Accurate and Robust Eye Tracking with Ultrasound: A Computational Study

Ning Lu
Stanford University
Palo Alto, CA, USA
ningluu@stanford.edu

Francesco LaRocca
Meta, Reality Labs Research,
Redmond, WA, USA
flarocca@meta.com

Sachin Talathi
Meta, Reality Labs Research,
Redmond, WA, USA
talathi@gmail.com

2023 IEEE International Ultrasonics Symposium (IUS) | 979-8-3503-4645-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IUS51837.2023.10306398

***Abstract*—This paper presents an ultrasound simulation platform to synthesize realistic ultrasound eye tracking data as a function of transducer/ system design, sensor noise, eye/ face occlusion, and headset slippage. Simulation data were synthesized using a single face with adjustable gaze and eyelid opening in the presence of headset slippage. The data generated using this face/eye model was input into a machine learning algorithm to jointly estimate gaze and headset slippage. We achieved gaze root-mean-square-error (RMSE) of $0.085°$ and $0.756°$ without and with headset slippage, respectively. We anticipate that the proposed end-to-end simulation pipeline will enable tractable design optimization of wearable ultrasound devices and facilitate further investigation of ultrasound sensing solutions as a complementary technology to camera-based eye-tracking for AR/VR applications.**

I. Introduction

Continuous tracking of eye movement is important for detecting psychological states and understanding their relevance to cognitive processes in both clinical and non-clinical contexts [1], [2]. More recently, eye tracking has also received increased attention for applications in augmented reality (AR) and virtual reality (VR) [3], [4]. Real-time eye tracking provides users' gaze and eye positions, thus enabling display enhancements and hands-free interactions in AR/VR.

Most current eye tracking systems rely on cameras to capture the position of the iris and/or light sources reflected from the cornea (glints) [5]. These camera-based methods are naturally sensitive to ambient light, high in power consumption, limited in frame rate, and computationally intensive due to the use of computer vision and machine learning algorithms for video processing [6], [7]. Ultrasound sensing for eye tracking has recently been proposed to inherently mitigate the limitations of camera-based sensors such as sensitivity to ambient light [8], [9]. One of the significant advantages of ultrasound sensing is that ultrasound's reflectivity at all eye surfaces is more than 99.9% due to the high acoustic impedance mismatch at the air/eye interface [8]. This offers up to $50\times$ better reflectivity compared to optical-based techniques (2% reflection at the air/cornea interface [10]). Furthermore, recent breakthroughs in Micro-Electro-Mechanical-Systems (MEMS)-based ultrasound transducers enable ultrasound devices to be very small, low cost at scale, and low power [11], [12]. These advantages, along with the low latency offered by ultrasound's propagation speed, make ultrasound-based eye tracking very promising for head-mounted wearable devices.

Several studies have demonstrated the feasibility of ultrasound-based eye tracking in the past few years [8], [9], [13]. Kaputa and Enderle first explored the possibility of using non-contact ultrasound sensors to track fast eye movements using a finite element simulation model with four transducers positioned perpendicular to the cornea [13]. Golard and Talathi proposed a ring architecture, which used a constellation of discrete Capacitive Micromachined Ultrasonic Transducers (CMUTs) distributed on the inner ring of a glasses frame, and machine learning to estimate gaze, achieving a gaze root-mean-square-error (RMSE) of 0.97 ± 0.18 degrees [8]. Sun *et al.* integrated piezoelectric micromachined ultrasonic transducer (PMUT) arrays on the lenses of glasses that were lightweight (< 25 mg), compact (millimeters in size), high-speed (response time < 1 ms) and had low power consumption (71μW) with qualitatively robust eye tracking and blink monitoring on human volunteers [9].

Although these studies have shown the potential of ultrasound as a viable sensor for eye tracking, none of them have provided quantitative evidence that ultrasonic eye tracking can meet or exceed the precision and robustness of conventional camera-based eye trackers and certainly not in a commercially viable AR/VR form factor. We anticipate that the performance of an ultrasound-based eye tracking system is highly dependent on many variables and thus optimizing the system design experimentally would be very challenging, time-consuming, and expensive. Therefore, we frame the problem of ultrasound eye-tracking system design using computational methods in the simulation domain. In this work, we present the first ultrasound simulation platform for eye tracking based on acoustic full-wave propagation. This simulator takes into consideration various acoustic configurations, sensor noise, eyelid/face occlusion, and headset slippage-induced variability. Synthetic pulse-echo data generated by the simulator are fed into an end-to-end (E2E) machine-learning algorithm to estimate gaze and headset slippage. Results from the E2E algorithm are analyzed for various system configurations to understand how performance can be improved with component and system-level changes and to provide direction on optimal designs for ultrasound-based eye tracking.

The first author conducted this work as a Research Intern at Meta Reality Labs.

II. Methods

A. Acoustic simulation

We performed ultrasound simulations using k-Wave, an open-source acoustic toolbox for acoustic wave propagation that can account for medium heterogeneity, absorption, and nonlinearities [14]. Using a k-space pseudo-spectral method, k-Wave requires fewer spatial and temporal grid points to calculate fast Fourier transforms (FFTs) of wave fields and is more computationally efficient compared to conventional finite-difference time-domain (FDTD) models. In this study, the simulation spatial and temporal step sizes were set to 0.17 mm (equivalent to $\lambda_{min}/4$, where $\lambda_{min} = 0.686$ mm, which is the wavelength of acoustic waves in air at 500 kHz) and 21 ns, respectively.

To accelerate simulations, we optimized other simulation parameters (i.e. use of small prime factors for grid size to speed up FFTs, single datatype in data-casting) and implemented the C++ version of k-Wave with multi-threading on a distributed computing cluster. Each simulation was run in parallel over 96 compute cores (Xeon Platinum 8259CL 2.5GHz, Intel, California, USA) with 40 GB of collective memory. These optimizations provided a 10x speed improvement compared to the MATLAB-only version of k-Wave reducing the average simulation execution time to within 30 minutes per gaze.

B. Simulation setup

1) Acoustic eye/face model: For the eye model, we used a simple sphere-on-sphere eye with an eyeball radius of 12.2 mm, a cornea radius of 8.7 mm, and a 4.47 mm distance between the centers of the two spheres. Different mediums such as air, cornea, sclera, and skin (at 20°C) were assigned sound speeds of 343, 1553, 1583, and 1624 m/s and densities of 1.2754, 1024, 1048, and 1020 kg/m^3, respectively [15], [16]. Attenuation α in k-Wave was modeled as a power law.

We simulated 2-directional gazes [yaw, pitch] $\in \pm 30^\circ$ with a 5° step size (resulting in 169 gaze angles in total). In order to simulate occlusion during eye tracking, we used a set of anatomically-similar CAD models of the face and eyelids and centered the eyelid opening with the center of the eye along the optical axis as shown in **Fig.**1. The CAD models shared the same geometric profile but with different eyelid openings of 15, 30, or 45°. Eyelash occlusions were not included herein.

2) Ultrasound transducer: Previous studies have investigated ultrasound transducers with center frequencies (f_c) between 500 kHz to 1.7 MHz for eye tracking [8], [9]. Higher frequencies offer better temporal resolution but lower signal-to-noise-ratio (SNR) because ultrasound signals attenuate exponentially in air with increasing frequency. Therefore, in this study, we implemented all simulations at $f_c = 500$ kHz, which has been shown to be capable of providing an SNR ≥ 40 dB using MEMS-based ultrasound transducers [9]. Each ultrasound transducer was modeled as a 1.22×1.22 mm square array that consists of 4×4 elements with an element radius of 0.2 mm and a half-wavelength spacing (0.34 mm) between elements.

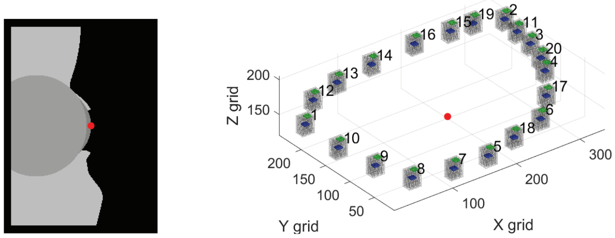

Fig. 1. K-Wave simulation setup. Left: Cross-section of the medium volume on sagittal plane. Right: The ring pattern for ultrasound transducers. Blue: Nominal transducer locations; Green: Effective transducer locations for one random slippage value; Black: point sensors where signals are being recorded. The red dot represents the apex of cornea surface at [yaw, pitch] = [0,0].

3) Ring configuration: Our previous study demonstrated the feasibility of using a circular ring of ultrasound transducers for eye tracking [8]. In this study, we investigated a non-circular ring pattern consisting of 20 transducers distributed on the inner ring of a realistic glasses frame with varying eye reliefs (**Fig.**1). Note that the large number of transducers in this ring pattern was not used because all transducers were necessary to achieve accurate/robust eye tracking, but because increasing the number of receivers did not increase simulation time and can be useful for down-selecting the best locations of a smaller number of transducers in post-processing. All transducers were effectively oriented towards the apex of the cornea when the eye was pointing straight ahead (gaze angle [yaw, pitch] = [0,0]) using electronic focal steering. This transducer orientation reduced the impact of occlusions, since most of the transmitted and received energy from the transducers were focused on the eye rather than the surrounding face and eyelids. For simplicity, only one transducer (index 1 in **Fig.**1) was used for transmission while all transducers were used for receiving.

4) Headset slippage: Head-mounted devices often move around or slip during normal use. In our study, we assumed that this slippage falls within the range of $\Delta x = \pm 1.0$ mm, $\Delta y = \pm 0.7$ mm, $\Delta z = \pm 2.0$ mm from the nominal location [17]. To investigate how headset slippage may impact eye tracking performance, we considered the following 2 cases:

1) **Worst-case slippage**: Reflection signals were received by the transducers after they slipped with the largest possible offsets within our slippage range.
2) **Random slippage**: Reflection signals were received by the transducers after applying offsets Δx, Δy, and Δz that were randomly chosen within our slippage range.

To include slippage without significantly increasing simulation runtime, we recorded from 20800 point sensors in a $2 \times 1.4 \times 4$ mm volume centered at the nominal location of each transducer (**Fig.**1). We then summed the waveforms received by a subset of sensors corresponding to a particular amount of slippage. Synthetic data with slippage was mixed together with synthetic data recorded at nominal transducer locations for both training and testing data for the machine learning model. It should be noted that all transducers were shifted uniformly with identical offsets for each slippage sample.

C. Pipeline for synthetic data generation and analysis

In k-Wave, a 3-cycle sinusoidal burst was applied to the ultrasound transducers with an anti-aliasing low-pass filter (cutoff at 2 MHz) and transmitted to the eye/face through the air. The reflection signals from the eye/face were received by point sensors and summed across elements on each transducer to produce the received signals on the transducer arrays. A 3rd-order Butterworth bandpass filter with a bandwidth equal to 20% of the transducer's resonant frequency was applied to these waveforms to compensate for the device's bandwidth. Bandlimited Gaussian noise (due to additive Gaussian noise from the pre-amplifier) was then added to the received signals to produce 50 synthetic recordings per gaze angle.

The simulator was then used to synthesize the dataset used for training/testing the ML model by concatenating data from 169 gaze angles $\times 4$ eyelid occlusions ($15°$, $30°$, $45°$, and no occlusion) $\times 50$ recordings per sample. Different values of system-level variables, including SNR, number of features, number of receivers, location/number of transmitters, headset slippage, etc., were tested in simulation to investigate how they impact eye tracking performance.

Peak amplitude and time-of-flight (ToF) were extracted as features from the synthetic dataset using the same methods described by Golard and Talathi [8]. We also investigated alternative approaches to analyze ToF such as the time-of-arrival for the front edge of the ultrasound signal (using the time at which the slope of the envelope is largest) and the difference of ToF across transducers. These features were then fed into 2 regression trees trained with gradient boosting methods [18] to independently estimate gaze along the horizontal and vertical axes with a train/test split ratio of 80/20 [8]. For datasets that involved slippage, 3 regression trees were trained to independently estimate slippage in 3 dimensions simultaneously. The following parameters were used for the ML model: max tree depth = 7, number of regression trees = 100, and the learning rate was fine-tuned in the range of 0.05 - 0.3 (default value = 0.1). The performance of eye tracking using synthetic datasets was quantified using the adjusted R^2 scores, root-mean-square-errors (RMSE), and the precision of gaze and slippage prediction.

III. Results

A. SNR

In Table I we summarize the performance for gaze estimation using the ultrasound synthetic data. With 60dB SNR (Sim 1.0.0), the ML algorithm was able to produce a gaze RMSE error of $0.085°$ with an adjusted R^2 score of 99.99%. However, when SNR decreased to 16 dB, the precision of gaze estimation degraded, resulting in a gaze RMSE error of $0.82°$ with an adjusted R^2 score of 99.81%. Though the adjusted R^2 scores were relatively high for both SNR values, the percentage of estimated samples that fell within $1°$ and $0.1°$ of the actual gaze was reduced significantly. We observed a larger reduction in gaze in the pitch direction (up-down) compared to the yaw direction (left-right). This follows the fact that eyelid occlusions occur mostly near the top and bottom of the eye, thus hindering gaze estimation more severely along the pitch direction.

TABLE I
Gaze estimation performance metrics

	1.0.0	1.0.1	1.1.1	1.1.2	1.2.0	1.2.2
SNR (dB)	60	16	60	60	60	60
Slippage	No	No	worst	random	No	random
Features	4	4	4	4	2	2
Adjusted R^2	0.999	0.998	0.999	0.998	0.999	0.993
RMSE	0.085	0.823	0.074	0.756	0.112	1.591
1° precision, yaw	100.0	97.9	100.0	98.9	99.9	94.7
1° precision, pitch	99.8	79.1	99.7	86.2	99.8	45.4

B. Headset slippage

Our results showed that the worst-case slippage (Sim 1.1.1) had a negligible impact on the performance for gaze estimation, compared to the baseline simulation with no headset slippage (Sim 1.0.0). However, major performance degradation was observed for random slippage when we added synthetic data using 8 different values of slippage that were randomly distributed within the previously specified range (Sim 1.1.2). For random slippage, the estimated gaze RMSE error increased from $0.0850°$ to $0.7562°$ and the precision within $0.1°$ decreased noticeably to 59.37% and 17.36% for yaw and pitch gazes, respectively. Such performance degradation was expected, since headset slippage changes the distance between the transducers and the eye, leading to changes in ToF that can be difficult to distinguish from ToF differences due to variations in gaze. When feeding the amplitude and time-of-flight features in 3 additional gradient-boosted trees, we were able to jointly estimate headset slippage in 3 dimensions together with gaze, achieving a total slippage RMSE error of 0.12 mm and a slippage estimation precision of 87.55% within 0.2 mm (Fig.2)

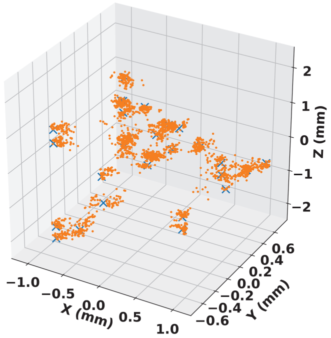

Fig. 2. Nominal (blue X) vs estimated (orange dot) headset slippage.

C. Number of features and receivers

Table.II summarized how the number of transducers and features impacted the performance of gaze estimation. With a fixed number of transducers, the ML model achieved better performance using 4 features than 2 features for both synthetic datasets with and without slippage (Sim 1.2.2 and 1.2.0, respectively), particularly for pitch gaze estimation. This suggested that more features (i.e., ToF of the envelope peak, ToF of the front-edge, and difference in ToF across transducers) can be helpful even though they are different methods to describe the same type of feature (i.e. ToF) in the synthetic data.

To investigate the robustness of gaze estimation with fewer transducers, we down-selected from all 20 transducers in the ring pattern to the 4 following subsets: (1) 15: Transducers 1-15; (2) 9: Transducers 1-6, 8, 10, and 14; (3) 6a: Transducers 1, 8, 9, 10, 12, and 14, which were selected as the "most important receivers" based on the importance map given by the gradient boosted trees; (4) 6b: Transducers 1-6. Our results demonstrated that, the first (15) and third (6a) subsets achieved comparable precision values to the full ring pattern for gaze estimation, while the second (9) and the fourth (6b) subsets resulted in noticeable performance degradations (Table.II). This suggested that a sparse set of carefully placed transducers provides significantly better eye tracking performance than a dense set of sub-optimally placed ones.

TABLE II
THE IMPACT OF NUMBER OF TRANSDUCERS AND FEATURES ON GAZE ESTIMATION.

Transducer subset		20	15	9	6a	6b
yaw	2 features	94.7	94.6	92.4	93.4	69.4
	4 features	98.9	99.1	98.8	99.0	84.6
pitch	2 features	45.4	51.0	43.7	50.3	31.0
	4 features	86.2	88.2	85.8	88.0	54.3
overall	2 features	41.8	46.9	39.4	46.4	20.9
	4 features	84.8	87.0	84.6	86.9	45.4

IV. Discussion

In this work, we presented the first ultrasound simulation platform to synthesize ultrasound eye tracking data as a function of transducer/system design, sensor noise, eye/face occlusion, and headset slippage. Compared to the ray-tracing model that was used in our previous study [8], this new platform implemented full-wave simulation to take into consideration reflections and interference effects at eye/face surfaces, allowing for the synthesis of more realistic ultrasound signals. Though this simulator was only demonstrated for ML-based E2E eye tracking herein, it can also be generalized to explore other non-ML models for ultrasonic eye tracking, such as the traditional beam-forming approaches using phased arrays.

The results from our E2E algorithm showed that gaze can be estimated with high precision (RMSE of $0.085°$) using ultrasound when headset slippage is not present and that adding random slippage degrades gaze estimation significantly (RMSE of $0.7562°$) but still within $1°$, which is comparable to camera-based eye tracking. We also demonstrated that, in addition to gaze, headset slippage can be estimated by the ML model as well (RMSE of 0.12 mm). Finally, we showed that this simulation framework was helpful in determining useful features within the ultrasound signal from the eye/face and in reducing the number of transducers with minimal impact on eye tracking performance.

Ideally, an apple-to-apple comparison between the ultrasound and optical-based eye-tracking platforms would be beneficial to fully demonstrate the advantages of using ultrasound approaches for eye tracking. However, to date, most optical-based eye trackers were evaluated with different methodologies, so the performance numbers can only be used to give us a ballpark understanding of how our platform compares to theirs. A commercially available VR headset was evaluated in the visual perimetry at 25 target positions spanning a range of $\pm 26.6°$ in a head-fixed condition, resulting in an average accuracy of $4.16°$ [19]. Another video-based eye tracker was shown to have a pupil-tracking accuracy of $0.5°$ [20]. Future investigations will be necessary to evaluate the performance of different eye tracking platforms using standardized metrics.

V. Conclusion

We presented an end-to-end simulation pipeline that enables tractable design optimization of wearable ultrasound devices and shows the feasibility of accurate, robust eye tracking using ultrasound for a variety of applications including AR/VR. Future studies will be required to investigate the impact of other variabilities on eye tracking performance, such as different Tx waveforms, improved features, eyelash occlusions, etc.

References

[1] M. Macaskill, C. Graham, T. Pitcher, D. Myall, and et al., "The influence of motor and cognitive impairment upon visually-guided saccades in Parkinson's disease," *Neuropsychologia*, vol. 50, no. 14, p. 3338, 2012.

[2] J. Reilly, R. Lencer, J. Bishop, S. Keedy, and J. Sweeney, "Pharmacological treatment effects on eye movement control," *Brain and Cognition*, vol. 68, no. 3, pp. 415–435, 2008.

[3] S. Ahn and G. Lee, "Gaze-Assisted Typing for Smart Glasses," *UIST19*.

[4] V. Clay, P. König, and S. König, "Eye Tracking in Virtual Reality," *Journal of Eye Movement Research*, vol. 12, 2019.

[5] A. Kar and P. Corcoran, "A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[6] Y.-m. Cheung and Q. Peng, "Eye Gaze Tracking With a Web Camera in a Desktop Environment," *IEEE Trans. Hum.-Mach. Syst.*, vol. 45, no. 4, pp. 419–430, 2015.

[7] N. Panigrahi, K. Lavu, S. Gorijala, P. Corcoran, and S. Mohanty, "A Method for Localizing the Eye Pupil for Point-of-Gaze Estimation," *IEEE Potentials*, vol. 38, no. 1, pp. 37–42, 2019.

[8] A. Golard and S. Talathi, "Ultrasound for Gaze Estimation-A Modeling and Empirical Study," *Sensors*, vol. 21, no. 13, p. 4502, 2021.

[9] S. Sun, J. Wang, M. Zhang, Y. Yuan, and et al., "Eye-Tracking Monitoring Based on PMUT Arrays," *J Microelectromech Syst*, vol. 31, no. 1, pp. 45–53, 2022.

[10] S. Hayes, P. Lewis, M. Islam, J. Doutch, T. Sorensen, and et al., "The structural and optical properties of type iii human collagen biosynthetic corneal substitutes," *Acta Biomaterialia*, vol. 25, pp. 121–130, 2015.

[11] J. Jung, W. Lee, W. Kang, E. Shin, J. Ryu, and H. Choi, "Review of piezoelectric micromachined ultrasonic transducers and their applications," *J. Micromech. Microeng.*, vol. 27, no. 11, p. 113001, 2017.

[12] J. Joseph, B. Ma, and B. Khuri-Yakub, "Applications of Capacitive Micromachined Ultrasonic Transducers," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 2, pp. 456–467, 2022.

[13] D. Kaputa and J. Enderle, "An ultrasound based eye tracking system," *J. Biomed. Eng. Med. Dev*, vol. 1, pp. 1–4, 2016.

[14] B. Treeby and B. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *J. Biomed. Opt.*, vol. 15, no. 2, p. 021314, 2010.

[15] "Speed of Sound IT'IS Foundation."

[16] K. Takemura and K. Yamagishi, "A hybrid eye-tracking method using a multispectral camera," in *2017 IEEE SMC*, pp. 1529–1534.

[17] S. Stojanov, S. Talathi, and A. Sharma, "The benefits of depth information for head-mounted gaze estimation," in *ETRA '22*.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD '16*, p. 785–794.

[19] I. Schuetz and K. Fiehler, "Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability," *J. eye mov. res*, vol. 15, no. 3, 2022.

[20] C. Sheehy, Q. Yang, D. Arathorn, P. Tiruveedhula, and et al, "High-speed, image-based eye tracking with a scanning laser ophthalmoscope," *Biomed. Opt. Express*, vol. 3, no. 10, pp. 2611–2622, 2012.