# WHAT IS THE PATIENT LOOKING AT? ROBUST GAZE-SCENE INTERSECTION UNDER FREE-VIEWING CONDITIONS

*Ahmed Al-Hindawi* [⋆,†] *Marcela P Vizcaychipi* [†,∫] *Yiannis Demiris* [1,⋆]

[⋆]Personal Robotics Laboratory, Imperial College London
[†] Magill Department of Anaesthesia, Chelsea & Westminster Hospital NHS Foundation Trust
[∫] Anaesthetics, Pain Medicine and Intensive Care, Imperial College London

## ABSTRACT

Locating the user's gaze in the scene, also known as Point of Regard (PoR) estimation, following gaze regression is important for many downstream tasks. Current techniques either require the user to wear and calibrate instruments, require significant pre-processing of the scene information, or place restrictions on user's head movements.

We propose a geometrically inspired algorithm that, despite its simplicity, provides high accuracy and $O(J)$ performance under a variety of challenging situations including sparse depth maps, high noise, and high dynamic parallax between the user and the scene camera. We demonstrate the utility of the proposed algorithm in regressing the PoR from scenes captured in the Intensive Care Unit (ICU) at Chelsea & Westminster Hospital NHS Foundation Trust [a].

***Index Terms*—** Eye Tracking, Point of Regard estimation, Gaze Cue estimation

## 1. INTRODUCTION

We are interested in the target of a user's gaze in the scene - often called the Point of Regard (PoR) or the gaze cue - given the user's location and gaze vector from a quasi-egocentric view. This is of interest in the continuous detection of the location of gaze targets in an eye tracked user in an environment where the scene view is offset/parallaxed from the user's perspective. Such environments have been deployed in hospitals but similar environments exist in other healthcare environments, vehicles, marketing, and human-robot interactions [1, 2, 3, 4, 5].

The location of the user's gaze in the scene provides cues that facilitate further analysis of internal cognitive states and attention behaviour. Cumulative gaze locations can then be used to generate a temporal heat map to indicate attention in objects in the scene or further analysed into scan paths [6, 7].

Fixed offset points of view, such as ones in eye-tracking glasses, resolve PoR estimation through calibration. This procedure creates a mapping between a user's eye direction and the location in the scene. This mapping, usually a linear regression mapping with learnable parameters, provides sufficient accuracy due to the low and fixed offset between the user's viewpoint and the scene camera's viewpoint [8, 9]. This calibration procedure is participant and session-specific and can be prone to error, drift and failure [9]. However, in situations where the user's head pose and the scene camera's pose are not static, and thus exhibit dynamic parallax, such mappings become highly non-linear and become intractable to find as existing mappings do not exist or, worse, produce incorrect results [10]. Thus, these solutions do not work if used in circumstances where the head-pose of the participant has a dyanmic offset from the scene camera, such as free-viewing conditions without instrumentation.

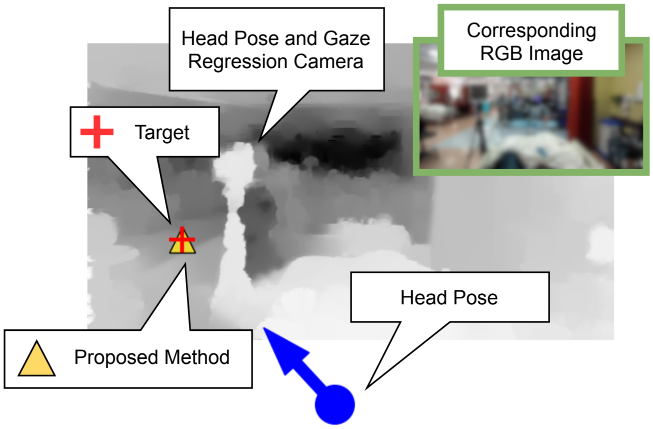


**Fig. 1**: An in-painted and rectified depth image with head position and gaze direction indicated as the blue circle and arrow respectively. The red cross is the target label with the detected gaze-scene intersection point labelled as a yellow triangle (RGB image blurred to preserve privacy).

If a scene depth map and a gaze vector are given *a-prior*, a brute-force solution is to project $\mathbb{R}^{u,v,d} \rightarrow \mathbb{R}^{x,y,z}$ using the camera's intrinsics matrix to a point cloud which can then be used to find the point closest to the gaze ray [11]. How-

[1] YD is supported by a Royal Academy of Engineering Chair in Emerging Technologies.

[a] Code available from www.imperial.ac.uk/personal-robotics/software/

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

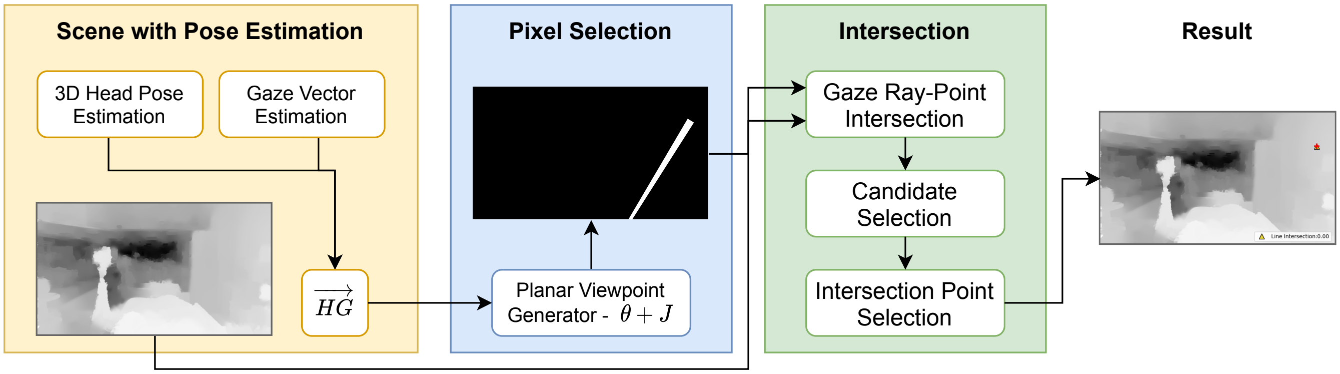


**Fig. 2**: Scene information including depth image (RGB image blurred) is provided where the Head Pose $H$ and Gaze vector $G$ are estimated. A planar cone is then rasterised starting at $H$ with direction $G$ with a central angle of $\theta$ and subsamples $J$. This rasterised image intersects the depth image. Gaze vector $\overrightarrow{HG}$ then intersects the set where candidate pixels are then classified whether they are inside the cone of gaze. The successful candidates are sorted and the intersection point selected.

ever, the step to convert the depth image into a point cloud and then initialising an appropriate dichotomisation scheme is computationally expensive. Thus, image-space techniques have been developed that are aimed at not using point clouds but using depth images directly.

Fang *et al.* used a synthetic depth map generated from a single RGB image to generate the scene's structure in an architecture called Dual Attention Guided Gaze Target Detection in the Wild (DAM) [12]. The depth map is reformulated to facilitate the intersection between the depth map and a generated Field of View (FOV) map. This intersection is then used as an input to two neural networks that classify the location of gaze vector in the scene [12]. The algorithm has two key hyper-parameters, the width ($\sigma$) of the FOV generator and the depth subsection parameter $\delta$ which were empirically set to 6 (representing $60°$ FOV) and $0.3$ respectively. Its reliance on synthetic depth data, and the requirement of the presence of the participant to be present in the frame, makes its immediate use unsuitable for use in an environment where the scene camera is located in a quasi-egocentric manner.

Our contributions are:

1. A geometrically inspired approach to scene dichotomisation that facilitates the calculation of the intersection between gaze vector.

2. Hyper-parameters that are intrinsically explainable and linked to the performance and accuracy of the gaze tracker.

3. A fixed performance envelope of $O(J)$ complexity using subsampling of $J$ points in the cone of gaze.

We demonstrate that this approach is accurate under sparse depth images and is robust to large parallax between the head pose and scene camera position from data obtained in a deployed gaze tracking, clinical, environment. Figure 1 demonstrates the results of the proposed architecture on an example scene.

## 2. METHOD

As per Figure 2, the first stage of the architecture requires the RGB and depth images, the head location and gaze direction of the participant in scene coordinates to be known. The second stage involves rendering a planar cone map emanating from the viewpoint of the participant which is parameterised by the accuracy of the gaze tracker. This map then subsets the depth map resulting in a small number of pixels tested to intersect with the gaze vector in the third stage. This third, and final stage, intersects each point with the gaze vector, classifies potential candidates as being inside or outside the gaze cone, and selects the most promising candidate to output.

Let $I_D \in \mathbb{R}^{1 \times h \times w}$ represent a optionally sparse and optionally rectified depth image with width $w$, height $h$ with a single channel representing depth. This map is to be acquired from a quasi-egocentric viewpoint with known camera intrinsics matrix $K$. The inverse of this matrix ($K^-1$) is used to convert world points to pixels in camera-space assuming a Pinhole camera model [13].

Let $\overrightarrow{\mathrm{HG}}$ be a vector that defines to start of the gaze at the headpose $H$ in coordinates relative to the scene-camera and with direction $G$. This is given *a-priori* from an eye tracker that also outputs the position of the head relative to the scene camera.

### 2.1. Planar Viewpoint Generator (PVG)

The PVG, parameterised by a central angle $\theta$ representing the gaze vector of the user from the perspective from the scene-camera starting at $\overrightarrow{HG}$ , rasterises a subset of the scene that is viewable by the participant into set $S_L$. Geometrically, $\theta$

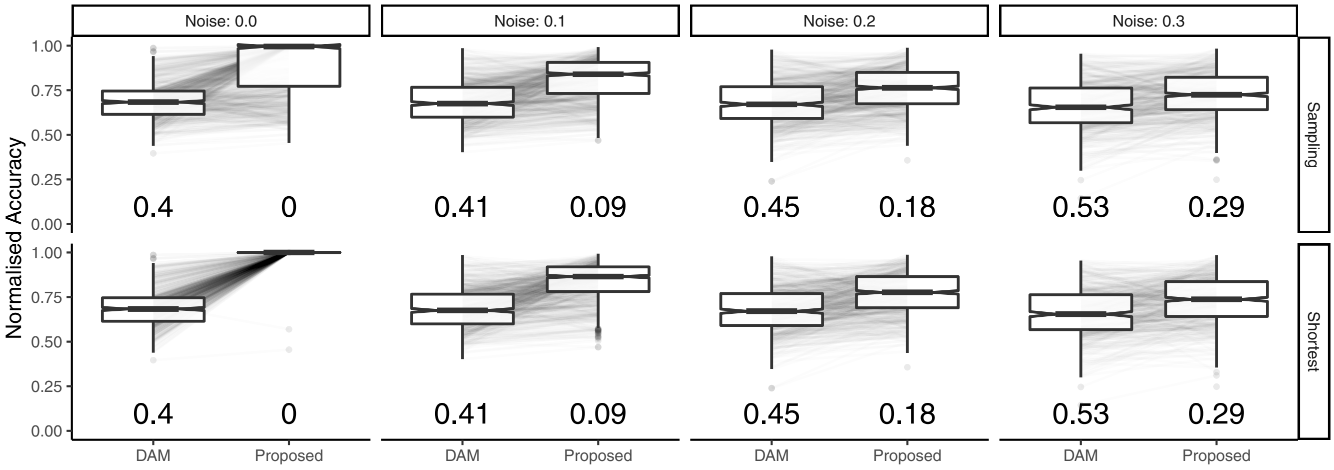


**Fig. 3**: The proposed algorithm's intersection point is compared against Dual Attention Module with increasing noise. Fractions below each plot represent the failures in each technique as a fraction of the overall number of runs. Notches in the box plot that do not overlap indicate statistical significance. Plots are split according to reduction strategy where the Sampling strategy $N$ was set to $5$.

represents the accuracy of the eye tracker. Then, $S_L$ is used to create a limited set of 3D world points such that:

$$S_L = \{\mathbb{N}_1, \mathbb{N}_2, ..., \mathbb{N}_m\} \subseteq I_D$$
$$I_d = I_D \cap S_L$$

where $\mathbb{N}_m$ is the index of $I_D$ and $m$ is the number of elements of the rasterised image and can be intersected with $I_D$ such that the resulting image $I_d$ is a subset of $I_D$.

As the number of elements in $I_d$ ultimately dictate the performance of the architecutre, rather than using the entire set, we instead choose to subsample $J$ points. A 2D Gaussian distribution centred on $H$ is drawn and multiplied by the result of the PVG and then sampled from. We empirically set the scale of the Gaussian distribution to be half the maximal depth, *i.e.* $^{\max(I_D)}/_{2.0}$. This ensure that points closer to $H$ are sampled at a higher density.

### 2.2. Intersection

Using a Pinhole camera model with intrisincs matrix $K$ [13], the world coordinates of pixels in $I_d$ can then be calculated:

$$W_L = \{K^{-1} \cdot [u, v, I_{D_{u,v}}]^T\} \quad \forall (u, v) \in I_d$$

where $u$ and $v$ are pixel indices and $I_{d_{u,v}}$ is the depth of that pixel at $u$,$v$. $W_L$ now contains a set of points in the same coordinate system as the gaze vector $\overrightarrow{HG}$. The distance between the gaze vector $\overrightarrow{\text{HG}}$ and each point $p$ in set $W_L$ can be found by first reformulating the vector $\overrightarrow{HG}$ to be a line in the form of $HG = H + tG$ where $t$ is a scalar that gives the locus of the line along direction vector $G$ and is set to $\max(I_d)$. The distance between $p$ and $\overrightarrow{HG}$ can thus be given as the normalised projection onto that line:

$$distance(p, HG) = \|(p - H) - ((p - H) \cdot G)G\|$$

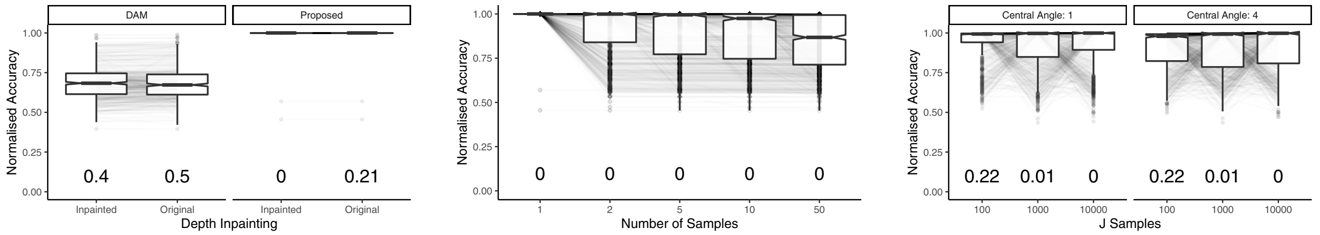
Each potential candidate is then classified as lying inside the cone of gaze into set $C_l$:

$$\{distance : distance \leq sin(\theta) \times \|p - H\|\} \quad \forall p \in W_L$$

Multiple candidates in $C_l$ could be the potential intersection point but only one is required. Several plausible reduction strategies exist, the simplest is to pick point $p$ with the smallest distance in set $C_l$. However, this has the potential effect of aliasing whereby a gaze vector *skimming* the actual intersection point to land on a much further location (*i.e.* a wall) which has a smaller distance. To tackle this potential problem, a further reduction strategy is explored where set $C_l$ is sorted in ascending order and the top $N$ candidates are selected. The point $p$ with the shortest $\|p - H\|_2$ is chosen as the intersection point potentially alleviating the aliasing issue. The second strategy collapses into the first strategy when $N = 1$.

## 3. DATASET

The above algorithm is validated using depth images acquired from a StereoLabs ZED2 camera in an Intensive Care Unit (ICU) setting at Chelsea & Westminster Hospital NHS Foundation Trust (HRA & REC Approval 20/LO/0162) [1]. DAM's esimated depth map was replaced with the actual



(a) Ablation study into the effect of depth inpainting on accuracy. Comparison intervals overlap and thus inpainting has no effect on accuracy but does reduce the number of failures.

(b) The number of potential candidates to sample from set $C_l$ relating to accuracy. The naive strategy has the highest accuracy under experimental conditions where increasing $N$ decreases accuracy.

(c) The number of samples $J$ tested against $\overrightarrow{HG}$ against normalised accuracy. As $J$ increases, the number of failure decreases.

**Fig. 4**: Accuracy of the proposed algorithm under various ablation conditions. Fractions under each box plot indicate algorithmic failure as a fraction of all runs.

depth map in the dataset. The resolution of the depth image is $1920 \times 1080$. The dataset consists of 23 recordings of depth from a quasi-egocentric view each lasting 10 minutes. The depth image of first frame is taken from each recording in this experiment. Due to lacking ground truth data of the PoR of each participant, for each recording, we chose to simulate head pose $H$ and gaze vector $G$.

Gaze target location in pixels was sampled randomly as $G_t \sim \mathcal{U}_{u,v}(0, L_{D_{width}} : L_{D_{height}})$ whilst the head-pose, located in world coordinates is sampled into pixel space by $H_{u,v} = K^{-1} \cdot [\sim \mathcal{N}(0, 0.2), \sim \mathcal{N}(1, 0.2), 1]^T$. These distributions are sufficiently wide enough to induce high random offsets between the gaze direction of the participant and the scene camera. Experimentally, increasing Gaussian noise ($\sim \mathcal{N}(0,1) \times diag(L_D)$) is added to $G_t$ to ascertain the reliability of the proposed algorithm against noise. Image locations where either algorithm failed were recorded. A random seed of 0 was used for reproducibility.

For each recording, $1,000$ $G_t$ gaze locations are sampled and processed to result in the proposed gaze-scene intersection point $p$. The Euclidean distance in pixels between $G_t$ and $p$ normalised by the maximal possible distance was taken as the accuracy $= 1 - {\|G_t - p\|_2}/{diag(L_D)}$ where $\| \cdot \|_2$ is the euclidean distance and $diag(\cdot)$ is the diagonal length of an image. This accuracy term can then be used to evaluate our proposed algorithm.

Comparison interval on box plots were calculated for statistical significance by $median \pm 1.57 \times {IQR}/{\sqrt{n}}$ where IQR is the inter-quartile range defined by the $25^{th}$ and $75^{th}$ percentiles and $n$ is the number of data points. Non-overlapping notches thus indicate non-parametric statistical significance.

## 4. RESULTS

Figure 3 demonstrate the accuracy of the proposed algorithm at increasing noise levels. At low levels of noise, the proposed algorithm provides near-perfect accuracy in a statistically significant manner which naturally converges to the accuracy achieved by [12] whilst also increasing the number of answered queries compared to DAM. The statistical significance is demonstrated by the non-intersection of Figure 3 median's notch.

### 4.1. Ablation Studies

Depth images often have 'holes' where depth estimation fails. Figure 4a demonstrates that in-painting does not increase the overall accuracy but does reduce the number of failures. Inpainting of the depth image is performed as per [14].

The number of candidates considered following classification ($N$) was experimentally validated. Figure 4b demonstrates that as the number of samples ($N$) increases, the normalised accuracy decreases. It is worth noting that the median increase of accuracy results from points increasing from areas of low accuracy to high accuracy, rather than a general decrease of accuracy. This suggests the simplest strategy of the closest point to vector $\overrightarrow{HG}$ is the one that provides the highest accuracy and thus aliasing is not a likely scenario in this dataset.

The number of samples $J$ dictates the number of intersections between $\overrightarrow{HG}$ and $p$ $\forall p \in W_l$ and the size of set $C_l$ which can be tuned for performance and accuracy. Figure 4c demonstrates the normalised accuracy achieved with increasing the number of $J$ samples across two cone angles. A $J$ value of 1000 is an optimal trade-off between performance and accuracy with only 1% of simulated runs not producing a result.

## 5. CONCLUSION

We presented an algorithm for the estimation of PoR that is accurate and robust under a variety of difficult situations in a quasi-egocentric manner. Extensive ablation studies demonstrate the importance of each sub-module while the architecture's hyper-parameters being derived from system requirements.

## 6. REFERENCES

[1] Ahmed Al-Hindawi, Marcela P Vizcaychipi, and Yiannis Demiris, "Continuous Non-Invasive Eye Tracking In Intensive Care," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Nov. 2021, pp. 1869–1873.

[2] Tobias Fischer and Yiannis Demiris, "Markerless perspective taking for humanoid robots in unconstrained environments," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 3309–3316.

[3] Christina Ohm, Manuel Müller, Bernd Ludwig, and Stefan Bienk, "Where is the Landmark? Eye Tracking Studies in Large-Scale Indoor Environments," in *2nd International Workshop on Eye Tracking for Spatial Research co-located with the 8th International Conference on Geographic Information Science (GIScience 2014)*, Vienna, Austria, 2014, vol. Vol-1241, pp. 47–51.

[4] Pierluigi Vito Amadori, Tobias Fischer, Ruohan Wang, and Yiannis Demiris, "Decision Anticipation for Driving Assistance Systems," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, Sept. 2020, pp. 1–7.

[5] Kevin Cortacero, Tobias Fischer, and Yiannis Demiris, "RT-BENE: A Dataset and Baselines for Real-Time Blink Estimation in Natural Environments," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[6] O. Špakov and D. Miniotas, "Visualization of Eye Gaze Data using Heat Maps," *Elektronika ir Elektrotechnika*, vol. 74, no. 2, pp. 55–58, Feb. 2007.

[7] Dirk Brockmann and Theo Geisel, "The ecology of gaze shifts," *Neurocomputing*, vol. 32–33, pp. 643–650, June 2000.

[8] Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer, "The influence of calibration method and eye physiology on eyetracking data quality," *Behavior Research Methods*, vol. 45, no. 1, pp. 272–288, Mar. 2013.

[9] C.H. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," in *2002 International Conference on Pattern Recognition*, Aug. 2002, vol. 4, pp. 314–317 vol.4.

[10] N. Ramanauskas, "Calibration of Video-Oculographical Eye-Tracking System," *Elektronika Ir Elektrotechnika*, vol. 72, no. 8, pp. 65–68, Oct. 2006.

[11] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu, "Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6801–6809.

[12] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai, "Dual Attention Guided Gaze Target Detection in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11390–11399.

[13] Kenneth M. Dawson-Howe and David Vernon, "Simple pinhole camera calibration," *International Journal of Imaging Systems and Technology*, vol. 5, no. 1, pp. 1–6, 1994.

[14] M. Bertalmio, A.L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, vol. 1, pp. I–355–I–362, IEEE Comput. Soc.