EYESIGHT: Eye Examination System with Intelligent Guidance and Human Tracking

Alexander D. Smith[1,2*], *IEEE Student Member*, Wenzhou Ding[3*],
Yuanyi Feng[1], and Kris Hauser[1], *IEEE Senior Member*

***Abstract*—Automated retinal inspection systems may enable broader access to consistent diagnostic eye examinations. In this work, we present an automated, contactless posterior eye examination system that can perform examinations at safe distances on freestanding patients. The system uses seven cameras, with two tracking the head, four tracking the pupil being examined, and one collecting diagnostic images through a wide-angle indirect ophthalmoscope lens to capture a 56° retinal field of view. The sensor system is mounted on a robot arm, which tracks the center to the patient's head motion, locking onto the pupil, and capturing images of the retina. Feasibility studies on a moving phantom and members of the research team indicate that the system can generate high quality images suitable for diagnostic purposes.**

***Index Terms*—Automation in Health Care; Motion Planning and Control; Big Data and Deep Learning**

I. Introduction

Clinicians must examine the posterior segment of the eye to observe the health of the retina. The retina serves as an important screening tool in medicine to identify and monitor several diseases including hypertension, diabetes, neurological disorders, and ocular disorders [1]. Examination techniques are limited in efficacy, particularly in routine primary care medical examinations by non-ocular specialists [2]. As such, many retinal conditions are missed due to inconsistent techniques or lack of posterior segment examination. Several retinal imaging systems have been developed to perform more comprehensive retinal screening, such as scanning laser ophthalmoscopes [3] capable of generating 200° retinal images. In setups like these, patients place their head on a chinrest and keep their eyes still for 0.5 s to ensure that images are captured with minimal artifact. As such, these tools are limited to eye clinics, involve physical contact with the patient for stabilization, and must be operated by skilled medical staff.

Recent work has demonstrated the feasibility of robot-stabilized optical coherence tomography (OCT) in freestanding patients [4, 5] with recent versions attaining with up to a 32° field of view at a 86 mm working distance [6]. However, no robot-stabilized system has been developed to perform posterior segment wide-angle imaging under white light. The wide-angle lensing of white light retinal imaging poses two technical challenges: 1) closer working distance (6 cm in our setup) and 2) interference with in-line pupil tracking cameras used for alignment in prior work. In prior work, environment-fixed cameras for head tracking cause a short loss of active tracking during the approach phase, when the sensor head occludes the patient's face [4, 5]. This leads to a risk of procedure failure and increased patient risk, which is exacerbated at closer working distances. Moreover, when pupil and head tracking cameras are mounted off-axis, the performance of tracking methods tends to degrade.

We propose EYESIGHT, a robot-mounted modular sensor system to perform fully automated mydriatic retinal inspection on freestanding patients (Fig. 1). Our system uses a single robot-mounted sensor package capable of head and pupil tracking and sub-two millimeter visual servoing [7] accuracy for contactless and safe control within 6 cm of the patient. To address the aforementioned weakness of prior systems, EYESIGHT leverages recent advances in tracking algorithms that allow cameras to track objects from closer and more extreme angles [8]. It performs sensor fusion over six cameras to track the head and eyes across a wider workspace than prior work [4, 5, 6]. Once pupil lock is obtained, an imaging sequence from the clinical sensor is then automatically processed using a deep learning model to identify retinal images that are clear, free from motion artifacts, and free from iris obstruction. We present experiments evaluating our system's accuracy on both a phantom and members of the research team, demonstrating adequate tracking performance. We also conduct a pilot evaluation of the full EYESIGHT system on a hand-moved phantom, demonstrating that it produces a series of clear images of the retina with 56° field of view even under constant target movement.

Our system is demonstrated in the supplemental video that accompanies this paper.

II. Related Work

A. Robotic Eye Examination Systems

Robotic ophthalmological procedures date back to as early as 2005, focusing on reducing tremor in delicate eye surgical procedures [9, 10]. Robotic ophthalmology recently expanded to eye examinations, including teleoperated robots

This research is supported by NIH Grant #R01 EY035106 A.

*Authors contributed equally to this work

[1]Department of Computer Science at the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ads10@illinois.edu, kkhauser@illinois.edu

[2]Carle Illinois College of Medicine at the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[3]Department of Mathematics at University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

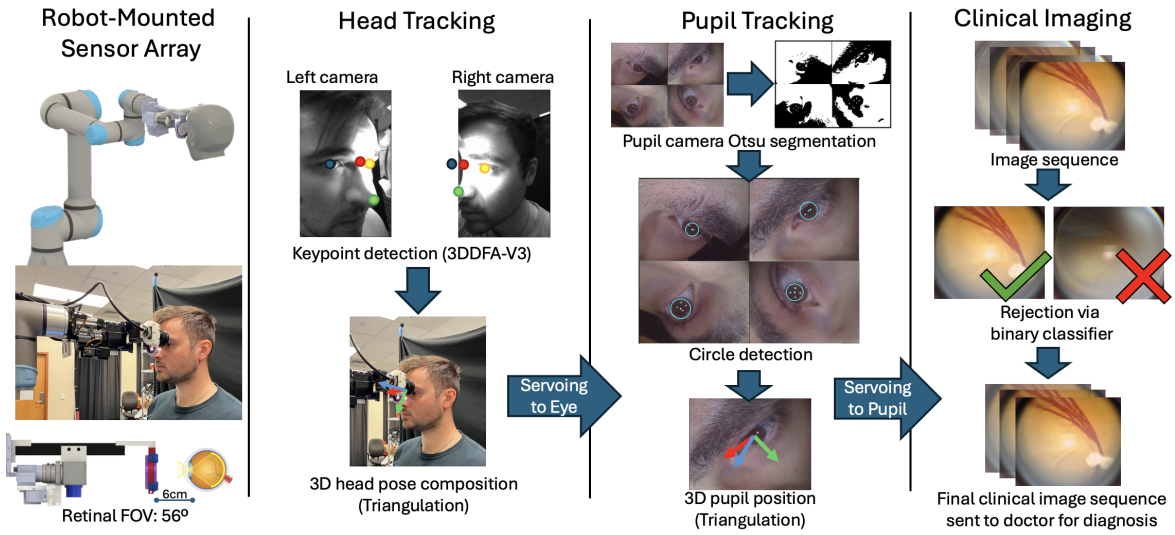

Fig. 1. Proposed system, starting with head PBVS, followed by pupil PBVS, then clinical imaging of the retina. Head tracking uses 3DDFA-V3 for keypoint detection and triangulation to resolve the 3D head keypoint positions. Pupil tracking uses Otsu segmentation and a Hough circle transform for pupil detection and a triangulation to resolve 3D pupil position. We use binary classification to reject low quality images from the clinical image sequence, generating a high-quality retinal image sequence that can be visualized for diagnosis by an ophthalmologist.

[11] in which a remote eye doctor provides remote commands to a system in the patient room with the assistance of a technician, which makes eye examinations available to more patients [12]. Recent autonomous eye examination systems aim to conduct repeatable, contactless eye exams without patient fixation. The first such system uses optical coherence tomography (OCT), which is used to capture 3D imaging of the eye to monitor a range of ocular disorders [4].

For head tracking, prior work used environment-mounted cameras [5, 13]. This causes a limitation where the patient's face is occluded by the sensor head as it approaches, leading to unreliable tracking [5]. In the approach phase, an open-loop motion is required to bridge the head tracking and pupil tracking phase. If the patient moves during this movement, pupil tracking can fail or the robot can collide with the patient. Our robot-mounted head tracking system allows for continuous closed-loop control during the approach phase.

For pupil tracking, prior work used an inline beamsplitter to combine the pupil camera and OCT optics [5]. In indirect ophthalmoscopy, inline pupil imaging is not possible at the same time as retinal imaging due to the intervening fundoscope lens. Instead, pupil tracking cameras must be mounted at an oblique angle, which makes tracking more challenging due to interference with the patient's nose, eyelash, and cheek.

B. Head Tracking

Head tracking has progressively advanced with new machine learning approaches. Prior robotic eye exam systems used the OpenFace2.0 algorithm, which becomes unreliable during occlusion and at oblique viewing angles [13]. Newer face tracking methods have the potential to work more reliably in close-range head tracking, such as 3DDFA-V2, 3DDFA-V3, and Google MediaPipe Face Mesh [8, 14, 15]. These deep learning approaches can be used for 3D head tracking with well-calibrated stereo cameras and state estimation. Recent work has explored 3D head tracking with a simultaneous localization and mapping methodology, though the approach is limited to flat-plane backgrounds due to use of a SIFT keypoint detector [16]. Face keypoint detectors appear promising as an alternative for head tracking. Triangulation is a naïve approach to 3D head tracking from 2D keypoints, and one recent approach includes the use of an unscented Kalman filter to estimate T_h from tracked keypoints [17]. In our work, we analyze several head tracking models on our stereo head tracker and perform 3D head pose estimation using triangulation.

C. Pupil Tracking

First and fourth Purkinje reflections can be used to determine eye gaze vector [18], and a recent work has digitized this approach for eye gaze vector tracking [19]. Deep learning approaches have become commonplace in pupil tracking and gaze vector estimation [20], based on a variety of open-access datasets. Existing eye gaze datasets are either too specific to particular hardware, or too general to transfer well to specific camera orientations and hardware. Simulated eye gaze datasets provide excellent ground truth, but have camera pose configuration limitations and simulation-to-real gaps, particularly for blink, saccade, and occlusion simulations [21, 22]. Real eye gaze datasets are often limited in range of eye orientations, expensive to generate, and ground truth is error-prone [23]. Our testing indicates that convolutional neural networks trained on such data tend to degrade in performance when out of distribution, in particular, in the oblique and out-of-center images frequently seen in our system.

A fast approach for pupil detection uses the Hough circle transform [24], which can serve as an accurate pupil center detector. However, Hough circle transforms require hardware-specific tuning, and can be sensitive to patient appearance variation. Pupil tracking in previous robotic OCT work relies on inline and oblique cameras [4, 5, 6]. Our work uses oblique cameras only, which are positioned at the outer edge of the clinical objective lens. Our method uses

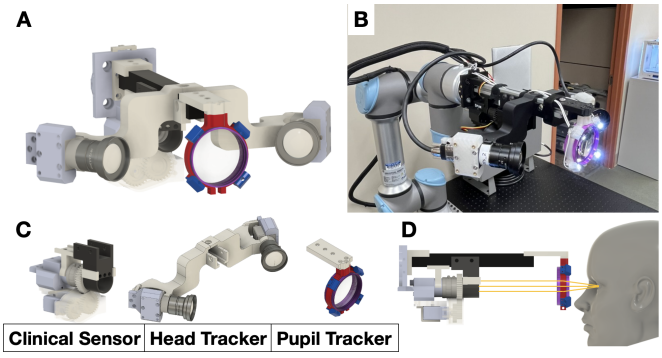

Fig. 2. End effector sensor array design. A) End effector sensor array CAD image. B) Assembled sensor array mounted on UR5e at the home position. C) Modular components for clinical imaging, head tracking, and pupil tracking. D) Retinal illumination design (head tracker removed for clarity), with illumination element paraxial to the clinical sensor.

Otsu segmentation [25], a Hough circle transform, and inter-camera consistency checking to achieve robust 3D tracking.

III. System Design

Our system consists of a robotic arm affixed to an optical table, and uses a central computer for tracking the patient and commanding the robot. Our sensor design differs from prior works [4, 5, 6] in that it contains all tracking components on the robot's end-effector, allowing occlusion-free motion relative to the patient and closer working distances. The assembly consists of a pupil tracker, head tracker, and clinical sensor (Fig. 2C). Our pupil tracker module contains four small cameras with respective camera centers aligned toward the focal point of an indirect ophthalmoscope lens. Our head tracker module contains two robot-mounted head tracking cameras, allowing a point-based visual servoing (PBVS) target to be seamlessly transferred from head tracking to the four robot-mounted pupil tracking cameras. A seamless fallback to head tracking occurs if pupil tracking fails, and if both fail, the robot retracts the sensor to a safe home position. Our clinical imaging camera, with a light source near the axis of the camera, is designed to view through the lens, capturing diagnostic retinal images.

Because indirect ophthalmoscopy requires a lens placed at a close working distance to the eye for retinal examination, an inline pupil camera is infeasible for PBVS and simultaneous clinical imaging. In our work, due to our unique camera arrangement, we use a classical approach to pupil detection via Otsu segmentation and a Hough circle transform [24, 25], which is both fast and provides robust detection on the tested subjects.

Our clinical sensor is designed to capture images through an indirect ophthalmoscope, and is based on the existing binocular indirect ophthalmoscope headset [26]. We use a single camera with a telephoto lens to magnify the view and use a paraxial ophthalmoscope light for illumination.

A. Design Requirements

Our system must perform PBVS relative to landmarks on the head with high enough accuracy to visualize the retina of a free-standing patient with a wide-angle indirect ophthalmoscope lens. To assess these features of our system, we propose the following constraints:

1) **Lighting Safety:** must be safe for >10 second exposures at the corneal surface as defined by ANSI 2000 [27]
2) **Safety:** head tracking should be sufficiently accurate during normal physiologic head motion, and robot motion response latency should be below 200 ms to avoid collisions at far distances (>10 cm), and below 50 ms during close distances (<10 cm).
3) **Servoing Accuracy:** target should remain within 1 cm for head tracking, and 4 mm for pupil tracking at a 6 cm working distance to view through an 8 mm pupil.
4) **Diagnostic Image Quality:** diagnostic images should clearly show the retina in >75% of reported images.

B. System Architecture

Our system hardware consists of a computer with 16 GB RAM and an NVIDIA RTX 4060 Ti, a UR5e robot, a Volk Pan Retinal 2.2 BIO indirect ophthalmoscope lens, a lighting array to illuminate the face and eye, seven cameras attached to the robot's end effector, and 3D printed mounting materials. The software consists of Ubuntu 22.04 LTS, Python 3.10.12, a Redis database to handle multi-process communications between tracking and control scripts, a behavior tree developed with PyTrees, and the Klamp't robot interface layer to handle robotic control and simulation [28].

Our sensor array contains the tracking cameras and lights necessary to perform retinal viewing (Fig. 2) in a payload that weighs 1.150 kg. Our pupil tracker consists of four Dothecamera 3.9 mm endoscope camera modules mounted on a ring that holds the lens. The endoscopes have built-in LEDs used for face illumination. Our head tracker consists of two FLIR Blackfly S BFS-U3-23S3M cameras with Kowa LM3NC1M lenses mounted at 30° off-axis of the effector (Fig. 2C). Our clinical sensor consists of one FLIR Blackfly S BFS-U3-51S5C camera with an Edmunds 25mm fixed focal length lens mounted further back on the effector to capture a complete view of the retina through the lens. The illumination element contains a Welch Allyn 04900-U LED bulb paraxial to the clinical sensor.

C. Keypoint Tracking

We use two keypoint trackers: a head tracker and a pupil tracker. The base of the robot is defined as the world origin, and we transform keypoints into world coordinates for use by our control system.

1) Head Tracking: Head tracking involves 2D keypoint detection and 3D keypoint triangulation. We define a head coordinate system T_h with respect to world coordinates. The origin of T_h is the anatomical nasion, the x'-axis is defined parallel to the line intersecting the two eye centers from left to right and originating from the nasion, the y-axis is defined along the line between the nasion and the nose tip, and the z-axis is the cross product of the x'- and y-axes. The x-axis is then defined as the cross product of the y and z axes to

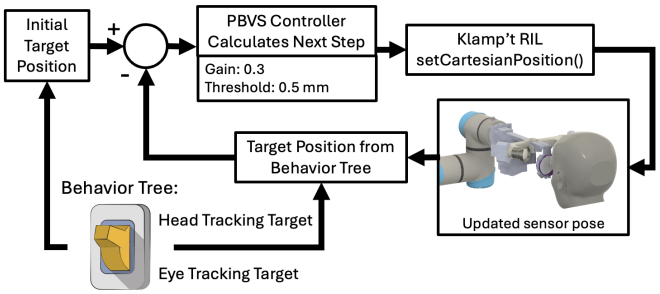

Fig. 3. PBVS diagram. The target position for PBVS is defined by the behavior tree, which will dynamically update the eye center target position from the head tracker to the pupil target position from the pupil tracker. In an error state, the behavior tree will send the most recently tracked position for 15 iterations before returning to the retracted state.

TABLE I
COMPARISON OF FACE KEYPOINT DETECTION METHODS

Method	Dropped (%) ↓	Quality (%) ↑	Rate (fps) ↑
3DDFA-V2 [14]	7.56	89.4	15.15±0.90
3DDFA-V3 [8]	**0.0**	**97.3**	11.46±0.80
OpenFace2.0 [13]	77.65	0.0	16.71±3.01
MediaPipe [15]	10.89	0.0	**24.70±2.83**

* all intervals reported as one standard deviation

account for non-orthogonality between the x'-axis and the y-axis, which can occur due to face asymmetry.

We fit the ldm106 facial landmarks [29] to the face and extract landmarks 52, 61, 75, and 84, corresponding to the nasion, nose tip, and center of the right and left eyes respectively. When keypoints are detected in both frames of our head tracker, we use keypoint triangulation to resolve the 3D position of the detected facial keypoints relative to the robot end effector, and use these keypoints to compose T_h in 3D (Fig. 1, Head Tracking). We use 3DDFA-V3 [8], a new model capable of face-fitting even in the context of extreme viewing angles, distortions, and expressions as seen in our system.

2) Pupil Tracking: Pupil tracking is used for short-range, higher-accuracy tracking. Like head tracking, pupil tracking requires keypoint detection and triangulation. We apply an adaptive Otsu threshold to perform binary segmentation of the image, effectively separating the darker regions (pupil and eyebrows) from brighter regions (skin and sclera). A Hough circle transform is empirically tuned for pupil detection. This approach results in a single circle surrounding the iris when unobstructed by the eyelid, with the center located approximately at the pupil in each of the endoscope camera views (Fig. 1, Pupil Tracking). When the pupil is detected in two or more cameras, we use triangulation relative to the robot end effector to resolve the pupil center. We then transform the pupil center into world coordinates to use in visual servoing.

D. System Control and Behavior Tree

We implement a PBVS controller for our eye-in-hand system in which the end effector approaches the 3D desired position estimated by either the head tracker or pupil tracker

TABLE II
ESTIMATED HEAD AND PUPIL TRACKER MEASUREMENT PRECISION RELATIVE TO FIXED AND FREESTANDING INDIVIDUALS

Dataset	Pupil P*	Head P*	Roll (°)	Pitch (°)	Yaw (°)
Rigid Phantom	±0.439	±0.141	±0.149	±0.404	±0.265
Held Phantom	±3.175	±2.522	±0.431	±1.132	±0.730
Chinrest Human	±0.575	±0.913	±0.480	±1.406	±0.618
Standing Human	±1.248	±2.001	±0.492	±1.393	±0.512

* P = Precision (mm) ** all intervals reported as one standard deviation

(Fig. 3) [7]. We define the desired end effector position as P_d, the current end effector position as P_c and the error as $e = P_d - P_c$. We use a proportional controller: $v = K \cdot e$, where v is the velocity of the end effector and K is the scalar gain. Two targets for P_d are defined as $C_e - t$, the eye center C_e minus the translation t from the end effector to lens focal point, and $C_p - t$, the pupil center minus the translation t. For smooth robot motion, we set a dead zone $||(C_e - t) - P_c||_2$ to 2.0 mm, and limit the maximum velocity to 25% of the robot's maximum Cartesian velocity. For servoing to the pupil, the dead zone for $||(C_p - t) - P_c||_2$ is tightened to 0.5 mm, while maintaining the same maximum velocity as in head-point servoing. Head-lock and pupil-lock are achieved when P_d is within the dead zone of the respective head and pupil servoing conditions. We perform translation-only servoing at 250 Hz, constraining end effector rotation to the identity matrix.

We address synchronization issues in software by associating tracking results with time-stamped camera poses, and we maintain a queue of past robot poses. When a tracking result arrives, we retrieve the robot pose with the closest timestamp in the queue to perform triangulation, ensuring that the position of the triangulated point accurately reflects the offset from the camera pose.

Our behavior tree contains four stages: move to home, approach C_e, head-lock to C_e, and pupil-lock to C_p. In normal behavior, a patient stands in front of the robot with the robot at the home position. When the head tracker detects T_h consistently for three seconds, the robot begins a safe approach to the patient's left eye C_e with an offset of 6.0 cm between the lens flange and the eye center, corresponding to the working distance of the lens. Once the robot reaches the head tracker dead zone, C_p tracking begins. If the pupil tracker detects C_p consistently for one second, the PBVS target switches from C_e to the more rapidly and accurately tracked C_p. After successfully tracking and servoing to C_p for two seconds, the clinical sensor begins recording a 10 second diagnostic image sequence of the retina.

If pupil detection fails for three or more cameras, servoing will fall back to C_e. If head tracking is also lost, the robot will return to the safe home position and restart the behavior tree. Once all clinical images are captured bilaterally, the system will reset for the next patient.

TABLE III
TRACKED PBVS TARGET POSITION DEVIATION

Servoing Stage	Target Translation (mm)
Home Head Detection	± 1.383 (world)
Approach Head Landmark	± 1.856 (world)
Head Servoing Lock Target Error	2.591 ± 1.996 (focal point)
Pupil Servoing Lock Target Error	1.262 ± 1.706 (focal point)

* all intervals reported as one standard deviation

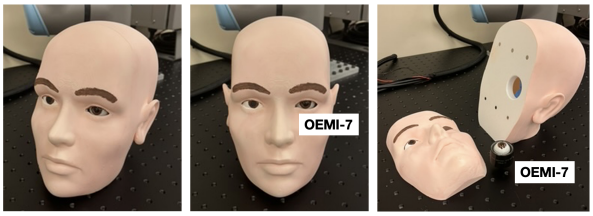

Fig. 4. Head and eye phantom with OEMI-7 Eye model embedded as the left eye. Used to test head and eye PBVS on the physical robot before testing on humans.

E. Clinical Image Filter

Due to patient movement and tracking error, retinal images during pupil lock may still exhibit incorrect focus, iris occlusion, and motion blur. To obtain a subset image sequence ready for examination by a medical professional, we train a binary classifier to filter out unclear images. We use a 489-frame video of the phantom eye collected by the clinical sensor during pupil-lock servoing. Annotations are manually provided by a medical student, with positive labels for images containing a well-illuminated retinal view in more than 75% of the lens. Using a 75-25 training-validation split, we applied a pretrained ResNet-34 model to the training frames [30]. A center crop is applied to the image sequence to eliminate optical artifacts at the edges. During training, random vertical and horizontal flips are applied to introduce variations in the training data. An Adam optimizer with a learning rate of 0.001 is used to minimize cross-entropy loss [31].

IV. EVALUATIONS AND RESULTS

To evaluate the system's adherence to our design constraints, we evaluate light safety, different keypoint detection approaches, 3D keypoint triangulation precision, PBVS accuracy, tracker latency, and image classification accuracy.

For tracking accuracy and retinal imaging experiments, we use a head and eye phantom with an Ocular Instruments OEMI-7 human eye model [32] (Fig. 4). We also perform tracking experiments on members of the research team.

A. Light Brightness Evaluation

Using a Thorlabs optical power meter, we measure the intensity of our lighting system at a maximum power lower than 1.5 W/cm^2 at the focal region of the lens, compliant with ANSI 2000 [27], indicating that our lighting system is safe for more than 10 s exposures incident at the cornea.

B. Head Tracker Keypoint Detection Evaluation

We compared performance of 3DDFA-V2 [14], 3DDFA-V3 [8], OpenFace2.0 [13], and MediaPipe [15] with our cameras on a 450-frame video sequence of large head movements of the phantom (approximately $\pm 10°$, ± 50cm in x, y, and z). We manually annotate the location of the nose tip, nasion, and the lateral canthus of the left and right eyes in all frames of the video, and calculate the distance between manual annotations and corresponding model prediction keypoints. We define high-quality frames when the mean distance between detected keypoints in the image and the ground truth is under 20 pixels and report the percentage of high-quality frames (Table I). OpenFace2.0 [13] and MediaPipe [15] are not capable of fitting keypoints to the face with our close-range head tracker. 3DDFA-V2 [14] often fails at the edges of the camera view.

C. System Precision and Accuracy Evaluations

Next, we evaluate the triangulation precision for the tracked points for both the head tracker and pupil tracker on both human participants and the phantom (Table II). The robot is positioned so that the left pupil is in front of the objective lens at approximately a 6 cm working distance. In the Rigid Phantom and Chinrest Human cases, respectively, the phantom and chinrest are mounted on a tripod during measurement. In the Held Phantom case, the phantom is held approximately still in a standing experimenter's hand, which mimics freestanding physiologic motion, and in Standing Human, the participant stands as still as possible. Precision is computed as the standard deviation of the position of C_e and C_p over a 1000-frame moving window. For the composed head transform in world coordinates, we evaluate the deviation in tracked roll, pitch, and yaw. The deviation is measured for both the head tracker and pupil tracker.

D. Latency Measurement

Processing time for each detector, from receipt of a keyframe to result, and overall tracker latency, from image acquisition to updating the PBVS target is measured and reported in Table IV. Data is collected from 1000 frames through a mixture of our phantom and members of the research team. Timing is broken down between head tracking and pupil tracking components. We observe that although computations for each camera are largely parallelized, and head detection is performed mostly on the GPU, there is still some synchronization overhead.

E. Servoing Accuracy

We report target position variations during the four stages of servoing motion: the measurement precision of C_e during head detection with the robot at the home position; the measurement precision of C_e while approaching the patient; the distance $||(P_c + t) - C_e||_2$ between the current lens focal point and the eye center provided by head tracking when head servoing; and the distance $||(P_c + t) - C_p||_2$ between the current lens focal point and the pupil center provided by pupil tracking when pupil servoing after target-transfer (Table III). This evaluation is conducted on both freestanding individuals and the handheld phantom for the duration of movements defined by the behavior tree.

TABLE IV
COMPONENT-LEVEL PROCESSING TIME AND LATENCY ANALYSIS

Component	Processing Time (ms)	Frequency (Hz)
Head Keypoint Detection	162.96 ± 13.76	-
Overall Head Tracker	180.81 ± 16.20	5.87 ± 1.47
Pupil Keypoint Detection	16.87 ± 2.29	-
Overall Pupil Tracker	33.78 ± 9.43	29.60 ± 2.42

* all intervals reported as one standard deviation

We show tracking regions in the XY and YZ planes in world coordinates to visualize the available tracking workspaces for the head tracker and pupil tracker (Fig. 5). We recorded a video sequence moving the phantom head throughout the workspace with the head in front of and facing toward the apparatus and report the regions as convex hulls of all detected keypoints projected into the respective planes. Monocular depth is estimated using the known distance between the eyes of the phantom, and triangulation results are used for the head and pupil tracking regions.

Observe that the pupil tracking region is contained entirely within the head triangulation region. This allows our method to seamlessly transition from head to pupil tracking, unlike the environment-mounted head tracking approaches of prior work [4, 5, 6]. We compare the reliable face tracking region for environment cameras as a pink cross-section in Fig. 5. These are determined by simulating two eye-level cameras 0.5 m away and oriented at a 45° angle to the left pupil when the head is at a nominal scanning location. Then, for each offset of the head in the XY and YZ planes, ray-casting is used to determine the percentage of the face occluded by the sensor array. Face tracking is considered reliable at a given offset if both cameras can see at least 90% of the face.

F. Binary Classifier Evaluation

The clinical image sequence binary classifier is evaluated on 2138 manually-annotated frames from three clinical image sequences collected during left-eye pupil-lock in a handheld phantom facing the sensor array. Our model demonstrates a classification accuracy of 90.97%. Our model achieves an F-statistic of 0.8833 and an AUC of 0.9896. In all frames selected by the model, the fundus of the phantom eye is visible, containing an in-focus optic disk and/or retinal vessels.

V. DISCUSSION

Our results show a robust tracking region for head tracking, low-latency pupil tracking, and collision-free stabilization at a 6 cm working distance. Qualitatively, our system is able to handle lost pupil tracking effectively, with a fallback to the head tracking method, and the ability to return to a safe retracted state if both detection methods fail (see Supplemental Video). Tracking is active during the entire approach phase, which reduces the risks of collision and failure to acquire the pupil target, and further improves the working range of the robot.

The accuracy of the system, measured on members of the research team, is sufficient for mydriatic (dilated) eye exams. Dilation is a common medical procedure, but comes with risks, including discomfort to the patient, altered vision, and more severe side-effects such as acute angle-closure glaucoma [33]. As a result, such exams are most often performed in clinics. To extend robot-assisted retinal imaging to routine screening with undilated patients, future work may place the patient in a darkened room, but more work is needed to perform tracking in the non-visible spectrum.

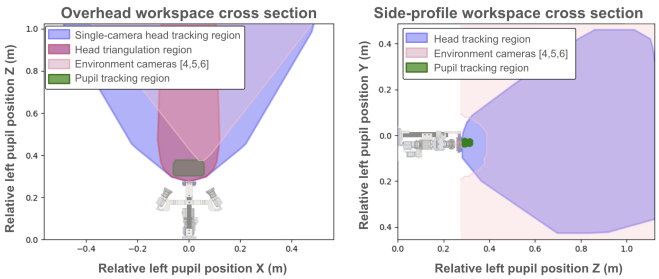

Fig. 5. Tracking region evaluation of the camera tracking modules with the phantom relative to the robot effector. The left is a top-down perspective of the tracking workspace in front of the effector, and the right is a side-profile of the workspace. The approximate robot servoing regions are illustrated as the position of the lens relative to the home position. Pink overlays represent head tracking occlusions with environment-mounted cameras, as seen in prior works [4, 5, 6], indicating the disconnect between visual servoing regions for head and pupil tracking.

VI. CONCLUSIONS AND FUTURE WORK

Our EYESIGHT system is a proof-of-concept wide-angle retinal examination tool for freestanding individuals. Our head and pupil trackers provides a step toward safe near-human robotic interaction, indicating that precise system motion is possible with 6 cm working distances for markerless, anatomy-based visual servoing. Notably, the complete tracking sensor package can be mounted entirely on a robot's end effector, enabling freestanding retinal imaging in a large workspace with low risk of tracking loss.

To extend our retinal inspection system to diverse populations, robustness to unexpected patient appearance and movement should be addressed. For example, pupil tracking may be adversely affected by hair occlusions and conditions that affect the external appearance of the cornea [34], requiring exploration of occlusion-robust pupil tracking. Head tracking can lose accuracy for patients with long hair, makeup, and abnormal facial physiology. Moreover, we are interested in augmenting the sensor head to include additional imaging modalities, e.g., slit-lamp examination. Non-dilated posterior segment examinations are also an exciting possibility, but such a system will need increased tracking accuracy and/or active management of pupillary response. Future work will also consider human-robot interactions, such as physician controls, verbal cues to assist patients, and aesthetic improvements to ease patient interaction.

VII. ACKNOWLEDGEMENTS

We thank Viktor Gruev and Yifei Jin from the UIUC BioSensors Lab for assisting in light safety testing with their light photometry devices.

References

[1] H. Schneiderman, "The Funduscopic Examination," eng, in *Clinical Methods: The History, Physical, and Laboratory Examinations*, H. K. Walker, W. D. Hall, and J. W. Hurst, Eds., 3rd, Boston: Butterworths, 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK221/ (visited on 10/31/2024).

[2] V. Biousse, B. B. Bruce, and N. J. Newman, "Ophthalmoscopy in the 21st century: The 2017 H. Houston Merritt Lecture," eng, *Neurology*, vol. 90, no. 4, pp. 167–175, Jan. 2018.

[3] C. Mishra and K. Tripathy, "Fundus Camera," eng, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK585111/ (visited on 10/31/2024).

[4] M. Draelos, P. Ortiz, R. Qian, B. Keller, K. Hauser, A. Kuo, and J. Izatt, "Automatic Optical Coherence Tomography Imaging of Stationary and Moving Eyes with a Robotically-Aligned Scanner," in *2019 International Conference on Robotics and Automation (ICRA)*, ISSN: 2577-087X, May 2019, pp. 8897–8903. [Online]. Available: https://ieeexplore.ieee.org/document/8793524 (visited on 11/13/2024).

[5] M. Draelos, P. Ortiz, R. Qian, C. Viehland, R. McNabb, K. Hauser, A. N. Kuo, and J. A. Izatt, "Contactless optical coherence tomography of the eyes of freestanding individuals with a robotic scanner," en, *Nature Biomedical Engineering*, vol. 5, no. 7, pp. 726–736, Jul. 2021, Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41551-021-00753-6 (visited on 10/28/2024).

[6] R. McNabb, P. Ortiz, K.-M. Roh, A. Song, M. Draelos, S. Schuman, G. Jaffe, E. Lad, J. Izatt, and A. Kuo, "Contactless, autonomous robotic alignment of optical coherence tomography for in vivo evaluation of diseased retinas," eng, *Research Square*, rs.3.rs–2 371 365, Jan. 2023.

[7] G. Flandin, F. Chaumette, and E. Marchand, "Eye-in-hand/eye-to-hand cooperation for visual servoing," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, ISSN: 1050-4729, vol. 3, Apr. 2000, 2741–2746 vol.3. [Online]. Available: https://ieeexplore.ieee.org/document/846442 (visited on 11/15/2024).

[8] Z. Wang, X. Zhu, T. Zhang, B. Wang, and Z. Lei, *3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation*, arXiv:2312.00311, Apr. 2024. [Online]. Available: http://arxiv.org/abs/2312.00311 (visited on 10/31/2024).

[9] G. Singh, W. W. J. Jie, M. T. Sun, R. Casson, D. Selva, and W. Chan, "Overcoming the impact of physiologic tremors in ophthalmology," en, *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 260, no. 12, p. 3723, Jul. 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9666294/ (visited on 10/28/2024).

[10] A. J. Thirunavukarasu, M. L. Hu, W. P. Foster, K. Xue, J. Cehajic-Kapetanovic, and R. E. MacLaren, "Robot-Assisted Eye Surgery: A Systematic Review of Effectiveness, Safety, and Practicality in Clinical Settings," en, *Translational Vision Science & Technology*, vol. 13, no. 6, p. 20, Jun. 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11210629/ (visited on 10/31/2024).

[11] L. J. Caffery, M. Taylor, G. Gole, and A. C. Smith, "Models of care in tele-ophthalmology: A scoping review," eng, *Journal of Telemedicine and Telecare*, vol. 25, no. 2, pp. 106–122, Feb. 2019.

[12] E. Tsui, A. N. Siedlecki, J. Deng, M. C. Pollard, S. Cha, S. M. Pepin, and E. M. Salcone, "Implementation of a vision-screening program in rural northeastern United States," eng, *Clinical Ophthalmology (Auckland, N.Z.)*, vol. 9, pp. 1883–1887, 2015.

[13] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 59–66. [Online]. Available: https://ieeexplore.ieee.org/document/8373812 (visited on 10/31/2024).

[14] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, *Towards Fast, Accurate and Stable 3D Dense Face Alignment*, arXiv:2009.09960, Feb. 2021. [Online]. Available: http://arxiv.org/abs/2009.09960 (visited on 10/31/2024).

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, *MediaPipe: A Framework for Building Perception Pipelines*, arXiv:1906.08172, Jun. 2019. [Online]. Available: http://arxiv.org/abs/1906.08172 (visited on 11/13/2024).

[16] S. Huang, K. Yang, H. Xiao, P. Han, J. Qiu, L. Peng, D. Liu, and K. Luo, "A new head pose tracking method based on stereo visual SLAM," *Journal of Visual Communication and Image Representation*, vol. 82, p. 103 402, Jan. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320321002698 (visited on 10/31/2024).

[17] X. Yu, Y. Zhang, H. Wu, and A. Wang, "An Improved Unscented Kalman Filtering Combined with Feature Triangle for Head Position Tracking," en, *Electronics*, vol. 12, no. 12, p. 2665, Jan. 2023, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2665 (visited on 10/31/2024).

[18] T. N. Cornsweet and H. D. Crane, "Accurate two-dimensional eye tracker using first and fourth Purkinje images," EN, *JOSA*, vol. 63, no. 8, pp. 921–928, Aug. 1973, Publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/josa/abstract.cfm?uri=josa-63-8-921 (visited on 10/31/2024).

[19] R.-J. Wu, A. M. Clark, M. A. Cox, J. Intoy, P. C. Jolly, Z. Zhao, and M. Rucci, "High-resolution eye-tracking via digital imaging of Purkinje reflections," en, *Journal of Vision*, vol. 23, no. 5, p. 4, May 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10166114/ (visited on 10/31/2024).

[20] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, vol. 199, p. 116 894, Aug. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422003347 (visited on 10/31/2024).

[21] E. Wood, T. Baltruaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '16, New York, NY, USA: Association for Computing Machinery, Mar. 2016, pp. 131–138. [Online]. Available: https://dl.acm.org/doi/10.1145/2857491.2857492 (visited on 10/30/2024).

[22] S. Porta, B. Bossavit, R. Cabeza, A. Larumbe-Bergera, G. Garde, and A. Villanueva, "U2Eyes: A Binocular Dataset for Eye Tracking and Gaze Estimation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, ISSN: 2473-9944, Oct. 2019, pp. 3660–3664. [Online]. Available: https://ieeexplore.ieee.org/document/9022577 (visited on 10/31/2024).

[23] Z. Yan, Y. Wu, Y. Shan, W. Chen, and X. Li, "A dataset of eye gaze images for calibration-free eye tracking augmented reality headset," en, *Scientific Data*, vol. 9, no. 1, p. 115, Mar. 2022, Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-022-01200-0 (visited on 10/31/2024).

[24] R. G. Bozomitu, A. Psric, V. Cehan, C. Rotariu, and C. Barabaa, "Pupil centre coordinates detection using the circular Hough transform technique," in *2015 38th International Spring Seminar on Electronics Technology (ISSE)*, ISSN: 2161-2064, May 2015, pp. 462–465. [Online]. Available: https://ieeexplore.ieee.org/document/7248041 (visited on 10/31/2024).

[25] J. Wang, G. Zhang, and J. Shi, "Pupil and Glint Detection Using Wearable Camera Sensor and Near-Infrared LED Array," eng, *Sensors (Basel, Switzerland)*, vol. 15, no. 12, pp. 30 126–30 141, Dec. 2015.

[26] I. Cordero, "Understanding and caring for an indirect ophthalmoscope," en, *Community Eye Health*, vol. 29, no. 95, p. 57, Feb. 2017. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5340107/ (visited on 11/14/2024).

[27] F. Delori, R. Webb, and D. Sliney, "Maximum permissible exposures for ocular safety (ANSI 2000), with emphasis on ophthalmic devices," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 24, no. 5, pp. 1250–1265, 2007.

[28] K. Hauser, "Robust Contact Generation for Robot Simulation with Unstructured Meshes," en, in *Robotics Research: The 16th International Symposium ISRR*, M. Inaba and P. Corke, Eds., Cham: Springer International Publishing, 2016, pp. 357–373. [Online]. Available: https://doi.org/10.1007/978-3-319-28872-7_21 (visited on 10/31/2024).

[29] Y. Liu, H. Shen, Y. Si, X. Wang, X. Zhu, H. Shi, Z. Hong, H. Guo, Z. Guo, Y. Chen, B. Li, T. Xi, J. Yu, H. Xie, G. Xie, M. Li, Q. Lu, Z. Wang, S. Lai, Z. Chai, and X. Wei, *Grand Challenge of 106-Point Facial Landmark Localization*, arXiv:1905.03469, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1905.03469 (visited on 10/31/2024).

[30] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385.

[31] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. [Online]. Available: https://arxiv.org/abs/1412.6980.

[32] A. R. Amorim, B. Bret, and J. M. Gonzlez-Mijome, "Opto-Mechanical Eye Models, a Review on Human Vision Applications and Perspectives for Use in Industry," en, *Sensors (Basel, Switzerland)*, vol. 22, no. 19, p. 7686, Oct. 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9573708/ (visited on 10/31/2024).

[33] D. Hong and K. Tripathy, "Tropicamide," eng, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK541069/ (visited on 11/16/2024).

[34] R. Shah, C. Amador, K. Tormanen, S. Ghiam, M. Saghizadeh, V. Arumugaswami, A. Kumar, A. A. Kramerov, and A. V. Ljubimov, "Systemic diseases and the cornea," en, *Experimental eye research*, vol. 204, p. 108 455, Jan. 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7946758/ (visited on 10/31/2024).