2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Shadow-Based 3D Pose Estimation of Intraocular Instrument Using Only 2D Images

Junjie Yang[1], Zhihao Zhao[1], Mathias Maier[2], Kai Huang[3], Nassir Navab[1], *Fellow, IEEE*, M. Ali Nasseri[2]

***Abstract*— In ophthalmic surgeries, such as vitreoretinal operations, surgeons rely on imaging systems, primarily microscopes, for real-time instrument monitoring and motion planning. However, novice surgeons struggle to extract 3D instrument positions from 2D microscope frames, necessitating extensive trial-and-error experience with the background that additional imaging modalities such as iOCT remain inaccessible in most operating rooms. Targeting intraocular assessment within the current surgical setup, this paper presents an image-based pose estimation method to obtain real-time instrument tip positions in a standard 12mm-radius spherical eyeball model, which links floating instruments with on-the-retinal objects based on the intraocular shadowing principle. We validate this estimation method in a Unity simulator and verify its depth estimation capability using a specially designed eyeball phantom. Both simulator and phantom experiments demonstrate an average needle-tip estimation error within [1.0, 2.0] mm using only 2D microscope frames.**

I. Introduction

Ophthalmic surgeries typically involve trocars as the penetration tunnels of surgical instruments, as illustrated in Fig. 1, leading to the instrument pivoting around trocars using a Remote Center of Motion (RCM) scheme. Meanwhile, surgeons heavily rely on microscope images for instrument manipulation. While there is a growing trend toward implementing intraocular Optical Coherence Tomography (iOCT) to enable micron-precision 3D intraocular perception, its adoption in ophthalmic operating rooms remains limited.

In the context of using microscopes, experienced surgeons can roughly estimate the instrument's intraocular pose after multiple trial-and-error movements, leveraging their implicit 2D-3D modeling skills. In contrast, novice surgeons cannot skillfully analyze instrument projections within microscope images and establish their methodology for 2D-3D transformations. This deficiency results in numerous trials to complete a single manipulation task, leading to undesirable surgical duration extensions. Consequently, there is a pressing need for an image-based pose estimation solution to aid surgeons in instrument manipulation.

When the instrument floats within the hollow vitreous area without direct contact with the retina, it is challenging to estimate the instrument-tip position only from its 2D projection, even with its shadow visible in the frame, to provide qualitative depth perception. This floating state persists for a significant portion of the surgical workflow and greatly influences subsequent motion planning for reaching specific intraocular targets. Consequently, achieving image-based 3D estimation poses a significant challenge, yet it holds substantial clinical significance.

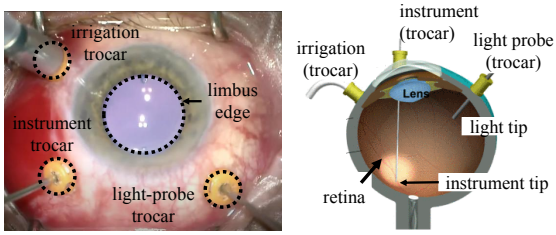


Fig. 1: Typical setup of ophthalmic surgeries [1].

This paper proposes an image-based method to analyze the 3D tip positions of intraocular objects within a standard 12mm-radius spherical eyeball model. Our approach utilizes the instrument's RCM-compatible motion trajectories to approximate the trocar's pixel position, serving as a reference for calculating the 2D-3D scaling ratio. Under the assumption of eyeball-microscope alignment, we estimate the lighted center on the retina to determine the light probe's 3D tip position. Then, we use light rays cast from the light-probe tip as the bridge to link floating instrument tips with relevant on-retina objects for the 3D position estimation. Meanwhile, we consider the surgical necessity of adjusting the microscope's magnification to obtain different Regions of Interest (ROI). Therefore, we decompose the estimation of invisible objects into the combination of visible image components in a hierarchical structure. Our method is tested in a Unity-based simulator, and its depth estimation capability is validated with a specially designed eyeball phantom. Our experiments consistently yield average absolute errors in the range of [1.0, 2.0] mm for both 3D position and depth.

The contributions of this paper are: 1) Emphasizing the crucial link between floating objects and their corresponding retinal objects through intraocular illumination. 2) A novel method that utilizes shadows to approximate the center point of the illuminated retinal area is introduced. 3) Proposing a comprehensive approach for estimating the 3D positions of floating points exclusively from 2D images.

II. Related Work

In recent years, a growing body of research has focused on leveraging iOCT's depth-sensing capabilities for intraocular 3D reconstruction and subsequent instrument navigation [2]–[10]. One of the persistent challenges in this domain is the

[1]School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany {junjie.yang,zhihao.zhao,nassir.navab}@tum.de
[2]Klinik und Poliklinik für Augenheilkunde, Klinikum rechts der Isar, 81675 Munich, Germany mathias.maier@mri.tum.de, ali.nasseri@tum.de
[3]School of Computer Science and Engineering, Sun Yat-Sen University, 510006 Guangzhou, China huangk36@mail.sysu.edu.cn

coordination of iOCT with the instrument to ensure that the instrument tip remains within the limited ROI. Despite efforts to introduce new modalities like spectrally encoded reflectometry (SER) [11], integrating these modalities into surgical practice still requires time-consuming validations. On the other hand, image-based approaches have been explored extensively. Many of these methods involve the use of laser or structured light to create target indications on the retina for instrument navigation [12]–[16]. However, these approaches raise concerns about illumination toxicity, instrument collisions, and additional sensor calibration. Some works have employed stereo-imaging systems for instrument pose estimation [17] or depth estimation [18]. These methods often rely on the calibration of stereo lenses to handle focus adjustments during surgery. While there have been attempts to use neural networks for automatic retina approaching and instrument trajectory prediction [19]–[21], these approaches face challenges related to diverse data collection and explainability. In contrast, Koyama *et al.* propose using surgical robots to explore depth information and approach a given retinal target [22]. However, this approach requires additional calibration between the robot and the eyeball and complex kinematic calculations for depth data generation. This paper introduces a 3D estimation method that offers a distinct advantage. This method only requires eyeball-microscope calibration and manual trocar-distance measurement at the initial surgery stage, making it fully compatible with the current surgical setup in the operating room.

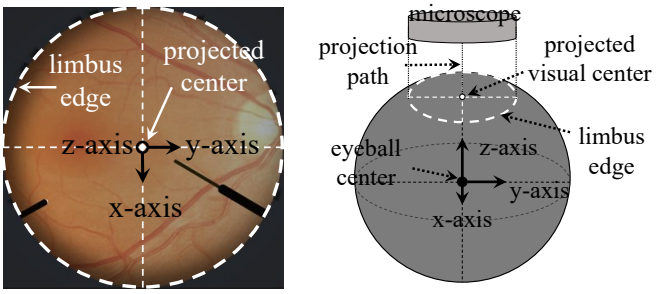
III. Prerequisite

A. Eyeball-Microscope Alignment

To ensure accurate visual projection from 3D instrument objects to 2D image components, maintaining the alignment between the eyeball and the microscope is paramount by keeping the eyeball static, with its transparent circle-shaped limbus area sharing the same center as the microscope's perspective, as illustrated in Fig. 2. This alignment crucially ensures that the vertical projection path of each object becomes perpendicular to the imaging plane, guaranteeing a direct correspondence between an object's 3D position and its pixel-based 2D position in the image. This eyeball-microscope alignment can be maintained during ophthalmic surgeries by combining feature-based rotation assessment with eyeball orbit control [23].

B. Spherical Motion Modeling

To accurately represent the instrument's motion inside the eyeball under the RCM constraint, a spherical coordinate system is employed with its origin at the trocar as shown in Fig. 3, where the instrument's pose is described by three axes: polar angle θ^v, azimuthal angle θ^h and radius r. The correspondence between the instrument's tip position in the local spherical coordinate system and the global Cartesian coordinate system is also established, with the latter's origin at the eyeball center. This spherical RCM-compatible motion modeling enables the characterization of changes in polar



(a) Top view of limbus area. (b) Eyeball-microscope alignment.

Fig. 2: The visual alignment between the microscope and the eyeball's limbus area.

angle, azimuthal angle, and radial length as distinct components: vertical RCM rotation (VRCM), horizontal RCM rotation (HRCM), and axial zooming (ZOOM), respectively. It is important to note that VRCM maintains the instrument's projected orientation while HRCM ideally preserves the same vertical distance from the instrument tip to the trocar.

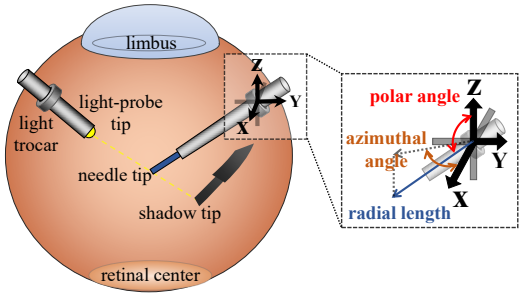


Fig. 3: Spherical modeling of RCM manipulations.

IV. Pose Estimation

A. Requirement and Assumption

In this paper, we consider the instrument a needle for method description. The necessary components of this estimation method are listed below:

- trocar: RCM point (upper half of eyeball surface);
- light probe: source of intraocular illumination (floating);
- lighted center: the intersection of the light probe's center line and the retinal surface as the center of the lighted retinal area (lower half of eyeball surface);
- needle: instrument responsible for interacting with tissues (floating);
- shadow: instrument's shadow on the retina (lower half of eyeball surface).

In addition, this method relies on the following assumptions: (a) the eyeball is a perfect 12mm-radius sphere; (b) camera distortion is calibrated; (c) the cone-shaped lighting range shares the same axial line with the light-probe shaft; (d) the needle's shaft and tip can be modeled together as a coaxial cylinder; (e) image segmentation is already finished with accessible components.

The general procedure of 3D-position estimation is depicted in Fig. 4. It is worth noticing that surgeons are required to measure the limbus diameter and the planar distance between two trocars as the ground-truth reference

of later 2D-3D scaling calculation. After the fixation of the magnification level, the 2D-3D scaling ratio is updated as the fixed 3D modeling parameter in the corresponding magnification scenario. Afterward, the light probe's trocar and the fixed lighted center are estimated using a proposed image-component decomposition method. Finally, the needle-tip estimation in the fixed magnification scenario can be achieved using the prior result of image segmentation. The implementation details of our method are described below.

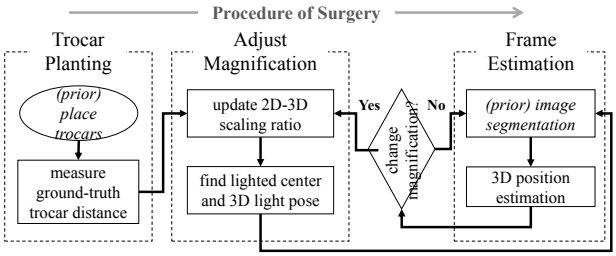


Fig. 4: The procedure of 3D-position estimation.

B. 2D-3D scaling

For objects situated on the spherical eyeball surface (including trocars $p_{lrcm/nrcm}$, needle's shadow p_{ns} and lighted center $p_{lcenter}$), their 3D position can be derived from their corresponding pixel position (x, y) in the image following Equation (1) using the 2D-3D scaling ratio $\sigma_{2d\to3d}$.

$$\begin{aligned} \text{Given } \sigma_{2d\to3d},\ x^{3d} &= (y - y_{ref}) \cdot \sigma_{2d\to3d} \\ y^{3d} &= (x - x_{ref}) \cdot \sigma_{2d\to3d} \\ z^{3d} &= \sqrt{r_{eye}^2 - (x^{3d})^2 - (y^{3d})^2} \cdot \rho_p \end{aligned} \tag{1}$$

(x_{ref}, y_{ref}) denotes the pixel coordinates of the limbus center, which also serves as the projected eyeball center in the image. Additionally, ρ_p denotes the location of points on the eye's surface, with $\rho_p = 1$ indicating the upper half of the eye and $\rho_p = -1$ the lower half. Under the assumption (b) of negligible camera distortion, Equation (2) ideally holds as the calculation method of the scaling ratio $\sigma_{2d\to3d}$.

$$\sigma_{2d\to3d} = \frac{d_{limbus}^{mm}}{d_{limbus}^{pixel}} = \frac{d_{lrcm\to nrcm}^{mm}}{d_{lrcm\to nrcm}^{pixel}}. \tag{2}$$

d_{limbus} is the diameter of limbus circle, while $d_{lrcm\to nrcm}$ denotes the planer distance between light-probe and needle trocars. The ground-truth distances d_{limbus}^{mm} and $d_{lrcm\to nrcm}^{mm}$ in millimeters can be manually measured during preoperative OCT scanning or the intraoperative trocar-placement phase. Subsequently, their corresponding pixel-based distances d_{limbus}^{pixel} and $d_{lrcm\to nrcm}^{pixel}$ are computed from microscope images during the surgical procedure.

Considering that surgeons may adjust the microscope's magnification level to focus on a smaller visible area as shown in Fig. 5, the 2D-3D scaling ratio $\sigma_{2d\to3d}$ requires updating by measuring either d_{limbus}^{pixel} or $d_{lrcm\to nrcm}^{pixel}$ depending on the visibility of the complete limbus edge. In a typical scenario where the limbus circle edge is completely visible, it is straightforward to segment the retinal area and calculate its pixel radius relative to the ground-truth radius (e.g., a 6-mm limbus radius in the ideal model). However, in scenarios where the limbus edge is not entirely visible, locating the latest trocar positions within the image and measuring the distance after performing 2D-3D conversion is necessary.

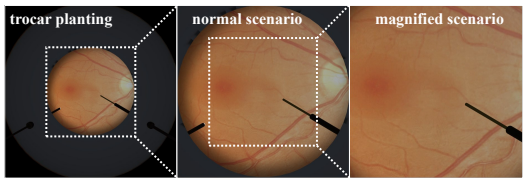


Fig. 5: Three scenarios of microscope magnification.

According to the RCM motion constraints, locating a trocar's 2D position in the imaging plane can be decomposed into a hierarchical structure shown in Fig. 6 until the leaf step of needle-shadow detection. First, the trocar's position

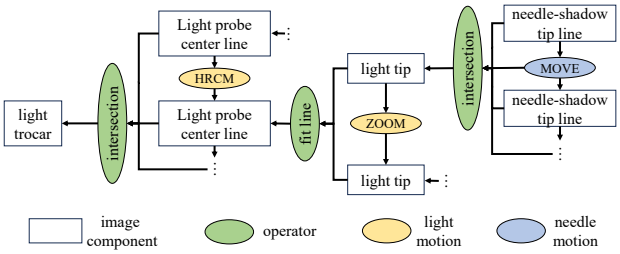


Fig. 6: The decomposition of locating the light trocar.

is represented as the intersection of multiple object (light probe or needle) lines in compliance with the global RCM definition. Then, each light-probe line is fitted with at least two light-probe tips obtained by ZOOM movement, which obeys the spherical coordinate modeling. Finally, the light tip is factorized into the intersection of multiple needle-shadow tip lines (the line passing both needle tip and shadow tip) guided by the principles of shadowing. Therefore, the visibility of the needle tip and its shadow tip should be guaranteed, while other invisible objects can be inferred by the decomposition above.

It is important to note that the light probe's HRCM motion is not strictly necessary, as any change in the light probe's orientation contributes to the trocar estimation, except for the light probe's pure VRCM rotation that overlaps all light-probe lines. Similarly, the needle movement for estimating the light-probe tip should avoid tracing the lighting path to obtain the same needle-shadow tip line. After accumulating a sequence of object lines $[l_0, ..., l_{m-1}]$, the trocar point p_{rcm} in the image is estimated as a least-square problem to be solved by Equation (3).

$$\begin{aligned} \mathbf{E}(p_{rcm}) &= \sum_{i=0}^{m-1} (p_{rcm} - p_i)^T (\vec{n}_i \vec{n}_i^T)(p_{rcm} - p_i) \\ \frac{\partial \mathbf{E}(p_{rcm})}{\partial p_{rcm}} &= 0,\ p_{rcm} = \left(\sum \vec{n}_i \vec{n}_i^T\right)^{-1} \left(\sum \vec{n}_i \vec{n}_i^T p_i\right) \end{aligned} \tag{3}$$

where $\vec{v}$ and $\vec{n}$ represent an object line's direction vector and normal direction vector. Subsequently, we convert this pixel trocar position p_{rcm} into its corresponding 3D position p_{rcm}^{3d}.

C. Finding Lighted Center

Based on the assumption (c), the light probe's lighting range is modeled by a cone with its apex at the light-probe tip, and its orientation is defined by the axial line of the light-probe shaft l_{lp}. This line l_{lp} forms a light ray starting from the light-probe tip and hitting the retinal surface at a point defined as the lighted center $p_{lcenter}$.

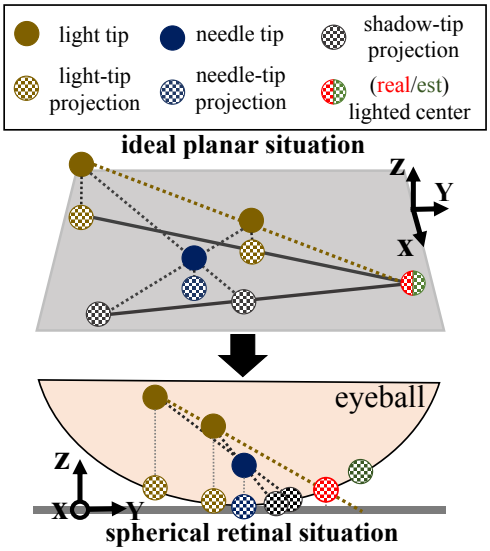


Fig. 7: Analysis of lighted-center estimation.

Leveraging the shadowing principle and assumption (b), we observe that the light-probe tip point p_{lp}, the needle tip p_n and its shadow tip p_{ns} always lie on the same light ray $l_{lp\to ns}$, which means the light probe determines the shadow position with a static needle placement. Considering the relationship between $l_{lp\to ns}$ and l_{lp}, two cases are defined based on the shadow's movement when the light probe axially zooms towards the retina without changing the lighted center:

- $l_{lp} \nparallel l_{lp\to n} \mapsto p_{lcenter} \neq p_{ns}$: shadow tip moves and its area scales.
- $l_{lp} \parallel l_{lp\to n} \mapsto p_{lcenter} = p_{ns}$: shadow tip remains stationary while its area scales.

In the case of $p_{lcenter} \neq p_{ns}$, the intraocular shadowing scenario is modeled as depicted in Fig. 7 with the condition that the light probe is positioned above the needle tip to ensure shadow visibility within the image. Assuming the retinal surface is flat, the shadow trajectory intersects with the light probe's projected insertion path at the lighted center, where the light probe's trajectory hits the retinal surface. In this flat-surface case, the estimated lighted center collides with the real lighted center at the abovementioned intersection point. However, the real retinal surface's spherical curvature introduces a deviation between the real lighted center (hitting the retina) and the estimated lighted center (hitting the plane). This deviation increases as the lighted center moves further from the retina center. In the context of a standard spherical eyeball model, the 6mm-radius visible limbus area results in a maximal z-axis deviation of 1.607 mm from the retinal center bottom to a visible point on the retina. This deviation trend also applies to the estimation of the shadow tip. Therefore, selecting the central retinal bottom as the range for lighted centers and shadow tips is recommended to enhance the accuracy of object position estimation. This selection also capitalizes on the approximate parallelism between the retinal and the planar shadowing surfaces in the central visible area.

Considering the above observations, a sequence of light probe's radial length $\{r_{lp}^0, ..., r_{lp}^{m-1}\}$ obtained by ZOOM generates a light-probe trajectory $\mathcal{T}_{lp} = \{p_{lp}^0, ..., p_{lp}^{m-1}\}$ to also obtain the corresponding shadow trajectory $\mathcal{T}_{ns} = \{p_{ns}^0, ..., p_{ns}^{m-1}\}$. Subsequently, $p_{lcenter}$ is approximated as the intersection point of two trajectories as Equation (4).

$$\begin{aligned} \mathcal{T}_{lp} &= \{p_{lp}^0, ..., p_{lp}^{m-1}\} = l_{lp} \\ \mathcal{T}_{ns} &= \{p_{ns}^0, ..., p_{ns}^{m-1}\} \approx l_{ns}^{shift} \\ p_{lcenter} &= l_{ns}^{shift} \wedge l_{lp} \end{aligned} \tag{4}$$

In the case of $p_{lcenter} = p_{ns}$, it is straightforward to utilize the shadow tip p_{ns} as the most accurate approximation for the lighted center $p_{lcenter}$.

D. Instrument Estimation

After converting the lighted center from pixel to 3D position, the light-probe's 3D orientation and its tip position are calculated as Equation (5) ($\sigma = \sigma_{2d\to 3d}$), provided that the tip is visible in the image.

$$\begin{aligned} \vec{l}_{lp}^{3d} &= \text{normalize}(p_{lcenter}^{3d} - p_{lrcm}^{3d}) \\ p_{lp}^{3d} &= p_{lrcm}^{3d} + \vec{l}_{lp}^{3d} \cdot (\|p_{lrcm} - p_{lp}\| \cdot \sigma \cdot \frac{\|\vec{l}_{lp}^{3d}\|}{\|\vec{l}_{lp}^{3d,xy}\|}) \end{aligned} \tag{5}$$

This calculation involves the conversion of the pixel-based distance between the light trocar to the light-probe tip into the millimeter-based planar distance and, subsequently, into the 3D distance, utilizing the ratio of the direction vector's 3D norm to its 2D norm (in the xy plane). In cases where the light probe's tip is not visible in the image, an alternative approach is employed to estimate the light probe tip as illustrated in Fig. 6.

Once the light probe tip's 3D position is obtained, we utilize the shadowing light ray to establish the connection between the needle tip and its shadow tip, following the proposed estimation guideline. Then, the same procedure of distance conversion is applied to estimate the needle-tip position as Equation (6) ($\sigma = \sigma_{2d\to 3d}$).

$$\begin{aligned} \vec{l}_{ns\to lp}^{3d} &= \vec{l}_{ns\to n}^{3d} = \text{normalize}(p_{lp}^{3d} - p_{ns}^{3d}) \\ p_n^{3d} &= p_{ns}^{3d} + \vec{l}_{ns\to n}^{3d} \cdot (\|p_n - p_{ns}\| \cdot \sigma \cdot \frac{\|\vec{l}_{ns\to n}^{3d}\|}{\|\vec{l}_{ns\to n}^{3d,xy}\|}) \end{aligned} \tag{6}$$

While it is possible to use p_{nrcm} for calculating the needle tip's 3D position, it is preferred to use the shadow to avoid introducing additional sources of error.

V. Experiment

A. Simulation

To assess the performance of our method, we employed a private Unity-based simulator as shown in Fig. 8 to generate microscope images and corresponding ground-truth position data. Both normal scenarios and magnified scenarios are tested with different datasets. The simulated images are sized as 1024x1024 pixels with the projected eyeball center at [512, 512] in the image.

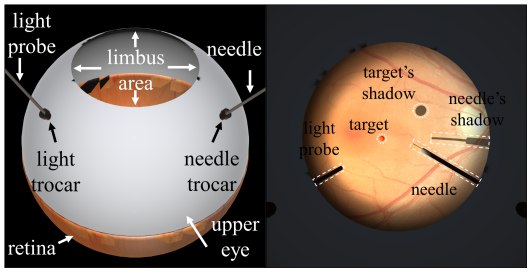


Fig. 8: The Unity-based simulator for testing.

For the normal scenario, a total of 1000 random scenarios are generated. Each scenario consists of 50 images (1024x1024 pixels). Although only 4 compulsory frames are needed for the lighted-center estimation according to the decomposition in Fig. 6, we add 6 more frames to form a set of 10 microscope images to relieve the error of intersection calculation during the fixed lighted-center estimation. Then, the remaining 40 images are used to estimate randomly placed needles, resulting in a total of 40000 cases for position estimation. In the normal scenario, the magnification level remains constant. Therefore, the light trocar's pixel and 3D positions are already calculated as [-170, 895] and [4.48828, -7.99219, 7.74469]^T, respectively. The light trocar's 3D-position estimation error is $\approx$ 0.01355 mm.

The successful rate of our method in 1000 simulated scenarios, encompassing 40000 cases, is presented in TABLE I. Note that the lighted center's estimation is a crucial prerequisite for estimating other objects. Hence, if the lighted center's position is already outside the spherical eyeball in a scenario, all random cases in this scenario are marked as failures. Cases with failed calculation (i.e., invalid shadow-trajectory approximation) are considered failures. Moreover, since the common scanning size of iOCT is 6 mm, cases with estimation errors exceeding 6 mm are also considered failures. The mean, variance, minimal, and maximal error of 40 cases in each scenario are collected to represent the method performance in that scenario, and the distribution of 3D estimation error for each object in 958 successful normal scenarios is listed in TABLE II. We succeeded in the estimation procedure in 95.8% scenarios and 94.995% random cases with all mean errors within 0.5 mm, which demonstrates the effectiveness of our proposed position estimation method in the normal scenario. However, the approximation of the shadow trajectory under the influence of retina curvature, as discussed in Section IV-C, leads to unstable lighted-center estimation and causes the maximal 5.8949-mm error. Meanwhile, the thickness of the retina model in the simulator is the leading cause of zero variance (too small) for the shadow-tip estimation error.

TABLE I: Summary of Simulated Normal Scenarios

	total	successful	failed
scenario	1000	958 (95.8%)	42(4.2%)
case	40000	37998 (94.995%)	2002(5.005%)

TABLE II: 3D Error (mm) of Successful Normal Cases

object	mean	variance	min	max
light trocar (static)	0.0136	-	-	-
light probe (tip)	0.2002	0.0803	0.0055	2.8420
lighted center	0.3878	0.2845	0.0008	5.8949
needle (tip)	0.0941	0.0240	0.0007	2.6104
shadow (tip)	0.0191	0.0000	0.0010	0.0322

For the magnified scenario, 1000 random scenarios are generated, each with varying microscope magnification levels ranging from 1x to 3x. In each magnified scenario, 50 images (1024x1024 pixels) are captured. Due to the variable random magnification, additional components such as the light trocar p_{lrcm}, needle trocar p_{nrcm}, and the sporadically invisible light-probe tip p_{lp} are also in need of estimation except for the lighted center. Therefore, we also add 4 frames to the compulsory 8 microscope frames to form a set of 12 frames for identifying essential components following the steps in Fig. 6. Subsequently, the remaining 38 images are used for the position estimation. Cases with failed 2D-3D calculation or estimation errors exceeding 6 mm are labeled as failures. The successful rate of magnified scenarios is presented in TABLE III, and a detailed performance analysis is available in TABLE IV. We succeeded in the estimation procedure in 81.7% scenarios and 71.066% random cases with all mean errors $\leqslant$ 2.1 mm, especially $\leqslant$ 1.0 mm for the needle tip even with the light probe occasionally invisible in the image.

TABLE III: Summary of Simulated Magnified Scenarios

	total	successful	failed
scenario	1000	817 (81.7%)	183(18.3%)
case	38000	27005 (71.066%)	10995 (28.934%)

B. Phantom Test

A phantom-based test is also conducted to rapidly validate the depth-estimation function in a normal scenario using the hardware setup as shown in Fig. 9. A spherical eyeball phantom (12-mm radius) is designed for the experiment, with its lower 1/4 part cut to facilitate side-view-based

TABLE IV: 3D Error (mm) of Successful Magnified Cases

object	mean	variance	min	max
light trocar	1.1006	0.8072	0.0047	5.9704
needle trocar	0.9992	0.7320	0.0019	5.1111
light probe (tip)	2.0359	2.2591	0.3217	5.9690
lighted center	1.5143	0.6794	0.3420	5.4916
needle (tip)	0.8644	0.6793	0.0085	5.8579
shadow (tip)	0.1511	0.0225	0.0012	1.1433

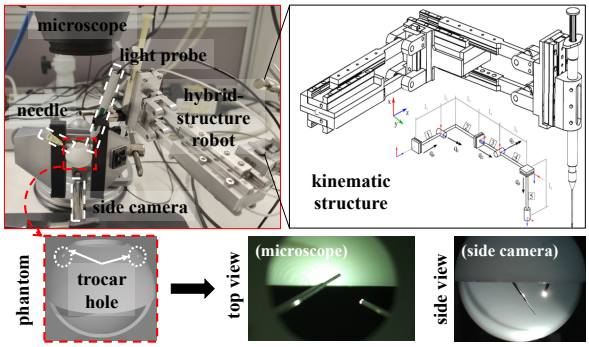


Fig. 9: The hardware setup of a 3D-printed phantom eyeball.

depth calculation. Meanwhile, the eyeball's top is also cut to enable a 6-mm-radius limbus circle area for the top-view monitoring. Two perspectives are employed to observe the instruments' positions within the eyeball phantom independently: a microscope from the top and a camera from the side (facing the 1/4-cut area). A simple manual camera-eyeball calibration is achieved to prepare the experiment environment quickly, which should be improved by rigorous marker-based calibration in the future.

TABLE V: Error (mm) of Phantom-Based Depth Estimation.

object	mean	variance	min	max
light tip	-1.5832	0.0485	-3.3903	-1.1840
needle tip	-1.5288	0.1939	-3.4914	-0.0432

62 images are collected along the light probe's radial insertion for the lighted-center estimation. After having the lighted center, the light probe remains static, and the needle tip is randomly placed to generate 100 images. All images are manually labeled to have precise pixel-based object positions. To facilitate scaling calculation for both views, we label the edge points of the limbus area in the top view and the edge points of the spherical retina in the side view. As for the performance evaluation, the depth estimation results from top-view images and side-view images are separately calculated, and the depth error is calculated as top-view depth minus side-view depth using the top view as the estimation and the side view as the ground truth. We provide a summary of estimation-reference error (top-view depth minus side-view depth) in Table. V, and its distribution is illustrated in Fig. 10. Although the depth estimated from the top view is averagely deeper than the side-view depth, which may result from camera distortion and light-range modeling, the average absolute depth error within [1.0, 2.0] mm still proves the method's effectiveness.

Compared with other methods involving additional modalities, the proposed image-only pose estimation method achieves an average estimation error within [1.0, 3.0] mm according to simulation and phantom tests. Suppose the needle tip is the focused measurement target due to its vital role during the instrument-tissue interaction. In that case, its average 3D estimation error is already reduced within [1.0, 2.0] mm with the potential of further improvement by better light modeling. Meanwhile, the estimation error of the magnified scenario is larger than the error of the normal scenario, which is caused by the invisibility of instruments in the image to force the decomposition of trocars and the light probe's estimation and amplify the estimation error level by level. Considering the precision of [1.0, 2.0] mm for position

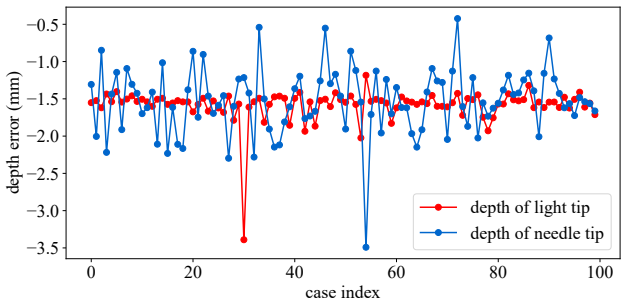


Fig. 10: The depth-error distribution by two camera views.

estimation, the proposed method still has room for improvement. However, since the surgical target is to estimate the floating instrument within the hollow vitreous area without touching the retina, intraoperative visualization and rough intraocular assessment can benefit from the presented 3D estimation.

VI. Conclusion

In this paper, we emphasize the importance of the light probe in connecting intraocular floating objects with their shadows on the retina. We propose a 3D estimation method using only 2D images that seamlessly integrates into the existing surgical workflow without involving additional modalities. Our approach can be further enhanced by using higher-resolution cameras to obtain a smaller millimeter-pixel ratio for better segmentation results. The assumption of perfect eyeball shape can also be relieved by using preoperative OCT/CT scanning to obtain the eyeball shape as a geometry prior. However, the static alignment between the microscope and the eyeball limits handling expansive retina areas when the orbit control is used to rotate the eyeball instead of maintenance. In conclusion, the presented approach is promising for enhancing ophthalmic surgery and benefiting novice surgeons and surgical automation, underscoring its meaningful contributions from clinical and engineering perspectives.

REFERENCES

[1] M. Simunovic. Sub-retinal tissue plasminogen activator for sub-macular haemorrhage. Youtube. [Online]. Available: https://www.youtube.com/watch?v=OKYqsVz3Maw&ab_channel=MatthewSimunovic

[2] M. Zhou, J. Wu, A. Ebrahimi, N. Patel, Y. Liu, N. Navab, P. Gehlbach, A. Knoll, M. A. Nasseri, and I. Iordachita, "Spotlight-based 3d instrument guidance for autonomous task in robot-assisted retinal surgery," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7750–7757, 2021.

[3] B. Keller, M. Draelos, G. Tang, S. Farsiu, A. N. Kuo, K. Hauser, and J. A. Izatt, "Real-time corneal segmentation and 3d needle tracking in intrasurgical oct," *Biomedical optics express*, vol. 9, no. 6, pp. 2716–2732, 2018.

[4] S. Dehghani, M. Sommersperger, P. Zhang, A. Martin-Gomez, B. Busam, P. Gehlbach, N. Navab, M. A. Nasseri, and I. Iordachita, "Robotic navigation autonomy for subretinal injection via intelligent real-time virtual ioct volume slicing," *CoRR*, vol. abs/2301.07204, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.07204

[5] M. Zhou, K. Huang, A. Eslami, H. Roodaki, H. Lin, C. P. Lohmann, A. Knoll, and M. A. Nasseri, "Beveled needle position and pose estimation based on optical coherence tomography in ophthalmic microsurgery," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2017, pp. 308–313.

[6] J. Weiss, N. Rieke, M. A. Nasseri, M. Maier, A. Eslami, and N. Navab, "Fast 5dof needle tracking in ioct," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 787–796, 2018.

[7] H. Yu, J.-H. Shen, R. J. Shah, N. Simaan, and K. M. Joos, "Evaluation of microsurgical tasks with oct-guided and/or robot-assisted ophthalmic forceps," *Biomedical optics express*, vol. 6, no. 2, pp. 457–472, 2015.

[8] J. P. Ehlers, A. Uchida, and S. K. Srivastava, "Intraoperative optical coherence tomography-compatible surgical instruments for real-time image-guided ophthalmic surgery," *British Journal of Ophthalmology*, vol. 101, no. 10, pp. 1306–1308, 2017. [Online]. Available: https://bjo.bmj.com/content/101/10/1306

[9] M. Zhou, X. Wang, J. Weiss, A. Eslami, K. Huang, M. Maier, C. P. Lohmann, N. Navab, A. Knoll, and M. A. Nasseri, "Needle localization for robot-assisted subretinal injection based on deep learning," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8727–8732.

[10] M. Sommersperger, J. Weiss, M. A. Nasseri, P. Gehlbach, I. Iordachita, and N. Navab, "Real-time tool to layer distance estimation for robotic subretinal injection using intraoperative 4d oct," *Biomed. Opt. Express*, vol. 12, no. 2, pp. 1085–1104, Feb 2021. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-12-2-1085

[11] E. M. Tang, M. T. El-Haddad, S. N. Patel, and Y. K. Tao, "Automated instrument-tracking for 4d video-rate imaging of ophthalmic surgical maneuvers," *Biomed. Opt. Express*, vol. 13, no. 3, pp. 1471–1484, Mar 2022. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-13-3-1471

[12] C. He, E. Yang, N. Patel, A. Ebrahimi, M. Shahbazi, P. Gehlbach, and I. Iordachita, "Automatic light pipe actuating system for bimanual robot-assisted retinal surgery," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 6, pp. 2846–2857, 2020.

[13] H. Moon, D. Lee, and D. Nam, "Trimanual technique using assistant-controlled light probe illumination and wide-angle viewing system in 23-gauge sutureless vitrectomy for diabetic tractional retinal detachment," *Ophthalmic surgery, lasers & imaging retina*, vol. 46, pp. 73–6, 01 2015.

[14] T. C. Hutchens, A. Darafsheh, A. Fardad, A. N. Antoszyk, H. S. Ying, V. N. Astratov, and N. M. Fried, "Characterization of novel microsphere chain fiber optic tips for potential use in ophthalmic laser surgery," *Journal of Biomedical Optics*, vol. 17, no. 6, pp. 068 004–068 004, 2012.

[15] B. C. Becker, S. Yang, R. A. MacLachlan, and C. N. Riviere, "Towards vision-based control of a handheld micromanipulator for retinal cannulation in an eyeball phantom," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 44–49.

[16] S. Yang, J. N. Martel, J. Louis A Lobes, and C. N. Riviere, "Techniques for robot-aided intraocular surgery using monocular vision," *The International Journal of Robotics Research*, vol. 37, no. 8, pp. 931–952, 2018. [Online]. Available: https://doi.org/10.1177/0278364918778352

[17] M. Dewan, P. Marayong, A. Okamura, and G. Hager, "Vision-based assistance for ophthalmic micro-surgery," vol. 3217-II, 09 2004, pp. 49–57.

[18] R. Richa, M. Balicki, R. Sznitman, E. Meisner, R. Taylor, and G. Hager, "Vision-based proximity detection in retinal surgery," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2291–2301, 2012.

[19] J. W. Kim, C. He, M. Urias, P. Gehlbach, G. D. Hager, I. Iordachita, and M. Kobilarov, "Autonomously navigating a surgical tool inside the eye by learning from demonstration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 7351–7357.

[20] J. W. Kim, P. Zhang, P. Gehlbach, I. Iordachita, and M. Kobilarov, "Towards autonomous eye surgery by combining deep imitation learning with optimal control," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 2347–2358. [Online]. Available: https://proceedings.mlr.press/v155/kim21a.html

[21] P. Zhang, J. W. Kim, P. Gehlbach, I. Iordachita, and M. Kobilarov, "Autonomous needle navigation in retinal microsurgery: Evaluation in ex vivo porcine eyes," 2023. [Online]. Available: https://arxiv.org/abs/2301.11839

[22] Y. Koyama, M. M. Marinho, M. Mitsuishi, and K. Harada, "Autonomous coordinated control of the light guide for positioning in vitreoretinal surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 1, pp. 156–171, 2022.

[23] Y. Koyama, M. M. Marinho, and K. Harada, "Vitreoretinal surgical robotic system with autonomous orbital manipulation using vector-field inequalities," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2023, pp. 1–7.