [55] uses half of the points for calibration, and then calculates the accuracy from all points (which includes the point the calibration has been performed on).

R1.6) I acknowledge the measure of the number of glints on accuracy. The degradation for a randomly sampled subset of points is actually quite large but appears to plateau. (a figure would have been better in my opinion). A thinning of the 2D distribution would have provided an even worse effect on accuracy. Could you please discuss the effects that a smaller (noise) OR thinner amount of points (effect of deflectometry screen distance) would have on precision and accuracy and what it means in terms of application and restriction to VR/AR usage?

Answer: We are happy to hear that R1 acknowledges our additional experiment, and, following R1's suggestion, we have added a plot of the values in the paper that makes the plateau more obvious to the reader.

We are not sure what R1 concretely means by a "thinner" distribution, but we assume that R1 means a "smaller" screen reflection that covers less surface of the eye (in mm^2). Such a smaller screen reflection can be caused by an increased distance of the screen (as R1 states) or a smaller screen. In Deflectometry literature, we typically refer to this as reduced "angular coverage". A reduced angular coverage has indeed a negative effect on our results. We have discussed this effect in Sec. 4 of our first revision. Triggered by R1's comment, we have added a few more clarifying statements to our discussion to make related tradeoffs more obvious to the reader. We also made sure to specifically comment about AR/VR setups, as this tradeoff is critical in this application: Placing the deflectometry display close to the eye surface would provide better coverage but could obstruct the user's field of view more. A possible solution is to place the display further away and use a curved screen, along with additional cameras embedded in the headset frame to capture reflections from multiple angles (see our discussion in Sec. 4).

Reviewer 2 (R2):

We appreciate R2's thoughtful comments and valuable suggestions since the first round of review. We are very glad R2 acknowledges our revision and supports the publication of our paper.

R2.1) it would be beneficial if the authors could include a discussion on phone screen flatness, as mentioned in comment R2.5, in the revised manuscript.

Answer: Thank you for the suggestion, we have happily added a discussion about the influence of phone screen flatness. The topic is most relevant for the calibration of our system, which is why we added the discussion in Sec. 2.

Reviewer 3 (R3):

We appreciate R3's valuable comments and suggestions to help us improve our paper. We thank R3 for accepting our paper as it is for publication.