INTEL REALSENSE = REAL LOW COST GAZE

Mark Draelos, Qiang Qiu, †Alex Bronstein, and Guillermo Sapiro

Duke University, Durham, NC 27708, USA
†Tel Aviv University, Ramat Aviv 69978, Israel

ABSTRACT

Intel's newly-announced low-cost RealSense 3D camera claims significantly better precision than other currently available low-cost platforms and is expected to become ubiquitous in laptops and mobile devices starting this year. In this paper, we demonstrate for the first time that the RealSense camera can be easily converted into a real low-cost gaze tracker. Gaze has become increasingly relevant as an input for human-computer interaction due to its association with attention. It is also critical in clinical mental health diagnosis. We present a novel 3D gaze and fixation tracker based on the eye surface geometry captured with the RealSense 3D camera. First, eye surface 3D point clouds are segmented to extract the pupil center and iris using registered infrared images. With non-ellipsoid eye surface and single fixation point assumptions, pupil centers and iris normal vectors are used to first estimate gaze (for each eye), and then a single fixation point for both eyes simultaneously using a RANSAC-based approach. With a simple learned bias field correction model, the fixation tracker demonstrates mean error of approximately $1\,\text{cm}$ at $20 - 30\,\text{cm}$, which is sufficiently adequate for gaze and fixation tracking in human-computer interaction and mental health diagnosis applications.

Index Terms— gaze tracker, fixation tracker, depth camera, mental health, human-computer interaction

1. INTRODUCTION

Gaze provides a wealth of information for human-computer interaction, particularly as an indicator of attention in consumer (e.g., video games), commercial (e.g., advertising), and medical (e.g., autism screening [1, 2] and child emotion studies [3]) applications. In such applications, detecting gaze to a computer monitor quadrant, on-screen window, or relatively large body part is all that is needed. The advent of now ubiquitous infrared structured illumination depth cameras, such as Microsoft's Kinect and Intel's RealSense 3D camera [4], capable of imaging the eye's exterior surface, enable new gaze estimation techniques based on eye surface geometry. The newly announced RealSense 3D camera claims better short-range resolution than other low-cost platforms such as PrimeSense and is embedded into the screen lids of laptops and tablets. We demonstrate here that the RealSense 3D camera can be efficiently converted into a low-cost gaze tracker with our proposed simple yet non-trivial methods. Specifically, we consider 3D gaze estimation using infrared images and point clouds of the eye's scleral and iris surfaces acquired with RealSense.

Eye surface geometry is defined by the approximately spherical scleral and corneal surfaces (Fig. 1a). When imaging the human eye, however, infrared structured illumination depth cameras produce sclera and iris point clouds only because the cornea is transparent (Fig. 1b). The iris point cloud is of interest due its geometrical relationship with the eye's optical axis or gaze, which is the line passing through the fovea and the pupil's center. Notably, the iris is a shallow cone that surrounds the pupil and is oriented perpendicularly to the optical axis [5]. Thus, based on Fig. 1a, we propose a non-ellipsoidal eye surface model. If approximated as planar, the iris normal vector is parallel to the optical axis and consequently parallel to the eye's gaze. Therefore, the iris point cloud and the pupil center in 3D space provide sufficient information to estimate gaze.

Stated formally, let $\mathbf{p} \in \mathbb{R}^3$ represent the pupil center in the depth camera's reference frame and let $\hat{\mathbf{n}} \in \mathbb{R}^3$ represent the iris plane normal. The points lying along the eye's gaze are thus those satisfying

$$\mathbf{g} = \mathbf{p} + x\hat{\mathbf{n}}, \tag{1}$$

where $x \in \mathbb{R}$. As an extension, given $\mathbf{p}$ and $\hat{\mathbf{n}}$ for corresponding right and left eyes, the fixation target $\mathbf{f}$ is the point (or nearest point) of intersection of $\mathbf{g}_r$ and $\mathbf{g}_l$, where subscripts indicate right and left, respectively. Moreover, fixation (of both eyes concurrently) provides a convenient internal constraint during gaze estimation that can enhance precision of otherwise independent monocular gaze estimates. In specific applications, requiring the intersection to lie on a surface, such as a computer monitor or tablet computer screen, can further constrain gaze estimation.

Our contribution is a novel 3D gaze tracker that uses registered infrared and depth eye images to estimate gaze from eye surface geometry by computing $\mathbf{p}$ and $\hat{\mathbf{n}}$ in Eq. (1). This gaze tracking method is advantageous due to infrared structured illumination cameras' affordability, increasing availability,

Work supported by ONR, NGA, AFOSR, NSF, and ARO. We thank our team on the *Duke Information and Child Mental Health Initiative* for important feedback and comments. The work with that team partially motivated the analysis here reported. The authors would like to thank Intel for contributing a RealSense F200 depth camera to support this research.

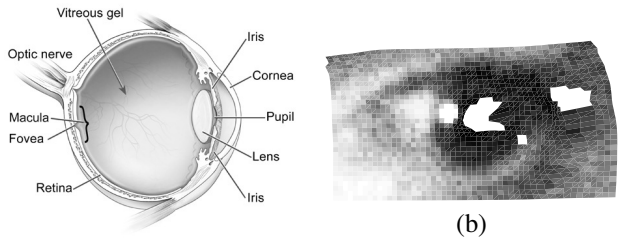


Fig. 1. (a) Eye cross-sectional anatomy [9], and (b) example eye exterior surface point cloud.

small form factor, and low power consumption. Unlike related techniques that employ 3D data in preprocessing steps [6, 7], require multiple camera units [6] or infer eye geometry from 2D images [8], our technique exploits direct measurement of eye surface geometry. Furthermore, we contribute a 3D fixation tracker that estimates the binocular fixation point on a known target surface using simultaneous intersecting gaze estimates from corresponding left and right eyes. We describe such a fixation target tracker and present an analysis of its performance.

2. METHODS

Given registered infrared images and point clouds, gaze estimation is performed in three phases to produce a final fixation estimate: infrared segmentation, point cloud fitting, and bias field correction. Infrared segmentation identifies pertinent point cloud regions for use in point cloud fitting, from which the final 3D gaze estimate is derived. The bias field correction step offsets fixation estimates to compensate for target surface and depth camera coordinate system variations. A computer monitor was chosen as the target surface for fixation estimation.

2.1. Infrared Segmentation

The goal of infrared segmentation is to identify the iris point cloud region for later iris plane normal estimation. The algorithm first finds the pupil region and then uses that blob to find the iris region.

Pupil detection: Pupil detection exploits poor reflection of infrared illumination (860 nm as used in RealSense) back through the pupil, which produces a dark pupil in infrared images. Given an infrared eye image, pupil detection identifies the largest dark blob as the pupil through thresholding and connected-components analysis (Fig. 2). Histogram equalization and discarding of all pixels with intensity greater than 50% is performed to increase the intensity magnitude between the dark pupil and the iris. All pixels with intensities less than 10% of the image's intensity range are considered candidate pupil pixels. Successive morphological erosion and dilation are applied to remove isolated pixels not part of the pupil region. Dilation is applied twice to include the pupil-iris boundary within the pupil region. The largest 8-connected blob is then marked as the pupil. Successive erosion and dilation with a larger structuring element is used to fill any glints within the pupil region.

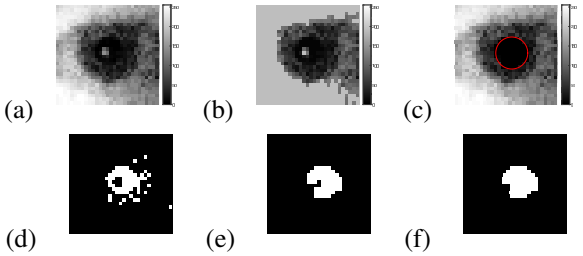


Fig. 2. Examples of pupil detection steps: (a) input image, (b) 50% thresholded histogram equalization, (c) final output with pupil region suppressed and circled (red), (d) 10% intensity thresholding, (e) erosion/dilation and area filtering, and (f) glint filling.

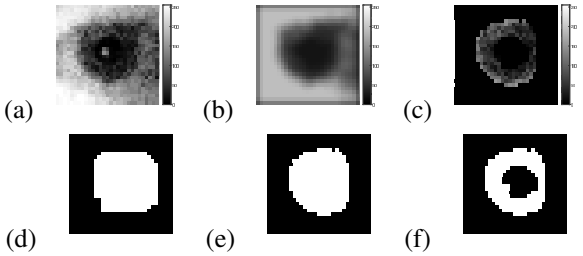
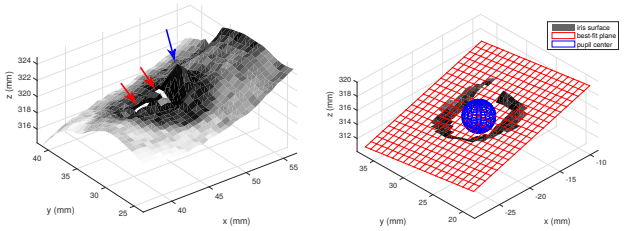


Fig. 3. Examples of iris segmentation steps: (a) input image, (b) gradient image for active contours, (c) final output, (d) initial active contour mask from inflated pupil region, (e) two-pass active contours output, and (f) iris segmentation.

Iris segmentation: Iris segmentation exploits the pupil-iris topographic relationship to expand the pupil region to include the iris using active contours [10]. Preprocessing with histogram equalization as in pupil detection is first performed. A blur is applied to reduce pixel intensity noise, particularly within the iris region, for later active contour evolution. Active contour segmentation is performed in two identical passes starting with the inflated pupil region as the initial mask and using a negative contraction bias. This causes the initial active contour to expand from the pupil region to the next hard edge: the iris-sclera or iris-eyelid edge. An increased smoothness factor is specified to promote segmentation of an approximately elliptical iris region. Between the first and second active contour passes, all pixels with intensities greater than 75% are suppressed to reduce influence of glints within the iris region on active contour evolution. The active contours initial mask is the pupil region inflated to ensure the initial active contour is between the pupil-iris and iris-sclera/iris-eyelid boundary; otherwise, the active contour may stabilize on a sharp iris-pupil edge rather than the desired iris-sclera/iris-eyelid boundary.



(a) (b)

Fig. 4. (a) Eye point cloud surface colored with infrared image showing distorted position near the pupil (blue arrow) and gaps within the pupil (red arrow). (b) Best-fit plane (red) and pupil mapping (blue) for iris region point cloud surface.

2.2. Point Cloud Fitting

Once the pupil and iris segmentation is complete, point cloud fitting is performed to estimate the eye's gaze using a RANSAC-based approach. This phase of processing relies upon the previously described non-ellipsoid eye surface model with a planar iris and the binocular fixation constraint that requires fixation estimates to lie on a target surface (the monitor plane in this case). First, the pupil is mapped from infrared image coordinates into the point cloud to localize the pupil search space. Second, a RANSAC-based approach evaluates potential fixations targets while allowing for small variation in pupil positioning.

Plane-based pupil mapping: Despite correspondence between a given infrared image and its derived point cloud, mapping the 2D pupil center image coordinates into the point cloud is nontrivial resulting from poor infrared reflectivity through the pupil. Consequently, the pupil region in the point cloud is frequently filled with invalid or distorted depth values as shown in Fig. 4a. Thus, interpolation is required for x, y, and z coordinates of the eye point cloud. First, the best-fit planes for the eye point cloud x and y components as a function of image coordinates are computed separately. The pupil center image coordinates are then mapped to x and y spatial coordinates by evaluating the best-fit plane at the pupil center. Using a best-fit plane helps reject the influence of distorted points like those in Fig. 4a on x and y interpolation. Next, the z spatial coordinates as a function of x and y spatial coordinates are interpolated using the iris region point cloud best-fit plane which relies upon the prior non-ellipsoidal eye surface model. The pupil center x and y spatial coordinates are then mapped to z spatial coordinates by evaluating the iris best-fit plane at the pupil center, yielding the full pupil center spatial coordinates (Fig. 4b).

Joint binocular fixation estimation: As previously stated, binocular fixation (of both eyes) provides a convenient internal constraint during gaze estimation to improve monocular (independent) gaze estimates. In applications involving devices such as laptops and tablets, requiring that the fixation estimate lie on a surface further constrains gaze estimation. Consequently, fixation target estimation uses the pupil center and fixation target search spaces to estimate the binocular fixation target, which is the point where the optical axes of both eyes intersect (Alg. 1). First, one candidate fixation target and two candidate pupil centers, one for each eye, are randomly chosen from their respective search spaces. The fixation target search space is specified in advance assuming the fixation target lies on a target surface. The pupil center search space is a small box centered on the estimated pupil center for each eye and provides additional degrees of freedom. Next, a candidate iris plane normal is computed from the gaze vector from each candidate pupil center to the candidate fixation target. If more than 5% of total iris region points lie within a 0.5 mm normal displacement from the iris plane, the error metric is computed, which is the sum of squared normal displacements for all iris points from their respective planes. The sample with the least error metric across 100,000 iterations is considered the best fixation estimate (Fig. 5). In short, fixation estimation applies RANSAC [11] to randomly sample the pupil center and fixation target search spaces and then computes iris plane inliers and uses squared normal distances as error metrics.

input : iris points $\{\mathbf{i}_r\}$ and $\{\mathbf{i}_l\}$
output: best fix
best error $\leftarrow \infty$;
for $n \leftarrow 1$ **to** 100000 **do**
  randomly select fixation target $\mathbf{f}$, right and left pupil positions $\mathbf{p}_r$ and $\mathbf{p}_l$ from search spaces;
  gaze $\mathbf{g}_r = \mathbf{f} - \mathbf{p}_r$ and $\mathbf{g}_l$;
  plane normal distances $d_r = \mathbf{g}_r \cdot (\mathbf{i}_r - \mathbf{p}_r)$ and d_l;
  inliers $I = \{d_r || d_r| < T\} \cup \{d_l || d_l| < T\}$ where $T = 0.5$ mm;
  if $|I| > |d_r \cup d_l|/20$ **then**
    error $E = \sum d_r^2 + \sum d_l^2$;
    if $E <$ best error **then**
      best error $\leftarrow E$;
      best fix $\leftarrow \mathbf{f}$;
    end
  end
end

Algorithm 1: RANSAC-based fixation estimation algorithm.

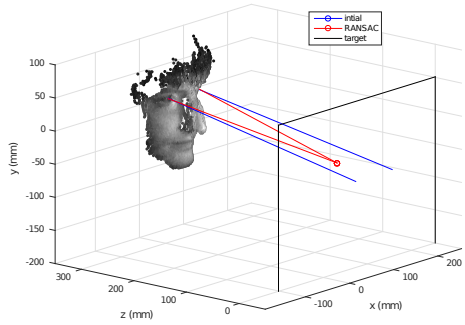


Fig. 5. Example point cloud depicting initial gaze estimates (blue) based on iris plane normals without the fixation constraint and the RANSAC fixation estimate (red) on the target surface (black).

Target	x (mm)	y (mm)	displacement (mm)
1	1.6 ± 16.1	-5.0 ± 6.9	8.8 ± 11.7
2	-8.6 ± 8.4	-6.3 ± 12.1	12.5 ± 10.8
3	0.2 ± 0.5	-2.3 ± 6.8	5.0 ± 9.2
4	9.5 ± 21.2	0.5 ± 1.0	6.3 ± 15.2

Table 1. Mean and standard deviation of x and y error and displacement from ground truth for fixation estimation examples in Fig. 6.

2.3. Bias Field Correction

As introduced above, the bias field correction step offsets raw fixation estimates to compensate for target surface and depth camera coordinate system variations. Because this work's focus is on the RealSense 3D camera's capability, the correction adopted here is simply a spatial displacement added to raw fixation estimates to form final fixation estimates. Bias offsets are derived from training fixation observations of fixation targets. The x and y errors for each training observation are considered to be samples of a bias field $\mathbf{C}(x, y)$ at the training observation location. The specific bias field $\mathbf{C}(x, y)$ is then learned using these observations. For this work, $\mathbf{C}(x, y)$ is linearly interpolated using the training dataset, although more sophisticated techniques such as random forests used in [6] will further improve performance. Ultimately, $\mathbf{C}(x, y)$ provides an additive offset for each test dataset raw fixation estimate such that $\mathbf{f}' = \mathbf{f} + \mathbf{C}(f_x, f_y)$, where $\mathbf{f}'$ is the final fixation estimate and $\mathbf{f}$ is the raw fixation estimate. Notably, camera calibration is skipped as reasonable performance is obtained without it.

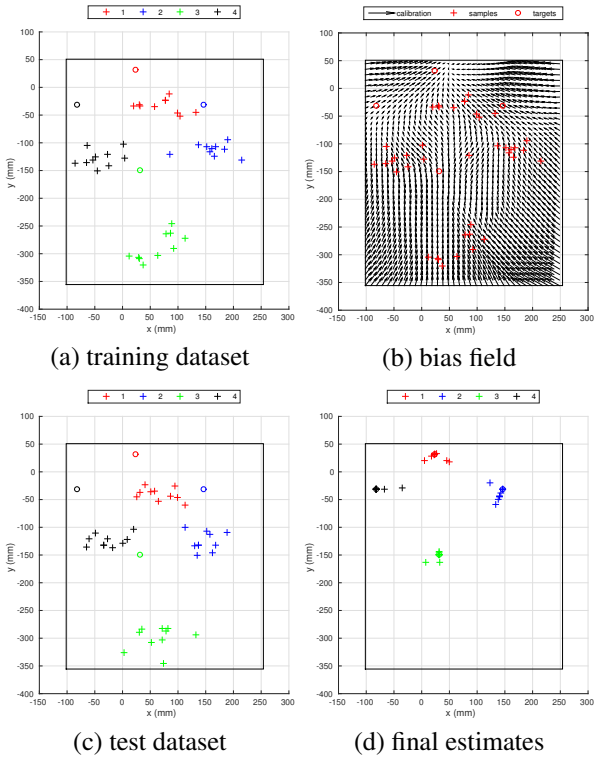
3. EXPERIMENTAL VALIDATION

3.1. Performance Evaluation

Fixation estimate accuracy and precision are evaluated using six metrics. These metrics consider the mean and standard deviation of the error in the x and y directions as well as displacement $r_e = \sqrt{x_e^2 + y_e^2}$ from the ground truth position along the target surface (e.g., computer screen), where x_e and y_e are the x and y error components, respectively.

3.2. Results

Corresponding infrared and depth images were taken at random fixation target positions in sets of ten images. Each set of ten images was captured successively without any intervening discarded frames. The human subject oriented their face towards the camera at a distance of $20 - 30$ cm, maintained binocular fixation on the specified target (a marker on the screen), and held the head still for capture of each ten-image set. This distance was chosen as it is representative of typical viewing distances for mobile devices (e.g., tablet computers) like those that the Intel RealSense 3D camera targets. Half of the dataset at each target was used for bias field training (Fig. 6a-b); the remaining half was used as the test dataset (Fig. 6c). Fig. 6d shows final fixation estimates for the test dataset. For purposes of illustration, Fig. 6 depicts four target locations from the captured dataset. Table 1 lists the associated mean and standard deviation of error for the x and y directions and the on-screen displacement.



(a) training dataset (b) bias field

(c) test dataset (d) final estimates

Fig. 6. Fixation estimate bias field computation and application within the fixation search space (black rectangle) at four example positions. (a) Training dataset fixation estimates (+) and targets (o) colored by target. (b) Bias field (black) derived from training dataset (red). (c) Test dataset fixation estimates (+) and targets (o) colored by target. (d) Test dataset fixation estimates corrected with training dataset.

4. CONCLUSION

We have demonstrated a low-cost 3D gaze and fixation tracker using the expected to be ubiquitous Intel RealSense 3D camera. On average, we achieve sub-centimeter mean fixation estimate error at typical usage distances. Although our algorithm will benefit from more sophisticated bias field correction techniques, the reported results have demonstrated the technique's feasibility and provide already sufficient accuracy for multiple important applications. Such applications include selection of a computer's active window, detection of the broad body part being observed in autism studies (e.g., face or hands), or detection of the quadrant being observed on a computer screen, with one stimulus in each, in anxiety studies.

5. REFERENCES

[1] W. Jones and A. Klin, "Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427–431, Dec 2013.

[2] J. Hashemi, T.V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro, "A computer vision approach for the assessment of autism-related behavioral markers," in *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, Nov 2012, pp. 1–7.

[3] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold, "The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): a new set of children's facial emotion stimuli," *Int J Methods Psychiatr Res*, vol. 20, no. 3, pp. 145–156, Sep 2011.

[4] Intel Corporation, "Intel RealSense 3D Camera," http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-depth-technologies.html, Jan. 2015, See also http://www.intel.com/content/www/us/en/events/intel-ces-keynote.html.

[5] P. Riordan-Eva, *Vaughan & Asbury's General Ophthalmology, 18e*, chapter Anatomy & Embryology of the Eye, The McGraw-Hill Companies, New York, NY, 2011.

[6] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3D gaze estimation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 1821–1828.

[7] K.A. Funes Mora and J. Odobez, "Gaze estimation from multimodal Kinect data," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, June 2012, pp. 25–30.

[8] H. Wu, Y. Kitagawa, T. Wada, T. Kato, and Q. Chen, "Tracking iris contour with a 3D eye-model for gaze estimation," in *Computer Vision ACCV 2007*, Yasushi Yagi, SingBing Kang, InSo Kweon, and Hongbin Zha, Eds., vol. 4843 of *Lecture Notes in Computer Science*, pp. 688–697. Springer Berlin Heidelberg, 2007.

[9] National Institues of Health, National Eye Institute, "Diagram of eye," https://www.nei.nih.gov/sites/default/files/nehep-images/eyediagram.gif, Jan. 2015.

[10] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal on Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, no. 6, pp. 381395, June 1981, doi:10.1145/358669.358692.