Low-cost Gaze and Pulse Analysis using RealSense *

Qiang Qiu
Duke University
Durham, NC 27707

Zhuoqing Chang
Duke University
Durham, NC 27707

Mark Draelos
Duke University
Durham, NC 27707

Jie Chen
University of Oulu, Finland

Alex Bronstein
Tel Aviv University
Ramat Aviv 69978, Israel

Guillermo Sapiro
Duke University
Durham, NC 27707

ABSTRACT

Intel's newly-announced low-cost and high precision RealSense 3D (RGBD) camera is becoming ubiquitous in laptops and mobile devices starting this year, opening the door for new applications in the mobile health arena. In this paper, we demonstrate how the Intel RealSense 3D camera can be used for low-cost gaze tracking and passive pulse rate estimation. We develop a novel 3D gaze and fixation tracker based on the eye surface geometry as well as an illumination invariant pulse rate estimation method using near-infrared images captured with RealSense. We achieve a mean error of 1 cm at $20 - 30$ cm for the gaze tracker and 2.26 bpm (beats per minute) for pulse estimation, which is adequate in many medical applications, demonstrating the great potential of novel consumer-grade RGBD technology in mobile health.

Keywords

Gaze tracker, pulse rate, depth camera, infrared, mobile health, mental health, human computer interaction.

1. INTRODUCTION

Gaze provides a wealth of information for human-computer interaction, particularly as an indicator of attention in medical diagnosis (e.g., autism screening [1, 2] and child emotion studies [3]). In such applications, detecting gaze to a computer monitor region (e.g., a quadrant of it), on-screen window, or body part is all that is needed. Those applications are often available only in lab settings using expensive equipments like Tobii eye trackers [4].

Heart rate (pulse) is a critical vital sign to assess the physiological state of a subject. In many applications, it is preferred or even required to measure the heart rate, e.g., of a patient, in a passive and remote manner. Recent studies validated the concept of detecting pulse passively from face color variation in a video [5, 6]. The cyclical movement of blood from the heart to the head via the abdominal aorta and the carotid arteries causes the head to move or face color to vary in a periodic way. However, those methods are known extremely sensitive to illumination variations. After all, given the numerous factors causing skin color fluctuation, changes from the cardiac pulse is subtle.

In this paper, we focus on low-cost, scalable and real-time analysis of human gaze and pulse for unique medical diagnosis applications, using Intel's newly-announced low-cost RealSense 3D (RGBD) camera. We thereby demonstrate that such devices provide an integrated (RGB, depth, and infrared) very low-cost sensor with critical health applications.[1] We first present a real-time gaze tracker from point clouds of the eye's scleral and iris surfaces acquired with RealSense. Unlike related techniques that employ 3D data in preprocessing steps [7, 8], require multiple camera units [7], or infer eye geometry from 2D images [9], we propose to exploit direct measurement of eye surface geometry captured by the RGBD camera. RealSense cameras infer depth information from a latent near-infrared (NIR) channel. We then experimentally demonstrate that pulse can be reliably estimated from faces in these near-infrared images. Such observation enables illumination-invariant passive heart rate estimation, and significantly extends its usage to low-light applications. The proposed framework for the measuring of critical health and medical signs is advantageous due to infrared structured illumination cameras' affordability, increasing availability, small form factor, low power consumption, and performance under low-light conditions.

2. METHODS

2.1 Gaze Tracking

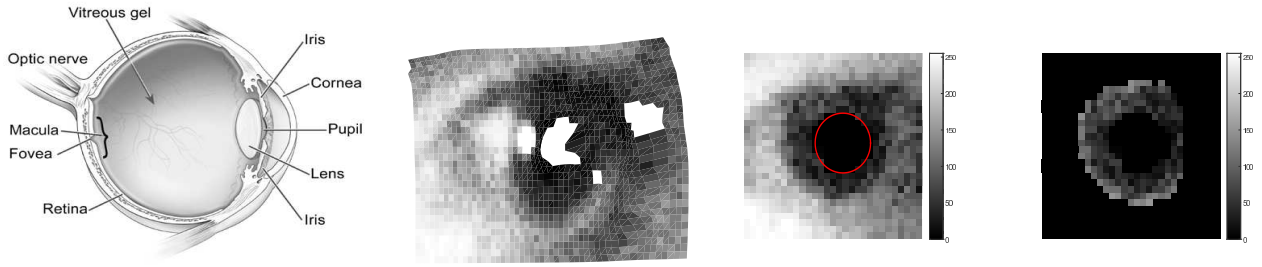
Eye surface geometry is defined by the approximately spherical scleral and corneal surfaces, as illustrated in Figure 1a. When imaging the human eye, however, infrared structured illumination in depth cameras produce sclera and iris point clouds only because the cornea itself is transparent (Figure 1b). The iris point cloud is of interest due to its geometrical relationship with the eye's optical axis or gaze, which is the line passing through the fovea center and the pupil's center. Notably, the iris is a shallow cone that surrounds the pupil and is oriented perpendicularly to the optical axis [11]. Thus, 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

*Work partially supported by NSF and DoD.

[1] While the exact production cost of the sensor is not public, is known to be in the order of low tens of dollars ($10-20), and known to have added very little cost, if at all, to current devices.

MOBIHEALTH 2015, October 14-16, London, Great Britain

DOI 10.4108/eai.14-10-2015.2261657



(a) Eye cross-sectional anatomy [10]. (b) Eye surface point cloud. (c) Pupil detection. (d) Iris segmentation.

Figure 1: Eye surface imaged by the RealSense camera.

in 3D space provide sufficient information to estimate gaze. We then develop and experiment a 3D fixation tracker that estimates the binocular fixation point on a known target surface, e.g., a computer screen or a car window, using simultaneous intersecting gaze estimates from corresponding left and right eyes.

Our gaze tracker consists of infrared segmentation and point cloud fitting. Pupil detection exploits poor reflection of (near) infrared illumination (860 nm as used in RealSense) back through the pupil, which produces a dark pupil in the infrared images. Given an infrared eye image, pupil detection identifies as the pupil the largest dark blob through thresholding and connected-components analysis (Figure 1c). Iris segmentation (Figure 1d) exploits the pupil-iris topographic relationship to expand the pupil region to include the iris using active contours [12]. 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 due to relatively 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 Figure 2a. Plane-based pupil mapping, Figure 2b, rejects the influence of distorted points, and constructs a non-ellipsoidal 3D eye surface model.

Binocular fixation (of both eyes) provides a convenient internal constraint during gaze estimation to further 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, the fixation target estimation, Figure 2c, 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. This nonlimiting constraint further improves gaze estimation. Finally, a bias field correction (computed only once) offsets the raw fixation estimates to compensate for target surface and depth camera coordinate system variations and calibration inaccuracies; Figure 3. This illustrates that the proposed approach can address the type of inaccuracies expected in commercial products and consumer scenarios that are not necessarily considered in professional laboratory devices such as Tobii high-end eye-trackers.

2.2 Pulse Rate Estimation

From the infrared and depth images, face detection and landmark tracking is performed to segment the cheek area of the face after which the average intensity is computed. A global self-similarity (GSS) filter and several temporal filters [5] are applied before computing the pulse frequency.

The Viola-Jones face detector [13] combined with a supervised decent method [14] is used to track facial landmarks in the infrared image (Fig. 4a). Next the connected components of the depth image are computed (Fig. 4b). The landmarks are then mapped from the infrared image to the depth image and the connected component, excluding background, containing the majority of facial landmarks is determined to be the face (Fig. 4c). Morphological operations such as erosion and hole filling are then performed on this region before using it as a mask to segment the face region in the IR image (Fig. 4d). Using the landmarks, the cheek is then extracted be selecting the area beneath the eyes and above the mouth (Fig. 4e). The temporal data from this region, averaged, is used for further analysis. Note that the 3D information is beneficial in tracking and detecting the region of interest for the IR measurements.

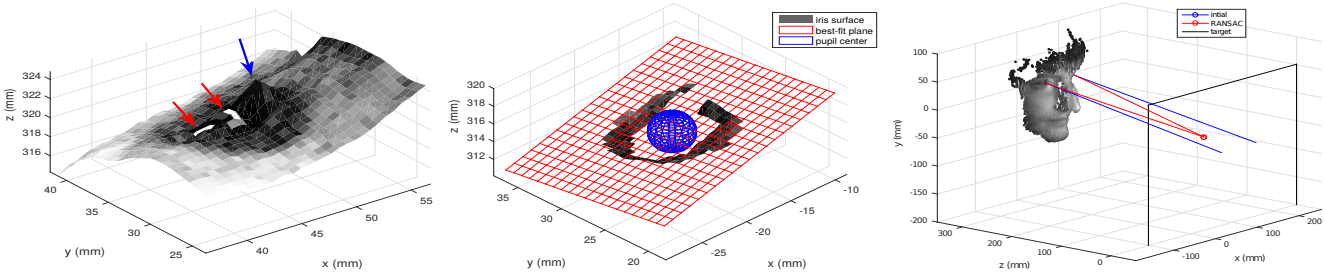
In the next step, Figure 5, a global self-similarity filter [15] is used to denoise the signal. Next, temporal filters [6] are applied. Firstly, a detrending filter [16] is used to reduce slow and non-stationary trends of the signal. A second moving-average filter is then applied to remove random noise. Lastly, a Hamming window based finite impulse response bandpass filter with cutoff frequency of [0.7, 4] Hz is used to exclude frequencies outside the [0.7, 4] Hz range, which corresponds to 42 to 240 bpm. After filtering, the pulse signal is converted to the frequency domain using FFT and its power spectral density distribution is estimated using Welch's method [17]. The frequency with the maximal power response is assumed to be the pulse frequency f_p, and the average pulse rate measured from the input video is computed as

$$P_{\text{video}} = 60 f_p.$$

3. EXPERIMENTAL RESULTS

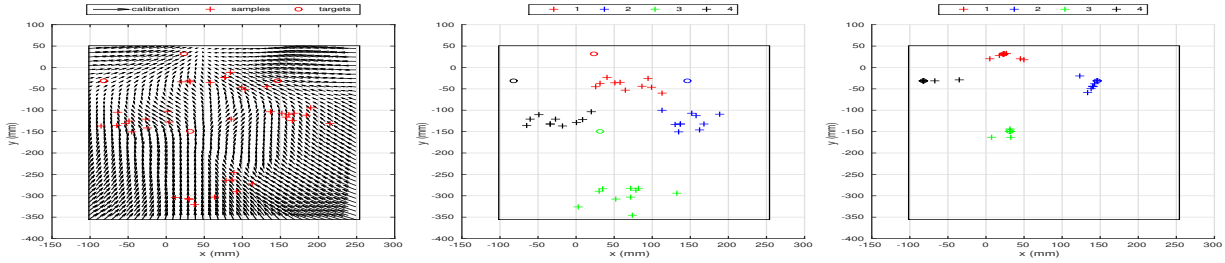
For testing the gaze tracking framework, corresponding infrared and depth images were taken at random fixation target positions in sets of ten images. In Fig. 3, half of the dataset at each target was used for bias field training, and the remaining half was used as the test dataset. Final fixation estimates are shown for the test dataset. For purposes of illustration, four target locations are depicted 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. The accuracy is more than sufficient for multiple tasks, e.g., child mental health tests (where pictures on the left or the right of the screen need to be selected, e.g., in autism or anxiety studies), and active-window selection in human computer interaction.

For testing the pulse framework, a dataset consisting of ten



(a) Pupil distortion in point clouds. (b) Plane-based pupil mapping. (c) Joint binocular fixation estimation.

Figure 2: Point cloud fitting in gaze estimation. (a) Eye point cloud surface colored with infrared image showing distorted position near the pupil (blue arrow) and gaps within the pupil (red arrows). (b) Best-fit plane (red) and pupil mapping (blue) for iris region point cloud surface. (c) Example point cloud depicting initial gaze estimates (blue) based on iris plane normals without the fixation constraint and the fixation estimate (red) on the target surface (black).



(a) Bias field. (b) Raw fixation estimates (c) Final fixation estimates

Figure 3: Fixation estimates at four example positions on the computer screen. (a) Bias field correction (computed only once) offsets raw fixation estimates compensating system inaccuracies. This provides an example of the importance of adjusting for challenges common in consumer scenarios. (b) Raw fixation estimates (+) and targets (o) colored by target. While precise estimations for important task is already obtained, this can be further improved. (c) Fixation estimates corrected with bias field; showing the high accuracy obtained with the RealSense low-cost RGBD camera. Sub-cm accuracy is obtained.

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. 3.

Method	$Me(SDe)$ (bpm)	RMSE (bpm)	Me_{Rate} (%)
Poh	-8.40 (27.98)	15.04	22.59
Balakrishnan	-5.91 (17.95)	10.33	15.07
Li	-1.45 (7.99)	4.56	6.32
Ours	2.26 (6.54)	3.66	5.34

Table 2: Mean and standard deviation of average pulse rate error for various methods.

subjects (1 female and 9 males) aged from 20 to 50 was collected using the RealSense camera. Each subject was recorded for about 90 seconds under varying illumination with slight head movements. Ground truth pulse rate was collected using a finger-oximeter. We compare our results with three previous methods: two color (RGB) based methods (Poh2011 [6], Li 2014 [5]) and one motion-based method (Balakrishnan2013 [18]). The mean error, standard deviation, root mean square error, and mean error rates are shown in Table 2. Our proposed framework not only achieves state-of-the-art performance but is the only one that worked in low-light and with varying illumination.

4. CONCLUSIONS

We have demonstrated the RealSense 3D camera is capable of gaze tracking and pulse estimation with good accuracy, all integrated in a single low-cost device. On average, we achieve sub-centimeter mean fixation estimate error for gaze and ±3 bpm for pulse, with added robustness to lighting conditions. Although our algorithms will benefit from more sophisticated processing techniques, the reported results have demonstrated the sensor and technique's feasibility and provide already sufficient accuracy for multiple important mobile health applications.

5. ACKNOWLEDGMENTS

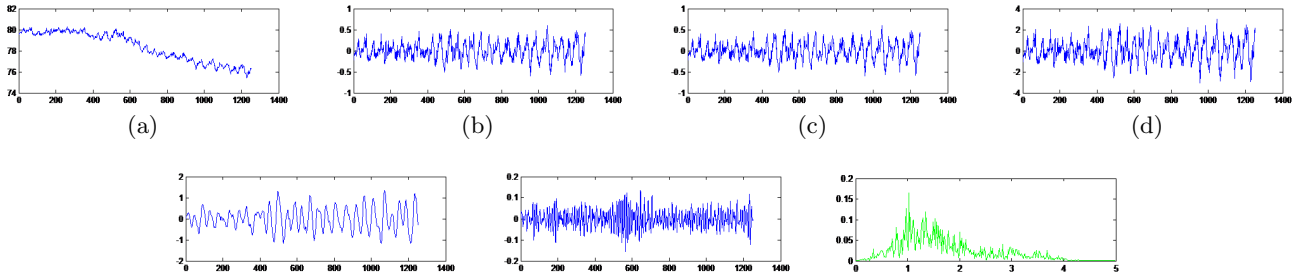
Work partially supported by NSF and DoD. Jie Chen was supported by University of Oulu, and the work was done while this author was visiting Duke University.

6. REFERENCES

[1] 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*, pages 1–7, Nov 2012.

(a) (b) (c) (d) (e)

Figure 4: Cheek region segmentation using infrared and depth image. (a) Facial landmarks tracked on infrared image. (b) Connected components in depth image. (c) The connected component containing the most landmarks is selected as the face. (d) Face region in infrared image. (e) Cheek area is selected as the region between the eyes and mouth landmarks.



(a) (b) (c) (d) (e) (f) (g)

Figure 5: Pulse rate measurements for cheek region in infrared image. (a) Infrared intensity. (b) GSS filtering. (c) Detrending. (d) Normalization. (e) Moving average filtering. (f) Band pass filtering. (g) FFT.

[2] W. Jones and A. Klin. Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431, Dec 2013.

[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*, 20(3):145–156, Sep 2011.

[4] Tobii. Eye tracking technology. *http://www.tobii.com/*.

[5] Xiaobai Li, Jie Chen, Guoying Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4264–4271, June 2014.

[6] Ming-Zher Poh, D.J. McDuff, and R.W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *Biomedical Engineering, IEEE Transactions on*, 58(1):7–11, Jan 2011.

[7] 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*, pages 1821–1828, June 2014.

[8] 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*, pages 25–30, June 2012.

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

[10] National Institues of Health, National Eye Institute. Diagram of eye, January 2015.

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

[12] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[13] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[14] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[15] Thomas Deselaers and Vittorio Ferrari. Global and efficient self-similarity for object classification and detection. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 1633–1640, 2010.

[16] M.P. Tarvainen, P.O. Ranta-aho, and P.A. Karjalainen. An advanced detrending method with application to hrv analysis. *Biomedical Engineering, IEEE Transactions on*, 49(2):172–175, Feb 2002.

[17] Peter D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *Audio and Electroacoustics, IEEE Transactions on*, 15(2):70–73, Jun 1967.

[18] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3430–3437, June 2013.