

ColibriDoc: An Eye-in-Hand Autonomous Trocar Docking System

Shervin Dehghani^{1*}, Michael Sommersperger^{1*}, Junjie Yang², Benjamin Busam¹, Kai Huang³
Peter Gehlbach⁴, Iulian Iordachita⁵, Nassir Navab^{1,5} and M. Ali Nasser^{1,2}

Abstract—Retinal surgery is a complex medical procedure that requires exceptional expertise and dexterity. For this purpose, several robotic platforms are currently being developed to enable or improve the outcome of microsurgical tasks. Since the control of such robots is often designed for navigation inside the eye in proximity to the retina, successful trocar docking and inserting the instrument into the eye represents an additional cognitive effort, and is therefore one of the open challenges in robotic retinal surgery. For this purpose, we present a platform for autonomous trocar docking that combines computer vision and a robotic setup. Inspired by the Cuban Colibri (hummingbird) aligning its beak to a flower using only vision, we mount a camera onto the endeffector of a robotic system. By estimating the position and pose of the trocar, the robot is able to autonomously align and navigate the instrument towards the Trocar’s Entry Point (TEP) and finally perform the insertion. Our experiments show that the proposed method is able to accurately estimate the position and pose of the trocar and achieve repeatable autonomous docking. The aim of this work is to reduce the complexity of robotic setup preparation prior to the surgical task and therefore, increase the intuitiveness of the system integration into the clinical workflow.

Index Terms—Medical Robots and Systems; Surgical Robotics; Planning; Computer Vision for Medical Robotics.

I. INTRODUCTION

Vitreoretinal surgery is known to be one of the most challenging and delicate surgical procedures, requiring surgeons to have sufficient expertise and exceptional hand stability to manipulate microsurgical instruments. The demand for specialized retinal surgeons is high, as more than 300 million [1] patients are affected by visual disorders caused by various retinal diseases. These diseases are mainly treated through vitreoretinal interventions aiming to preserve or restore vision. Since vitreoretinal surgery is a minimally invasive procedure, trocars are placed as insertion ports on the sclera (sclerotomy) prior to surgery allowing access to the operating area. The surgeon then docks an infusion line, an illuminator and a surgical instrument into these ports to start the vitreoretinal intervention (see Fig. 1).

* The first two authors contributed equally to this paper. Corresponding author: M. Ali Nasser (ali.nasser@mri.tum.de)

¹ S. Dehghani, M. Sommersperger, B. Busam, N. Navab and M. Ali Nasser are with Department of Computer Science in Technische Universität München, München 85748 Germany.

² J. Yang and M. Ali Nasser are with Augenklinik und Poliklinik, Klinikum rechts der Isar derc Technische Universität München, München 81675 Germany.

³ K. Huang is with Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Guangzhou, China.

⁴ P. Gehlbach is with Wilmer Eye Institute, Johns Hopkins Hospital, Baltimore, MD, USA.

⁵ I. Iordachita and N. Navab are with Laboratory for Computational Sensing and Robotics, Johns Hopkins University, Baltimore, MD, USA.

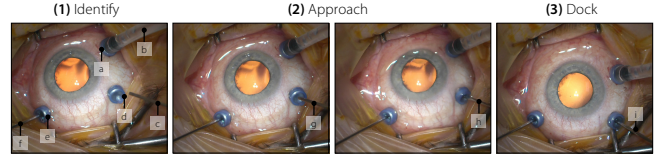


Fig. 1. In conventional vitreoretinal interventions, trocars are inserted into the sclera to dock various surgical instruments and allow access to the surgical site: a. Trocar I b. Infusion line c. Instrument approaching the trocar d. Trocar II e. Trocar III f. Illuminator g. Instrument docking the trocar h. Instrument aligning towards the trocar orientation i. Properly aligned instrument inserted into the trocar.

In conventional vitreoretinal interventions, surgeons rely on both their visual and haptic feedback to introduce instruments through the trocars. They approach by visually identifying the trocar and fine-tune the insertion alignment by sensing the forces during the docking procedure.

In recent years, robotic platforms have made promising strides towards facilitating and improving vitreoretinal procedures and could enable more surgeons to perform such complicated tasks [2]–[4]. Despite their significant technical improvements, robotic systems are still limited in interfacing and integration capabilities into the surgical workflow. For instance, in current robotic setups clinicians need to invest significant time and energy for the preparation of the system and the accurate manual positioning of the robot in order to adjust the appropriate docking orientation prior to the main procedure. Therefore, in minimally invasive robotic surgery, and specifically in vitreoretinal robotic surgery, the task of robot positioning, trocar docking, and instrument insertion poses additional cognitive demands on the surgeon. The main reason for this effort is that the control of such systems is designed for delicate procedures at micron-level scale inside the eye, with limited working volume, and is often manipulated using a controller such as a joystick [5]. Therefore, inserting the microsurgical instrument into the trocar using this control system naturally becomes more challenging and time-consuming compared to the conventional insertion in manual surgery. Automating the task of navigating a surgical tool from a safe distance towards the trocar and introducing it into the eye could thus relieve additional cognitive load on the surgeon and consequently reduce the complexity of employing a robotic system in the clinical workflow.

In this paper, we propose *ColibriDoc*, a system for autonomous trocar docking and instrument insertion. Our setup consists of an RGB camera, mounted on the endeffector of an ophthalmic robot. Its design is mainly inspired by the natural behavior of Cuban Colibris (hummingbirds), which hover at flowers and rely only on their stereo vision to align

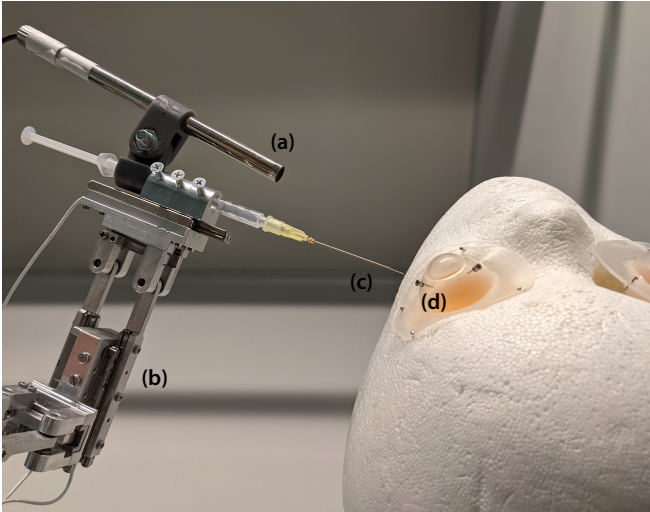


Fig. 2. Our setup for autonomous trocar docking consists of a monovision RGB camera (a) mounted on a microsurgical robot (b). The surgical instrument (c) is attached to the robot end effector and autonomously navigated to perform the docking procedure to the target trocar (d).

and insert their beak [6], [7]. Due to the space constraints in the operating room and the limited mounting options on robotic systems, contrary to the hummingbird we follow a mono-vision approach based on a single camera.

Our proposed method precisely identifies the homogeneous location of the Trocar Entry Point (TEP) and its orientation from RGB images using a two-stage learning-based approach. The robot then autonomously aligns the surgical tool with the trocar orientation and ultimately performs trocar docking and instrument insertion. For a proof of concept and experimental evaluation, we implement the proposed method on a robotic system designed for vitreoretinal surgery, which is described in [5]. Our experiments show that the TEP and trocar pose can accurately be determined using RGB images only. Most importantly, we show that such a system is capable of performing trocar docking and instrument insertion in a precise and repeatable manner, automating a task that in robotic surgery poses additional complexity on the surgeon. We demonstrate the concept for vitreoretinal surgery, however, this methodology could also be transferred to other types of minimally invasive robotic surgery. To the best of our knowledge, the system proposed in this paper is the first work towards automating the trocar docking procedure.

II. RELATED WORKS

Robotic Systems: In the last decade, many robotic systems have been designed and applied in surgeries as robotic assistants to promote autonomy and high accuracy such as [8]–[10]. These approaches are classified into the following types according to the interaction between the surgeon and the robotic system: 1) hand-held surgery instruments [11] fully controlled by the surgeon, 2) cooperatively controlled robotic systems [12] in which the surgical instrument is jointly controlled by the robot and surgeon and 3) teleoperation systems, in which the robot is remotely controlled by

the surgeon via a guidance device such as a joystick [13], [14]. In this work we employ a 5 DoF hybrid parallel-serial robot [13], [15] specifically designed for delicate vitreoretinal procedures, which can be controlled as a teleoperation system or by a software framework.

Trocar Entry Point Detection and Pose Estimation: Till date, only few works have been published with a motivation similar to the one for autonomous trocar docking proposed in this paper. Multiple works have focused on estimating the TEP and positioning the Remote Center of Motion (RCM) during surgery using a geometric approach [16]–[18] or an external stereo-vision system [19]. Birch et al. [20] recently published an initial paper on the development of an instrument with two integrated miniature cameras to detect the trocar position and an internal measurement unit to estimate the RCM point of the robot. Rather than performing the entire trocar docking procedure at an initial safe distance from the target, such approaches focus on repositioning the instrument to align the RCM with the TEP after instrument insertion. In contrary, we detect the TEP and its pose at a safe initial distance from the eye and autonomously navigate the robot to perform docking and instrument insertion.

In recent years, 6D Pose Estimation of objects has become a popular research topic. While depth-based methods show higher accuracy than monocular-based methods, the recent progress in the field of monocular pose estimation [21]–[23], demonstrates the high capabilities of using monocular camera instead of stereo or depth cameras. These approaches can be categorized into two main groups: indirect and direct methods. The indirect methods goal is to find $2D - 3D$ point correspondences to directly derive the pose from a PnP algorithm [24]. Direct methods, on the other hand, approach the problem as a regression [22], [23] or classification [21] task using a direct differentiable method. To estimate the pose of a microsurgical trocar as part of our proposed system, we use a mono-camera based direct approach, in order to better cope with the texture-less and symmetric properties of trocar and the limitations in the workspace of the robot, as described in more detail in section IV-B.

III. DATASET

A major challenge in the learning-based detection of the trocar and the estimation of its pose is the acquisition of a dataset with accurate ground truth information containing both the location and orientation of the TEP. Due to the lack of available datasets, we generate a purely synthetic dataset from a virtual setup, which contains virtual models of an eye, a trocar, a surgical needle, as well as a virtual camera. The camera and the tool are positioned relative to the eye and the trocar similar to the real setup. In this virtual environment the exact location and orientation of the trocar in relation to the camera is known. The parameters of the virtual camera are adjusted according to the parameters of the calibrated camera that is mounted on the real robotic system and the resolution of the rendered frames is adapted to match the image resolution of the robot-mounted camera. To generate the synthetic dataset, we randomize the position and orientation

of the trocar in the rendered frames, along with the location of the trocar on the eye model. Furthermore, we vary the metallic properties and glossiness of the trocar material, as well as the direction and intensity of the lighting in the scene. In 20% of the acquired frames the eye model is not rendered to avoid bias towards specific eye characteristics, but rather focus on learning the features of the trocar. In an effort to improve robustness towards changing peripheral areas when training deep learning models to estimate the TEP and pose, randomized images from the cocodataset [25] are rendered in the background of the virtual scene. The generated synthetic dataset $DS_{synthetic}$ contains 2000 images with ground truth information of the TEP in image coordinates and the 3D pose of the trocar in the virtual scene relative to the camera. For fine-tuning and bridging the gap between the synthetic and real data domain, we acquire two different labeled datasets, consisting of images extracted from videos captured with our proposed system. One of them, $DS_{trainTEP}$, is used for the TEP detection, while the other one, $DS_{trainPoseReal}$, is used for orientation estimation. To acquire these datasets, a microsurgical trocar (23G from RETILOCK) was inserted into a phantom eye and the robotic setup was positioned in a realistic distance from the trocar. Two different phantom eyes (VR Eye from Philips Studio and BIONIKO eye model with flex orbit holder) were used to replicate real surgical scenarios. Fig. 3 shows an example of a synthetic and a real image. For $DS_{trainTEP}$, six videos were acquired, from which 1280 frames were extracted to create the dataset. The ground truth location of the TEP was manually annotated by a biomedical engineering expert. $DS_{trainPoseReal}$ consists of 100 images, in which we obtain the ground truth orientation of the trocar in relation to the camera using a marker aligned with the trocar. With the same approaches, 160 images were extracted creating the dataset $DS_{testTEP}$ and 50 images creating $DS_{testPoseReal}$, which are used for the evaluations presented in section V.

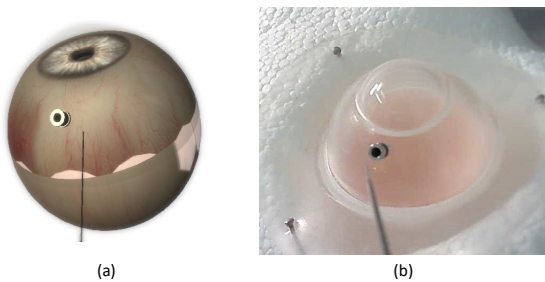


Fig. 3. A synthetic (a) and a real (b) image with similar trocar poses.

IV. METHOD

In this section, the proposed method for autonomous trocar docking is described in detail. Firstly, the setup consisting of an ophthalmic robot and a mono-vision camera mounted on the robot endeffector is illustrated. Thereafter, the components to detect the position and orientation of the trocar, as well as the post-processing steps to refine the estimations are presented. Finally, we outline the trajectory planning to autonomously align the instrument with the trocar orientation

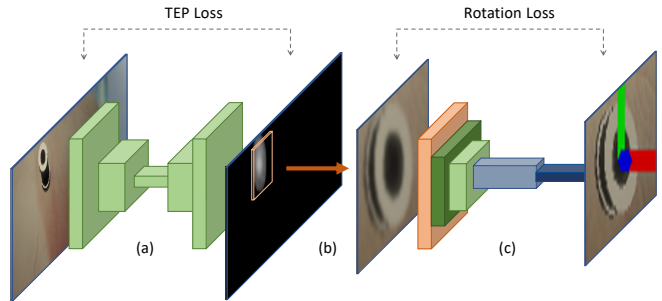


Fig. 4. The proposed pipeline for the 4DoF trocar pose estimation. **a.** A U-Net like network to segment the TEP, **b.** Cropping the image around TEP, **c.** A Resnet34 to regress the pose.

and navigate the tooltip towards the TEP to perform the docking procedure.

A. Setup

In this work, a 5DoF robotic micromanipulator [15], [26] is employed. The robot consists of two parallel coupled joint mechanisms for translation and rotation in two axes and a decoupled prismatic joint for Z movement of the endeffector. All joints are actuated by micron-precision piezo motors with integrated sub-micron optical encoders. To achieve autonomous trocar docking, a Teslong Multi-Function Soldering Magnifier Camera is calibrated and rigidly mounted on the endeffector of the robot. A sub-retinal cannula (23G with 40G tip) is attached to the robot endeffector and a 3D printed holder was specifically designed to mount the camera to the syringe in a suitable orientation. The proposed robotic setup is illustrated in Fig. 2. For this Eye-in-Hand [27] setup, in which the endeffector moves the instrument along the camera, a Hand-Eye calibration [28] is performed to acquire the camera pose w.r.t. to the robot.

To create realistic docking targets, conventional 23G trocars are placed 3.5 mm posterior to the limbus of two surgical training phantom eyes (described in section III) with proper scleral textures and deformability.

B. 4DoF Trocar Pose Estimation

In a typical vitreoretinal surgery, several trocars are placed to provide the ports for instruments and access to the surgical site. The first step towards autonomous robotic trocar docking is identifying the appropriate entry point and pose of the target trocar based on image frames captured by the camera mounted on the robot. Our goal is to estimate the projected 2D point of the closest trocar's TEP in the current image plane and the normal vector of its cross-section. Due to the nature of vitreoretinal surgery and the small working area of the surgical robot, the robot's endeffector is initially positioned in a reachable distance to the trocar. Thus, the closest trocar is the docking target, which is also directly visible in the camera's field of view. We employ a two stage 4DoF pose estimation neural network, to address both the TEP and trocar pose estimation from single RGB images. We initially locate the TEP by the first network, which extracts

the ROI that is subsequently forwarded to the second network, responsible for estimating the orientation. Obtaining the orientation along the TEP provides the homogeneous location of the trocar, which is the basis of our hummingbird-inspired docking approach, as explained further in section IV-C.

Trocar Entry Point Detection. Detecting the TEP is, firstly, relevant for extraction of a ROI to be later used for pose estimation, and controlling the robot to align the tooltip with the trocar orientation and to perform the trocar docking procedure. In addition, a small region of interest around the TEP is subsequently extracted from the input frame and used to estimate the trocar orientation. Rather than using an object detection network to estimate a bounding box around the trocar, it is instead important to estimate the exact location of the TEP, since, depending on the orientation of the trocar, the entry point cannot easily be obtained from a bounding box. For this stage, a U-Net-style network [29] using a Resnet34 [30] as a backbone feature extractor is trained to obtain the location of the TEP in the camera images. To generate the ground truth images for training the network, a Gaussian function is applied around the ground truth TEP with a maximum distance of 15 pixels from its center and a sigma of 1.

The network is first pretrained on $DS_{synthetic}$ and afterwards fine-tuned on $DS_{trainTEP}$. The last network layer uses the sigmoid activation function, which is followed by the binary cross entropy loss.

To derive the final image coordinates of the TEP location for each processed frame, we consider all pixel locations (x, y) in the network output, which satisfy the equation

$$pred(x, y) \geq 0.8 * max(pred) \quad (1)$$

where $pred(x, y)$ defines the confidence that the output image at the pixel location (x, y) is classified as TEP and $max(pred)$ is the overall maximum value in the network's output image. The TEP is then estimated as the median of the extracted candidate locations.

During inference, we apply further post-processing which combines the predicted locations of every seven consecutive frames to improve the robustness of the prediction. First, the median value of all seven predictions is calculated and the Euclidean distance between each individual prediction and the overall median value is determined. All predictions more than one quarter of a standard deviation distant from the overall median location are considered outliers. The remaining estimates are then averaged to produce the final TEP.

Orientation Estimation. While dealing with a texture-less and symmetric object as a trocar, we leverage a direct pose estimation method with symmetric loss to predict the normal vector of trocars' cross-section.

In order to have a continuous space of rotations in $\mathbf{SO}(3)$, we follow the method introduced in [31] to parameterize rotation angles. As demonstrated in [22], [23], the mapping function f to the 6-dimensional representation \mathbf{R}_{6d} is defined

as the first two columns of \mathbf{R}

$$f \left(\begin{bmatrix} | & | & | \\ R_1 & R_2 & R_3 \\ | & | & | \end{bmatrix} \right) = \begin{bmatrix} | & | \\ R_1 & R_2 \\ | & | \end{bmatrix} \quad (2)$$

Given a 6-dimensional vector $\mathbf{R}_{6d} = [\mathbf{r}_1 | \mathbf{r}_2]$, the unit and orthogonal rotation matrix \mathbf{R} is computed as

$$\begin{cases} \mathbf{R}_1 = \phi(\mathbf{r}_1) \\ \mathbf{R}_3 = \phi(\mathbf{R}_1 \times \mathbf{r}_2) \\ \mathbf{R}_2 = \mathbf{R}_3 \times \mathbf{R}_1 \end{cases}, \quad (3)$$

where $\phi(\bullet)$ denotes the vector normalization operation.

For the convenience of our problem, we use a permutation of columns of \mathbf{R} which changes the 6D representation to

$$\mathbf{R}_{6d} = [\mathbf{R}_Z | \mathbf{R}_Y] \quad (4)$$

Thereafter, we use a Resnet34-based [30] backbone to convert the input image, derived from a trocar-centered ROI extraction into its features, followed by fully connected layers to regress to the 6D representation of the rotation. Assuming the coordinate system of a trocar as illustrated in Fig. 5, it can be seen that a trocar is symmetric along its Z axis. In this regard, we design a loss which does not penalize the network for non-relevant rotations around the trocar's z axis. The angle between the ground truth \mathbf{R}_Z and estimation can be computed as:

$$\Delta_\theta = \arccos(\mathbf{R}_Z^{gt} \cdot \mathbf{R}_Z^{pred}) \quad (5)$$

For ℓ^2 -normalized vectors MSE is proportional to *cosine* distance, and due to this fact we use Eq. 6 as the loss function, which shows a more promising convergence.

$$\mathcal{L}_{rotation} = \text{MSE}(\mathbf{R}_Z^{gt}, \mathbf{R}_Z^{pred}) \quad (6)$$

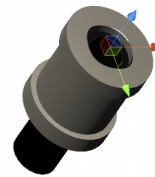


Fig. 5. Illustration of the trocar coordinate system. The trocar orientation can be defined only by the rotation around its x and y axis, which are visualized in red and green, respectively.

In order to be robust against noise, we average seven consecutive estimations using the method introduced in [32].

C. Trajectory Planning

Due to the non-unique Z value obtained from the TEP detection, it is not possible to apply a direct docking process. This creates the need for designing a method to dynamically find the appropriate trajectory for the robot. In this work we have designed a procedure, inspired by hummingbird's docking approach [6]. We first align the tooltip and trocar orientation, followed by the alignment of the XY translation of the tooltip with TEP. As drawn in Fig. 6, keeping the tooltip on the line which connects the trocar's entry location

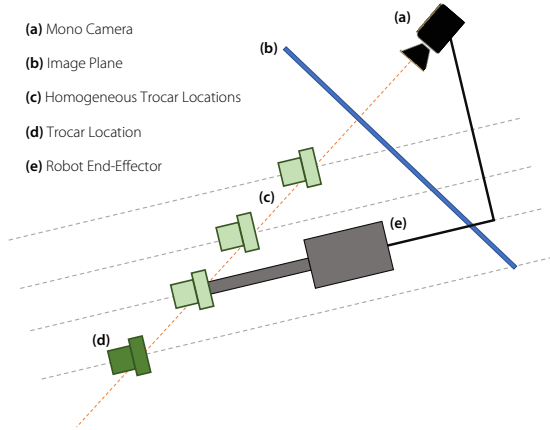


Fig. 6. Trajectory of the tooltip, compared to the camera and trocar’s homogeneous location, illustrated in a 2D projection.

to the camera, we start approaching the trocar with an adaptive speed. With this approach we ensure the tooltip is always aligned with the trocar, and the tooltip’s projected position is aligned with the TEP’s projected position in each image plane. By iteratively following this approach, we can also compensate for small movements of trocar.

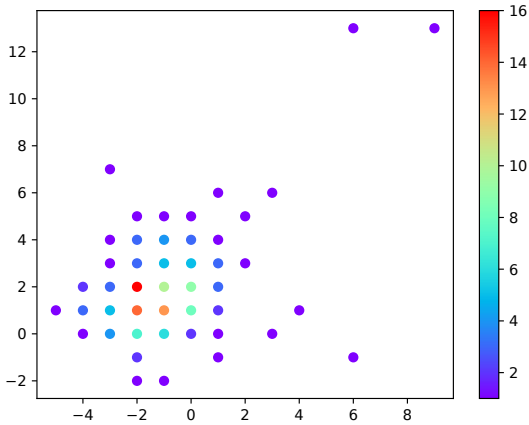


Fig. 7. The offset between the ground truth TEP and the detected TEP in x and y direction in pixel. The color scheme indicates the number of times each offset was attained. To aid illustration, two outliers with euclidean distances to the ground truth location of 320 and 216 pixel were omitted.

V. EXPERIMENTS AND RESULTS

To validate our proposed system, we first separately evaluate its sub-components and finally the overall autonomous trocar docking performance. In the following, we demonstrate the accuracy of the TEP detection and trocar orientation estimation. We finally recreate a surgical scenario using phantom eyes and demonstrate the validity of our system by reporting the success rate of our method. We additionally show that the autonomous docking procedure does not increase the operation time compared to the manual joystick-based approach. Besides the quantitative results presented in the following sections, we provide qualitative visual results of the TEP and orientation estimation as well as the

autonomous docking procedure as supplementary materials along with this paper.

A. Trocar Entry Point Detection

To evaluate the detection of the TEP we use our test set $DS_{testTEP}$ consisting of 160 images, which were extracted from a video captured with our proposed setup. The 2D image coordinates of the ground truth TEPs were manually annotated in the image frames. The detected location of the TEP is then compared to its respective ground truth. The overall achieved median, mean and standard deviation of the error in x and y image coordinates are 6.19, 2.82, and 30,29 pixel, respectively. Given an image resolution of 1280×720 pixels using our robot mounted camera, the offset error results in a mean of 0.74%, a median of 0.28%, and a standard deviation of 3.75% from the annotated entry point. Fig. 7 illustrates the detection error along with the number of occurrences of each offset evaluated in our test set. The corresponding euclidean distances in pixels are visualized in Fig. 8. The comparably high standard deviation is caused by two outliers, of which one showed a euclidean pixel distance of 320 and 216 pixels, respectively.

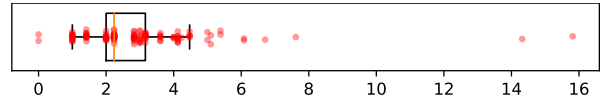


Fig. 8. The euclidean distances between the estimated TEPs and the respective ground truth locations in pixel. Similar to Fig. 7, two outliers were omitted for improved visualization.

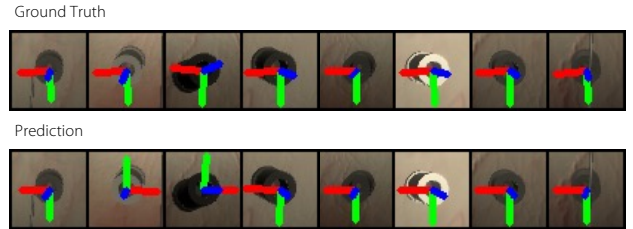


Fig. 9. Comparison between the ground truths and the pose regression network on the test set of the synthetic dataset.

B. Trocar Orientation Estimation

After training the pose estimation model on the $DS_{synthetic}$ and fine-tuning on $DS_{realPoseTrain}$, we evaluated the results on the real images of our test set $DS_{realPoseTest}$. As illustrated in the Fig. V-B, the model has achieved an accuracy of 80% to estimate the trocar orientation below 10° error, and 94% below 15° , which is achieved by fine tuning on a small set of real data. It is worth mentioning that the generation of a real dataset with ground truth orientation of the microsurgical trocar is challenging and prone to small errors due to the marker-based estimation. Therefore, this quantitative evaluation is limited to the quality of the generated ground truth orientation.

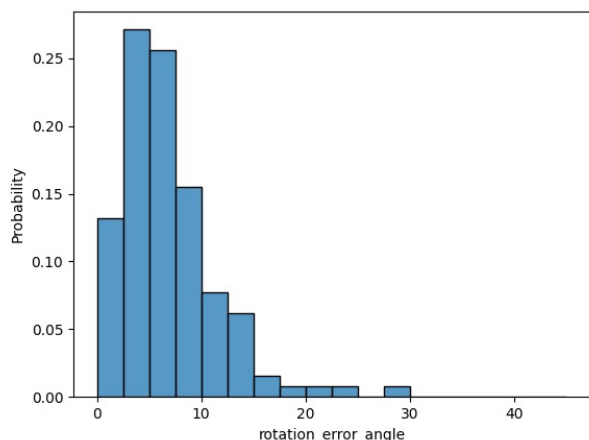


Fig. 10. Pose estimation error on test set of the real dataset

C. Autonomous Trocar Docking

To validate our entire system for autonomous trocar docking, we simulate a surgical environment consisting of the robotic setup and a VR Eye from Philips Studio, which is attached to a head phantom. For each trial we initially position the robot at a randomized location near the target trocar, ensuring that the docking target is located within the robot workspace and that the TEP can be reached during the docking process. From the initial position, the robot then aligns the instrument with the trocar orientation and moves the tip towards the TEP as described in Section IV-C. Our experiments showed that out of 11 attempts, the robot was able to successfully reach the TEP and perform docking in 10 cases. In one case, the instrument touched the edge of the trocar and missed the entry point. However, this was due to a hand-eye calibration error between the instrument tip and the camera caused by a deformation of the tooltip. The average time to completion 10 successful autonomous trials was 35.1 seconds with a standard deviation of 3.2 seconds. For comparison, two biomedical engineers trained to control the robot with a joystick performed the docking task with the same setup. The manual alignment took on average 40.8 seconds but showed a higher variance in the time to completion with a standard deviation of 12.75 seconds. The maximum time required for autonomous docking was 41 seconds compared to 57 seconds for manual alignment

VI. DISCUSSION AND FUTURE WORK

In our current experiments, one constraint of our method is the working space of the microsurgical robot. To perform successful docking, the robot has to be positioned in proximity to the trocar, so that the instrument can be aligned with the trocar and the entry point can be reached. However, the challenging act of aligning and inserting the instrument is still performed autonomously by the robot. When using a robot with a larger working space, autonomous docking could also be performed from a greater starting distance.

In this work, we used a sub-retinal cannula (23G with 40G tip) with a straight tip, therefore, the introduction of the cannula into the trocar was performed by a final "Z"

movement, following the proper docking alignment. For bent-tip cannulas, this method can be extended for more complex introduction trajectories. Here, if the trocar is out of detection range (due to movement of the eye during the procedure, or coverage by an external objects or blood), the docking procedure is stopped until the trocar is back to the detection range.

As our experiments have shown, the current setup is able to perform autonomous trocar docking and insert the instrument into the eye. Once the instrument has reached the desired insertion depth, the process is manually stopped. In future works, we will investigate suitable stopping criteria to autonomously insert specific tools to the desired depth within the eye (e.g. when the tip appears in the microscopic view). Additionally, our approach can be extended to dynamically adjust for changing trocar position during the insertion process. We will also further analyze the force applied during autonomous robotic trocar docking and compare it to the force applied during conventional manual insertion, as autonomous trocar docking could lower the force applied on the sclera, thereby reducing patient trauma.

VII. CONCLUSION

In this paper we proposed an autonomous docking system inspired by the natural behaviour of hummingbirds for minimally invasive robotic surgery. Goller, Altshuler et al. in [6] and [7], presented "Feeding hummingbirds use vision, not touch, to hover at flowers". They showed hummingbirds are identifying their target entry point and after approaching they fix their head and proceed in "Z" direction to insert their beak into a flower. Learning from this natural phenomena, our system for autonomous trocar docking and instrument insertion is based on a camera mounted to the endeffector of a micron-precision robot. The approach first obtains the position of the TEP and subsequently estimates the trocar's orientation from RGB images. The robot then aligns instrument with the trocar orientation and autonomously performs the docking procedure. Our experiments have shown that the TEP can be detected with a clinical grade accuracy with a mean euclidean error of 3 pixel given an image resolution of 1280×720 pixel. The detection of this point is extremely important for the subsequent pose estimation, since the input for the pose estimation network comprises a small region around the estimated TEP. Further experiments were performed to validate our system by evaluating the entire autonomous docking procedure using a conventional 23G trocar mounted on a surgical training phantom eye: Successful trocar docking could be achieved with high repeatability and an average time of 35 seconds, which indicates the potential of our method. We consider this work as a proof that the proposed method for autonomous robotic trocar docking is a valid approach towards automatizing a task, which in robotic surgery would otherwise require additional cognitive effort from the surgeon.

REFERENCES

- [1] I. F. de Oliveira, E. J. Barbosa, M. C. C. Peters, M. A. B. Henostroza, M. N. Yukuyama, E. dos Santos Neto, R. Löbenberg, and N. Bou-Chacra, "Cutting-edge advances in therapy for the posterior segment of the eye: Solid lipid nanoparticles and nanostructured lipid carriers," *International Journal of Pharmaceutics*, p. 119831, 2020.
- [2] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel *et al.*, "Medical robotics-regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.*, vol. 2, no. 4, p. 8638, 2017.
- [3] B. Lu, H. K. Chu, K. Huang, and L. Cheng, "Vision-based surgical suture looping through trajectory planning for wound suturing," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 542–556, 2018.
- [4] E. Vander Poorten, C. N. Riviere, J. J. Abbott, C. Bergeles, M. A. Nasser, J. U. Kang, R. Sznitman, K. Faridpooya, and I. Iordachita, "Robotic retinal surgery," in *Handbook of Robotic and Image-Guided Surgery*. Elsevier, 2020, pp. 627–672.
- [5] M. Zhou, Q. Yu, K. Huang, S. Mahov, A. Eslami, M. Maier, C. P. Lohmann, N. Navab, D. Zapp, A. Knoll, and M. Nasser, "Towards robotic-assisted subretinal injection: A hybrid parallel–serial robot system design and preliminary evaluation," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6617–6628, 2020.
- [6] B. Goller and D. L. Altshuler, "Hummingbirds control hovering flight by stabilizing visual motion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 51, pp. 18 375–18 380, 2014.
- [7] B. Goller, P. S. Segre, K. M. Middleton, M. H. Dickinson, and D. L. Altshuler, "Visual sensory signals dominate tactile cues during docked feeding in hummingbirds," *Frontiers in neuroscience*, vol. 11, p. 622, 2017.
- [8] F. Ullrich, C. Bergeles, J. Pokki, O. Ergeneman, S. Erni, G. Chatzipirpiridis, S. Pané, C. Framme, and B. J. Nelson, "Mobility experiments with microrobots for minimally invasive intraocular SurgeryMicro-robot experiments for intraocular surgery," *Invest. Ophthalmol. Vis. Sci.*, vol. 54, no. 4, pp. 2853–2863, 2013.
- [9] E. Rahimy, J. Wilson, T. C. Tsao, S. Schwartz, and J. P. Hubschman, "Robot-assisted intraocular surgery: development of the IRISS and feasibility studies in an animal model," *Eye*, vol. 27, no. 8, pp. 972–978, 2013.
- [10] A. Gijbels, E. Vander Poorten, B. Gorissen, A. Devreker, P. Stalmans, and D. Reynaerts, "Experimental validation of a robotic comanipulation and telemanipulation system for retinal surgery," in *2014 5th IEEE RAS EMBS Int. Conf. Biomed. Robot. Biomechatronics*. IEEE, 2014, pp. 144–150.
- [11] C. Song, P. L. Gehlbach, and J. U. Kang, "Active tremor cancellation by a "smart" handheld vitreoretinal microsurgical tool using swept source optical coherence tomography," *Opt. Express*, vol. 20, no. 21, pp. 23 414–23 421, 2012.
- [12] R. Taylor, P. Jensen, L. Whitcomb, A. Barnes, R. Kumar, D. Stoianovici, P. Gupta, Z. Wang, E. deJuan, and L. Kavoussi, "A steady-hand robotic system for microsurgical augmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI'99*, C. Taylor and A. Colchester, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 1031–1041.
- [13] M. Zhou, Q. Yu, K. Huang, S. Mahov, A. Eslami, M. Maier, N. Navab, D. Zapp, A. Knoll *et al.*, "Towards robotic-assisted subretinal injection: A hybrid parallel–serial robot system design and preliminary evaluation," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6617–6628, 2019.
- [14] A. Molaei, E. Abedloo, M. D. de Smet, S. Safi, M. Khorshidifar, H. Ahmadi, M. A. Khosravi, and N. Daftarian, "Toward the art of robotic-assisted vitreoretinal surgery," *Journal of ophthalmic & vision research*, vol. 12, no. 2, p. 212, 2017.
- [15] M. A. Nasser, M. Eder, S. Nair, E. C. Dean, M. Maier, D. Zapp, C. P. Lohmann, and A. Knoll, "The introduction of a new robot for assistance in ophthalmic surgery," in *Eng. Med. Biol. Soc. (EMBC), 2013 35th Annu. Int. Conf. IEEE*. IEEE, 2013, pp. 5682–5685.
- [16] J. Smits, D. Reynaerts, and E. V. Poorten, "Setup and method for remote center of motion positioning guidance during robot-assisted surgery," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1315–1322.
- [17] C. Gruijthuijsen, L. Dong, G. Morel, and E. V. Poorten, "Leveraging the fulcrum point in robotic minimally invasive surgery," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2071–2078, 2018.
- [18] L. Dong and G. Morel, "Robust trocar detection and localization during robot-assisted endoscopic surgery," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4109–4114.
- [19] B. Rosa, C. Gruijthuijsen, B. Van Cleynenbreugel, J. V. Sloten, D. Reynaerts, and E. V. Poorten, "Estimation of optimal pivot point for remote center of motion alignment in surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 2, pp. 205–215, 2015. [Online]. Available: <https://doi.org/10.1007/s11548-014-1071-3>
- [20] J. Birch, King's College London, K. Rhode, King's College London, C. Bergeles, King's College London, L. Da Cruz, and Moorfields Eye Hospital, "Towards localisation of remote centre of motion and trocar in vitreoretinal surgery," pp. 33–34. [Online]. Available: <https://www.ukras.org/publications/ras-proceedings/UKRAS21/pp33-34>
- [21] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1521–1529.
- [22] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16611–16621.
- [23] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–591.
- [24] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 292–301.
- [25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [26] M. Zhou, M. Hamad, J. Weiss, A. Eslami, K. Huang, M. Maier, N. Navab, A. Knoll, and M. Nasser, "Towards robotic eye surgery: Marker-free, online hand-eye calibration using optical coherence tomography images," *IEEE Robot. Autom. Lett.*, 2018.
- [27] 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)*, vol. 3, 2000, pp. 2741–2746 vol.3.
- [28] R. Horaud and F. Dornaika, "Hand-eye calibration," *The international journal of robotics research*, vol. 14, no. 3, pp. 195–210, 1995.
- [29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition. corr abs/1512.03385 (2015)," 2015.
- [31] Y. Zhou, C. Barnes, L. Jingwan, Y. Jimei, and L. Hao, "On the continuity of rotation representations in neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [32] F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman, "Averaging quaternions," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 4, pp. 1193–1197, 2007.