2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Hybrid Visual Servoing for Robotic Assistance in ENT Microsurgery: A Case Study on Middle Ear Access*

Mohamed-Aimen Boulala[1], Carlos Mateo-Agulló[1,†], Renato Martins[1], Alain Lalande[2], Cédric Demonceaux[1], Alexis Bozorg Grayeli[2,3]

***Abstract*— This paper introduces a robotic assistance system for minimally invasive middle ear surgery, focusing on precise access to the tympanic membrane. The system combines a 7-degree-of-freedom robotic arm with a hybrid visual servoing framework that integrates position-based and image-based control strategies. Dual visual feedback from a color camera and an endoscope enables robust 6-DoF pose estimation and sub-millimetric tool guidance. A model-based tracker and blob detection ensure accurate alignment and targeting, while a Quadratic Programming controller enforces safety constraints such as maintaining the field of view. The approach is validated through simulations and real-world experiments, demonstrating high accuracy, robustness to anatomical variability, and suitability for clinical integration. This work advances robotic microsurgery by providing a closed-loop, constraint-aware control architecture tailored for otologic procedures.**

I. INTRODUCTION

Minimally invasive surgery (MIS) has revolutionized modern medical practice by reducing patient trauma, postoperative pain, and recovery time. In the field of otolaryngology (ENT: Ear, Nose, and Throat), particularly in middle ear procedures, the demand for precision is exceptionally high due to the small anatomical structures and limited access pathways. Traditional manual techniques often suffer from limitations such as hand tremor, restricted visibility, and ergonomic fatigue, which can compromise surgical outcomes.

Robotic systems have emerged as promising tools to enhance precision and stability in microsurgical environments. In particular, the concept of Remote Center of Motion (RCM) has enabled the development of robotic manipulators capable of operating through narrow anatomical corridors without damaging surrounding tissues. However, most existing systems rely on open-loop control or pre-programmed trajectories, lacking the adaptability required for dynamic surgical environments.

To address these challenges, this work proposes a hybrid visual servoing framework for robotic assistance in middle ear surgery. The system integrates a 7-degree-of-freedom (7-DoF) robotic arm with dual-camera visual feedback — combining a color camera and an endoscope— to guide the surgical tool with sub-millimetric accuracy (Fig. 1). By leveraging a hybrid visual servoing scheme, the robot can adapt its motion in real time based on visual cues, ensuring safe and precise navigation toward the tympanic membrane.

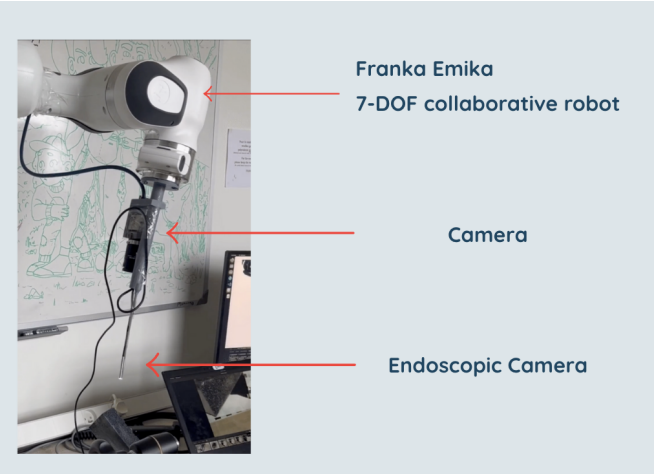


Fig. 1. System overview. General architecture of the robotic system for middle ear surgery, featuring a 7-DoF robotic arm and a dual-camera setup for hybrid visual servoing.

This paper presents a novel framework designed around the surgical workflow: approaching the ear, aligning with the tympanic membrane, and targeting a sub-millimetric penetration point. Each stage leverages specific visual and control strategies to ensure safety and precision. The system is validated in both simulation and real-world scenarios, demonstrating its potential for clinical application. Key contributions include: (1) A dual-camera visual servoing architecture specifically designed for ENT surgery; (2) The integration of a dual-tracking system—combining 3D object pose and image-based tracking—for robust 6-DoF pose estimation while maintaining precise 2D targeting; (3) A Quadratic Programming control framework that enforces safety constraints, such as preserving the field of view.

II. RELATED WORK

Minimally invasive surgery (MIS) has become a cornerstone of modern surgical practice due to its ability to reduce patient trauma, postoperative pain, and recovery time. In the field of otolaryngology (ENT), MIS techniques have been explored to access the middle ear through narrow and delicate anatomical pathways. Among these, endoscopic transtympanic approaches (ETTA) have enabled direct access

*This work was partially supported by grants ANER MOVIS from Conseil Régional BFC and ANR MANYVIS (ANR-23-CE23-0003-01), to whom we are grateful.

[1]Université Bourgogne Europe, ICB UMR CNRS 6303, Dijon, France
[2]Université Bourgogne Europe, ICMUB UMR CNRS 6302, Dijon, France
[3]Service ORL et de chirurgie cervico-faciale, CHU de Dijon, France
[†]carlos-manuel.mateo-agullo@ube.fr

to the ossicular chain for procedures such as ossiculoplasty using bone cement [1]. These interventions demand high precision due to the limited operating space and the fragility of surrounding structures.

A critical constraint in such procedures is the need for the surgical instrument to pivot at the tympanic membrane to avoid tissue damage. In robotics, this is formalized as the *Remote Center of Motion* (RCM) constraint. RCM is essential in robotic MIS, as it ensures that instruments move around a virtual pivot point located outside the robot's physical structure [2]. Two main strategies exist for implementing RCM: (1) design-based mechanisms that passively enforce the constraint through mechanical architecture, and (2) control-based approaches that actively regulate motion via software algorithms. Both are evaluated using metrics such as workspace, dexterity, accuracy, and stiffness—key factors when human safety and interaction forces are involved.

Several robotic systems have been developed to enhance precision in middle ear surgery. The RobOtol system [4], for example, was specifically designed to operate within the confined anatomy of the external auditory canal. It enables high-precision tool manipulation while preserving an unobstructed visual field, which is critical for microsurgical procedures such as stapedotomy. In [3], a clinical study demonstrated the feasibility of robotic assistance in cochlear implantation, showing improvements in surgical accuracy, ergonomics, and reduction of hand tremor and fatigue.

Despite these advances, many existing systems lack full automation and real-time feedback integration. This limitation motivates the use of visual servoing techniques, which enable closed-loop control based on visual information, allowing the robot to adapt to dynamic anatomical changes.

Visual servoing is a well-established technique in robotics that uses visual data from one or more cameras to control robot motion. It has been widely applied in industrial settings for tasks requiring high precision, such as part alignment in assembly lines [5]. In medical robotics, visual servoing enables real-time tool guidance based on intraoperative imaging. Several visual servoing paradigms exist: Image-Based Visual Servoing (IBVS) uses 2D image features (e.g., points, lines) to guide motion without estimating 3D pose. It is robust to calibration errors but may suffer from image singularities [6]. Pose-Based Visual Servoing (PBVS) estimates the 3D pose of the target and controls the robot in Cartesian space. It provides intuitive motion but requires accurate camera calibration [7]. Hybrid Visual Servoing combines IBVS and PBVS to balance robustness and accuracy, adapting to task-specific requirements [8]. Moment-Based Visual Servoing leverages global image features such as centroids or areas derived from image moments, which is useful when local features are unreliable [9]. Recent developments incorporate learning-based control to improve generalization. For instance, deep reinforcement learning has been applied to visual servoing in dynamic and partially occluded environments [10]. In surgical contexts, PBVS has enabled sub-millimetric precision using intraoperative camera feedback [11].

Our work adopts a hybrid visual servoing strategy that combines PBVS and IBVS to leverage the strengths of both methods. PBVS relies on accurate 3D pose estimation, which can be achieved using homography-based methods for planar scenes [8], or model-based tracking for general 3D objects. The latter aligns a known 3D model with image features such as edges or keypoints. KLT trackers are commonly used for 2D point tracking and are robust to out-of-plane motion. Examples include hybrid frameworks that integrate edge, texture, and depth information [12], Lucas–Kanade optical flow [13], and real-time markerless tracking systems [14]. More recently, learning-based methods such as FoundationPose [15], MegaPose [16], and GigaPose [17] have demonstrated strong performance in estimating 6D poses of novel objects in cluttered scenes. These approaches use dense template matching or transformer-based architectures and can generalize without retraining.

III. METHOD

The visual tracking process is divided into two components: 2D and 3D. The control strategy integrates a hybrid visual servoing scheme within a Quadratic Programming (QP) framework, as shown in Fig. 2. This approach ensures accurate trajectory tracking while satisfying physical and task-specific constraints.

A. Preliminars

The visual servoing framework guides the robot's end-effector to a desired pose relative to a target using visual features extracted from camera images. It is formulated as:

$$\dot{\mathbf{x}} = -\lambda \mathbf{L}_s^{+} \left(\mathbf{s}\left(\mathbf{I}\right) - \mathbf{s}^{*} \right) \tag{1}$$

where $\dot{\mathbf{x}} = \mathbf{v}$ denotes the desired end-effector velocity. The vector $\mathbf{v} = [\boldsymbol{v}^{\top}; \boldsymbol{\omega}^{\top}]^{\top}$ is six-dimensional, composed of the linear velocity $\boldsymbol{v} = [v_x; v_y; v_z]^{\top}$ and the angular velocity $\boldsymbol{\omega} = [\omega_x; \omega_y; \omega_z]^{\top}$. Here, λ is a positive control gain, $\mathbf{s}(\mathbf{I})$ represents the current feature vector, and $\mathbf{s}^{*}$ is the desired feature configuration. The interaction matrix $\mathbf{L}_s$ relates image features to Cartesian motion, and the superscript $(\cdot)^{+}$ denotes the generalized inverse.

In the hybrid scheme, both $\mathbf{s}(\mathbf{I})$ and $\mathbf{L}_s$ are defined according to their role in processing input data. For PBVS, 3D features extracted from the color camera provide the estimated 6-DoF pose of the ear, represented by the transformation matrix ${}^{c}M_o$. Thus, $\mathbf{s}$ corresponds to the current pose ${}^{c}M_o$, and $\mathbf{s}^{*}$ to the desired pose ${}^{c}M_o^{*}$ in the camera frame. The endoscopic camera supports IBVS by capturing image-plane features, specifically the 2D pixel coordinates (u, v) of anatomical landmarks. Here, $\mathbf{s} = (u, v)$ and $\mathbf{s}^{*}$ corresponds to the target penetration point. PBVS is employed during the ear-approach phase due to the complex geometry of the ear, where full 6-DoF pose estimation is advantageous. However, PBVS is sensitive to calibration errors; therefore, both intrinsic and extrinsic calibrations were performed for the color camera and endoscope.

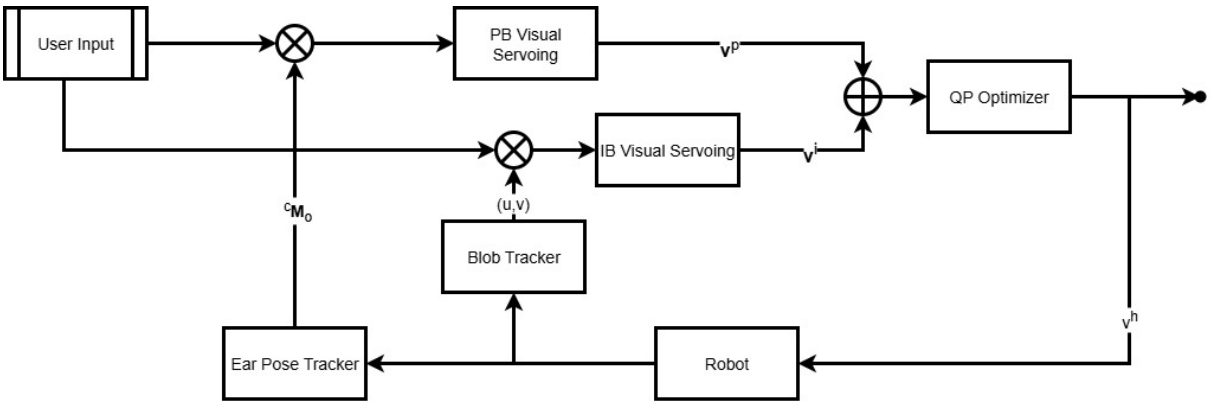


Fig. 2. Hybrid visual servoing workflow illustrating the phases of ear approach, tympanic membrane alignment, and sub-millimetric targeting.

IBVS is used for the membrane penetration phase. The nearly planar tympanic membrane makes IBVS robust to calibration errors, making it suitable for targeting a penetration point as small as 2 mm in radius. Since IBVS requires depth information, we estimate depth using the 3D pose from PBVS and the anatomical CAD model of the ear, ensuring accurate depth inference for the control law.

B. External Ear Auricle Tracking

To estimate the 6-DoF pose of the ear model, we employ the MegaPose framework [16], designed for 3D object pose estimation from a single RGB or RGB-D image. MegaPose addresses the challenge of generalizing to novel objects using a render-and-compare approach for pose refinement. Given the 3D CAD model of the target object (the ear), the system renders multiple synthetic views and uses them during both coarse and fine pose estimation steps. Its ability to generalize to unseen objects is particularly advantageous in surgical contexts, where anatomical variations are common. This generalization is enabled by training on a large-scale synthetic dataset of over 2 million photorealistic images and more than 20,000 3D object models.

In our system, MegaPose uses the ear's CAD model to obtain an initial pose estimate relative to the color camera, which is then fed to the PBVS controller for accurate end-effector positioning. A user initializes the process by selecting a bounding box around the ear in the image (Fig. 3).

C. Penetration Point Tracking

The tympanic membrane's flat structure and the small size of the penetration point (approximately 2 mm in radius) make blob tracking well-suited for target detection and tracking. We employ a blob tracker [18], which provides robust real-time performance (Fig. 4). The endoscopic camera offers a magnified, centered view of the ear canal, ensuring that the penetration point appears sufficiently large for reliable detection.

The tracker is initialized manually by selecting the penetration point in the endoscopic image. Once initialized, the blob tracker continuously follows the target and outputs its

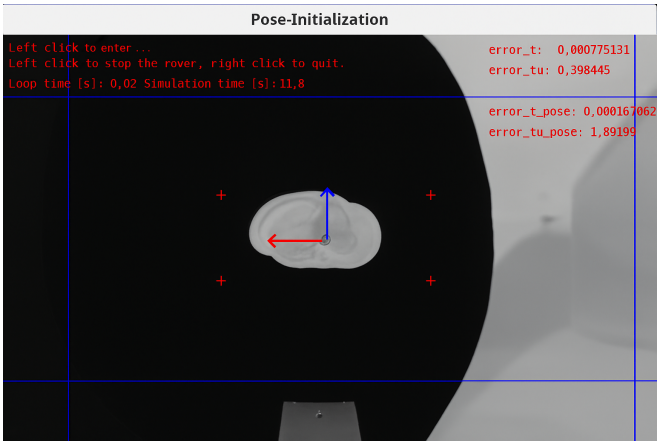


Fig. 3. Pose estimation results using the ear's 3D model and MegaPose-based tracking

pixel coordinates (u, v), which serve as the feature vector $\mathbf{s}$ for the image-based visual servoing (IBVS) control loop.

Although IBVS typically requires at least four feature points to fully constrain all six degrees of freedom, we simplify the control law by constraining all rotational motions and translation along the Z-axis. This allows the remaining degrees of freedom to be controlled using a single 2D feature point, significantly reducing complexity without compromising accuracy for this surgical task.

D. Maintaining Target Visibility via QP Optimization

In Position-Based Visual Servoing (PBVS), the robot computes motion based on the estimated 3D pose of the target. Although this approach enables intuitive Cartesian control, it suffers from a key limitation: the robot may follow a straight-line trajectory that causes the target to leave the camera's field of view (FoV). When this occurs, pose estimation fails and the visual servoing loop breaks.

To address this issue, we introduce a constraint-aware control strategy that maintains target visibility throughout the motion. The problem is formulated as a Quadratic Program (QP) that minimizes the deviation between the desired velocity $\mathbf{v}^*$ and the actual velocity $\mathbf{v}$, subject to linear inequality

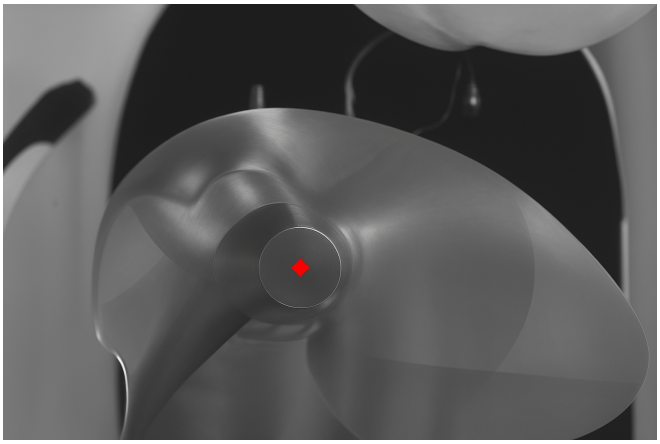


Fig. 4. Penetration point detection using blob tracking in an endoscopic view of the ear canal.

constraints:

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}} \quad \frac{1}{2}\|\mathbf{v} - \mathbf{v}^*\|^2 \quad \text{subject to} \quad \mathbf{C}\mathbf{v} \leq \mathbf{d} \tag{2}$$

The constraint matrix $\mathbf{C}$ and threshold vector $\mathbf{d}$ are constructed by projecting the 3D bounding box of the target onto the image plane using the pinhole camera model:

$$u = f_x \frac{x}{z} + c_x, \quad v = f_y \frac{y}{z} + c_y \tag{3}$$

For each projected point (u, v), the interaction matrix $\mathbf{L}_i = [\mathbf{L}_i^{(u)}; \mathbf{L}_i^{(v)}]^\top$ relates camera velocity to image motion:

$$\mathbf{L}_i^{(u)} = \left[-\frac{f_x}{z},\ 0,\ \frac{f_x x}{z^2},\ \frac{f_x xy}{z^2},\ -f_x\left(1 + \frac{x^2}{z^2}\right),\ \frac{f_x y}{z}\right] \tag{4}$$

$$\mathbf{L}_i^{(v)} = \left[0,\ -\frac{f_y}{z},\ \frac{f_y y}{z^2},\ f_y\left(1 + \frac{y^2}{z^2}\right),\ -\frac{f_y xy}{z^2},\ -\frac{f_y x}{z}\right] \tag{5}$$

Each bounding box corner contributes four constraints: one for each image border (left, top, right, bottom). These constraints are encoded in the matrix $\mathbf{C} \in \mathbb{R}^{32\times 6}$ as follows:

$$\mathbf{C}[i + 8k] = (-1)^{\delta_k} \cdot \mathbf{L}_i^{(j_k)} \tag{6}$$

where $i = 0, \ldots, 7$ indexes the corner point, and $k = 0, 1, 2, 3$ corresponds to the four borders. The functions δ_k and j_k are defined as:

$$\delta_k = \begin{cases} 1 & \text{if } k < 2 \quad \text{(left/top)} \\ 0 & \text{if } k \geq 2 \quad \text{(right/bottom)} \end{cases} \tag{7}$$

$$j_k = \begin{cases} u & \text{if } k \in \{0, 2\} \quad \text{(horizontal)} \\ v & \text{if } k \in \{1, 3\} \quad \text{(vertical)} \end{cases} \tag{8}$$

The corresponding threshold vector $\mathbf{d} \in \mathbb{R}^{32}$ defines the minimum distance each projected point must maintain from the image borders. A safety margin τ (typically 50 pixels) is applied, scaled by a factor $\alpha = 5$:

$$\mathbf{d}[i + 8k] = \begin{cases} \alpha(w - \tau - u) & \text{if } k = 0 \quad \text{(Left)} \\ \alpha(h - \tau - v) & \text{if } k = 1 \quad \text{(Top)} \\ \alpha(u - \tau) & \text{if } k = 2 \quad \text{(Right)} \\ \alpha(v - \tau) & \text{if } k = 3 \quad \text{(Bottom)} \end{cases} \tag{9}$$

This formulation guarantees that all bounding box corners remain within the visible region. The resulting velocity $\mathbf{v}_{\text{FOV}}$ respects both the PBVS objective and visibility constraints. Leveraging the 7-DoF redundancy of the robot, the controller adjusts joint configurations to maintain visibility without significantly deviating from the desired trajectory, improving robustness during critical phases.

E. Hybrid PBVS-IBVS Switching and Fusion Strategy

The control strategy consists of two sequential yet overlapping phases: ear canal approach and penetration point targeting. A smooth transition between these phases is essential for accuracy and patient safety.

In the first phase, PBVS guides the end-effector to a desired 6-DoF pose near the ear canal entrance, generating velocity $\mathbf{v}^{\text{p}}$. When the penetration point becomes visible in the endoscopic image, the user confirms the target and selects it via the console, initiating the second phase.

During the second phase, the control strategy fuses $\mathbf{v}^{\text{p}}$ with $\mathbf{v}^{\text{i}}$ —computed using IBVS— velocity vectors to compute the resulting velocity command $\mathbf{v}^{\text{h}} = \begin{bmatrix}\boldsymbol{v}_{xy}^{\text{h}} & \boldsymbol{v}_{z}^{\text{h}} & \boldsymbol{\omega}^{\text{h}}\end{bmatrix}^{\text{T}}$. Concretely, the designed fusion strategy is depicted as follows.

First, **rotational motions** are taken directly from $\mathbf{v}^{\text{p}}$, thus $\boldsymbol{\omega}^{\text{h}} = \boldsymbol{\omega}^{\text{p}}$. This preserves the perpendicular orientation of the tool with respect to the tympanic membrane plane, established during the first phase. Secondly, **lateral translations** along the X- and Y-axes are combined using a weighted sum of $\mathbf{v}^{\text{p}}$ and $\mathbf{v}^{\text{i}}$, such as,

$$\boldsymbol{v}_{xy}^{\text{h}} = (\beta_1)\, \boldsymbol{v}_{xy}^{\text{p}} + (1 - \beta_1)\, \boldsymbol{v}_{xy}^{\text{i}}. \tag{10}$$

This compensates potential patient micro-movements ($\mathbf{v}^{\text{p}}$) and aligns the tool tip precisely above the selected penetration point ($\mathbf{v}^{\text{i}}$). Finally, **depth translation** —along Z-axes— is manually controlled by the user to ensure gradual and safe insertion. The user receives live visual feedback on the current distance between the tool tip and the target. This distance is calculated from the 3D pose estimation from the 6-DoF pose tracker and the known anatomy from the 3D CAD model of the ear. Thus, the final z component of the linear velocity is composed as $v_z^{\text{h}} = (\beta_2)\, v_z^{\text{p}} + (1 - \beta_2)\, v_z^{\text{u}}$, where v_z^{u} is the user controller penetration speed.

To ensure all movements remain safe and within operational constraints (e.g., maintaining target visibility and avoiding collisions), the final control output –velocity vector– is passed through the Quadratic Programming (QP) optimization layer $\mathbf{v}^* = \mathbf{v}^{\text{h}}$ as described in the previous section. The solver minimizes deviations from the desired motion while satisfying physical constraints and visual field-of-view constraints.

F. Human Interaction Approach

To facilitate precise control during process, we incorporate interactive user guidance devoted to translation correction, rotation alignment, and approach motion. Each interaction modifies the target pose of the end-effector relative to the ear ${}^{o}M_{c}^{*} = [\mathbf{R} \quad \mathbf{t}]$. Where $\mathbf{R} \in \mathrm{SO}(3)$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector.

a) Interaction command `T`: The first interaction — Translation Correction— helps the robot to point the tool directly above the target by lateral displacement. This is achieved by setting the horizontal components of $\mathbf{t}$ vector to zero, while preserving the vertical component:

$$\mathbf{t} = \begin{bmatrix} 0 & 0 & t_z \end{bmatrix}^{\mathrm{T}} \tag{11}$$

b) Interaction command `R`: The second interaction — Rotation Alignment— guide the robot to point the tool perpendicular to the surface of the ear, e.i. parallel to z-axis of the tympanic membrane, by reorienting the end-effector. Let $\mathbf{z}_c = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$ represent the end-effector's local z-axis. We represent it in the ear's frame as,

$$\mathbf{z}_o = \mathbf{R} \cdot \mathbf{z}_c \tag{12}$$

To align $\mathbf{z}_c$ with $-\mathbf{z}_o$, a correction rotation matrix $\mathbf{R}_{\text{align}}$ is computed using the angle-axis method. Where, the rotation axis $\mathbf{u}$ is given by:

$$\mathbf{u} = \frac{\mathbf{z}_c \times (-\mathbf{z}_o)}{\|\mathbf{z}_c \times (-\mathbf{z}_o)\|} \tag{13}$$

and the rotation angle is defined as $\theta = \cos^{-1}\left(\mathbf{z}_c^{\top} \cdot (-\mathbf{z}_o)\right)$. Thus, the correction matrix $\mathbf{A}$ is then constructed as:

$$\mathbf{A} = \exp(\theta[\mathbf{u}]_{\times}) \tag{14}$$

where $[\cdot]_{\times}$ denotes the skew-symmetric matrix associated with a vector. An the resulting orientation matrix is obtained by,

$$\mathbf{R} \leftarrow \mathbf{A} \cdot \mathbf{R} \tag{15}$$

c) Interaction command `Z`: The third interaction — Approach Motion— helps to guide the robot towards the gradual descent of the end-effector along the z-axis. At each step, the vertical component t_z is decremental updated if $t_z > 0.01$,

$$t_z \leftarrow t_z - 0.05 \tag{16}$$

Once the threshold is reached, the descent is finalized by setting $T_z \leftarrow 0$.

The simplicity of the proposed T–R–Z interface was a deliberate choice guided by feedback from our clinical collaborator. During microsurgical procedures, surgeons face a high cognitive load, where even small additional stimuli—such as multiple displays or extra control inputs—can increase mental effort and reduce focus. Our design goal was therefore to minimize unnecessary information and provide an intuitive, low-stress interaction method. The interface was implemented on a PC using simple keyboard commands, but this minimal structure also facilitates later integration with medical-grade interfaces such as foot pedals or haptic controllers. Future work will extend this scheme toward more advanced modalities (e.g., haptic feedback or semi-autonomous control) while preserving its low cognitive demand.

IV. EXPERIMENTS

All experiments were conducted using a workstation equipped with an Intel Core i9-10900 CPU @ 2.8GHz with 130 GB RAM, and an NVIDIA RTX 3070 GPU. The robot platform used was a Franka Emika FR3 arm (7 DoF) . The sensors used where a camera IDS UI-3240CP-C-HQ with a optical lens FUJINON HF8XA-5M, and a generic endoscope (480p @ 30hz, 70 deg FoV). The system was implemented in ROS2 Humble and Python 3.10 on Ubuntu 22.04. Simulations were conducted in CoppeliaSim, and control algorithms were developed using ViSP 3.6.

A. Camera Calibration and Synchronization

To ensure accurate spatial alignment and temporal coherence between visual inputs, both cameras are calibrated and synchronized. Intrinsic and extrinsic parameters are estimated using Zhang's method [19], with respect to the robot base frame. Temporal synchronization is achieved through timestamp alignment.

B. Evaluation Protocol

The procedure begins with an initial estimation of the outer ear pose, which serves as a precursor to approaching the ear canal. A graphical user interface displays a live image from the color camera, prompting the user to select two points that coarsely bound the ear region. These selections provide a rough localization of the ear in the image and initialize the 3D ear pose tracker server to obtain ${}^{\mathbf{c}}\mathbf{M}_{\mathbf{o}}$.

Subsequently, the user start to interact with the robot via the interaction commands, with the sequence `T`, `R`, and `Z`. This interaction involves confirming the visibility of the tympanic membrane and identifying the penetration point. The user verifies the target location via the endoscopic camera and clicks on the desired point to initialize a blob tracker. The selected point is then projected onto the image, facilitating precise localization of the needle entrance.

Final positioning of the needle is conducted similarly but just two actions are needed. The user presses `T` to align the needle tip with the selected point and `Z` to descend, guided by distance feedback. Orientation correction is not required at this stage, as it was previously addressed during outer ear alignment. When the distance reaches zero, the needle tip is accurately aligned with the penetration point on the tympanic membrane.

V. EVALUATION

The proposed system was evaluated in two complementary stages: first in a simulated environment to allow controlled analysis of pose estimation and control performance, and then on a real robotic platform to validate robustness under physical constraints and real-world uncertainties.

A. Simulation-Based Evaluation

The analysis of the ear pose tracker —MegaPose— robustness (Table I) revealed that translation error remained generally below 1 mm for distances between 34 cm and 20 cm, with occasional transient spikes during large robot motions such as combined translation and orientation commands or descent steps. These spikes were observed at approximately 32 cm, 27 cm, and 22 cm, but the error quickly stabilized to sub-millimetric values after each perturbation. When the distance decreased below 20 cm, the system exhibited improved stability, and no significant spikes were detected. However, at 10 cm, pose estimation degraded noticeably because the object nearly filled the entire camera field of view, leading to overshoot and estimation failure. Orientation error remained robust throughout the process, typically around or below 1 degree, with degradation only at the closest distance limit. These results, illustrated in Fig. 5, confirm that MegaPose provides accurate pose estimates across most of the operational range, with limitations only at extreme proximity.

TABLE I

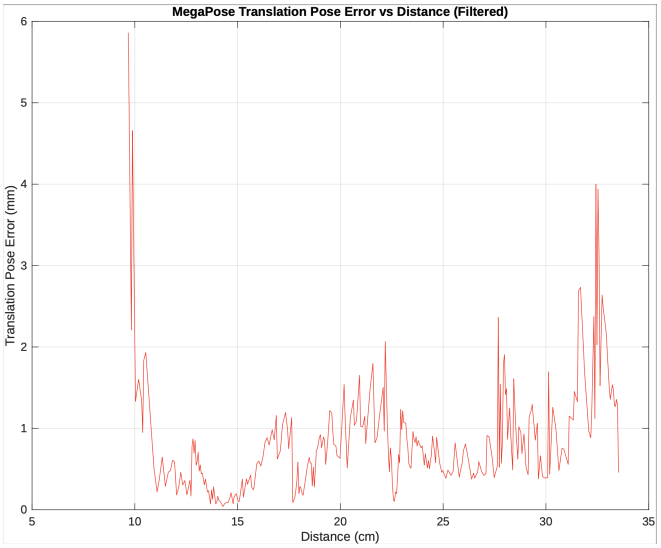
ERROR SUMMARY OF MEGAPOSE AT DIFFERENT DISTANCE RANGES

Distance Range	Translation Error	Orientation Error
$>$ 20 cm	$\approx$ 1 mm	$\approx$ 1 deg
10–20 cm	$<$ 1 mm	$\approx$ 0.8–1 deg
$\leq$ 10 cm	Estimation fails	Estimation fails

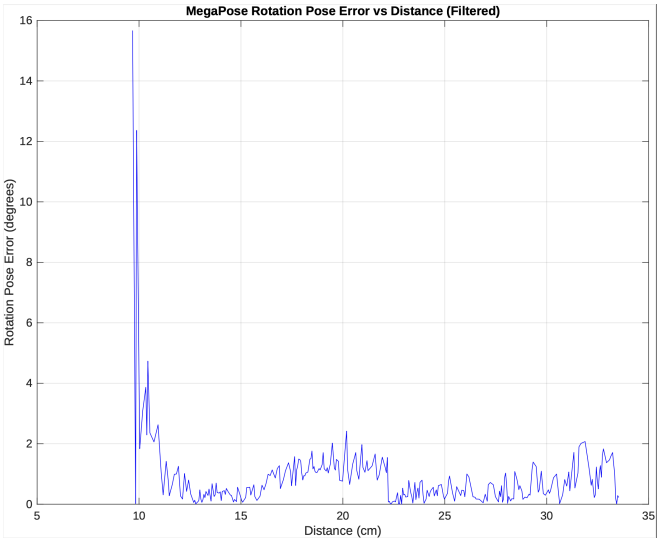
The performance of the visual servoing controllers was also analyzed in simulation. For the position-based component, the evolution of translation and rotation errors over time (Fig. 6) showed distinct peaks corresponding to specific control actions, such as initial translation correction, orientation alignment, and descent steps. Despite these transient deviations, the PBVS control law consistently drove both errors toward zero, achieving sub-millimetric precision after stabilization. Across ten trials, the translation error converged to a mean value of approximately 0.88 mm, with a minimum of 0.088 mm and a maximum of 2.61 mm. The rotation error stabilized around a mean of 0.71 deg, with a maximum of 4.94 deg. Both are summarized in Table II. For the image-based component, which was used to track the penetration point, the error in image space—defined as the difference between current and desired pixel coordinates—decreased smoothly over time. Both the individual pixel components and the Euclidean norm of the error converged to near zero, as shown in Fig. 7. The mean steady-state pixel error norm was approximately 0.52 px, corresponding to a spatial deviation of about 0.095 mm at an estimated depth of 7 cm (Table III). These results confirm that both controllers exhibit stable and precise convergence behavior under simulated conditions.

B. Real-Robot Experiments

The experimental protocol involved tracking the outer ear while introducing deliberate perturbations in both translation and rotation to challenge the robustness of pose estimation and control.



(a)



(b)

Fig. 5. Ear's pose tracker. (a) Translation error (mm) versus camera-object distance. (b) Orientation (degrees) versus camera-object distance.

The observations in the real setup were consistent with those obtained in simulation. MegaPose maintained accurate pose estimation across most of the operational range, with performance degradation occurring only when the camera approached the target at very close distances, where the object occupied nearly the entire field of view. This limitation, already identified in simulation, was confirmed as the primary factor influencing control performance. Despite these challenges, the PBVS controller successfully compensated for pose deviations and achieved sub-millimetric convergence after stabilization, provided that pose estimation remained reliable. Similarly, the IBVS controller demonstrated smooth and stable convergence of the image-space error, ensuring precise tracking of the penetration point even under moderate

TABLE II

PBVS CONVERGING ACCURACY

Error Type	Mean	Min	Max
Translation Error (mm)	0.8795	0.0880	2.6080
Rotation Error (deg)	0.7090	0.0000	4.9423

target motion.

Overall, these real-robot experiments validate the robustness and accuracy of the proposed system. The results confirm that the control laws are capable of achieving high precision in both translation and orientation, and that the main limiting factor is the reliability of pose estimation at extreme proximity rather than the control strategy itself.

TABLE III

STEADY-STATE IBVS PIXEL ERROR STATISTICS AND ESTIMATED SPATIAL ERROR

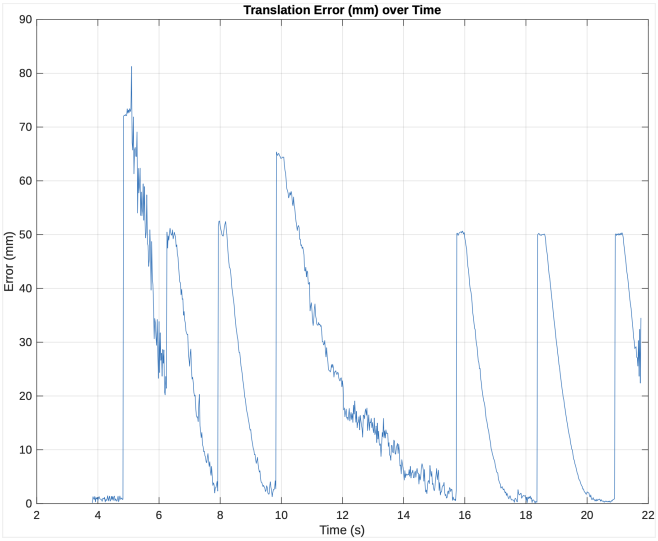
Component	Pixel Error [px]			Spatial Error [mm]		
	Mean	Min	Max	Mean	Min	Max
u	-0.042	-0.050	0.017	-0.008	-0.009	0.003
v	0.514	0.500	0.592	0.094	0.092	0.109
Norm	0.516	0.503	0.592	0.095	0.092	0.109

VI. CONCLUSION

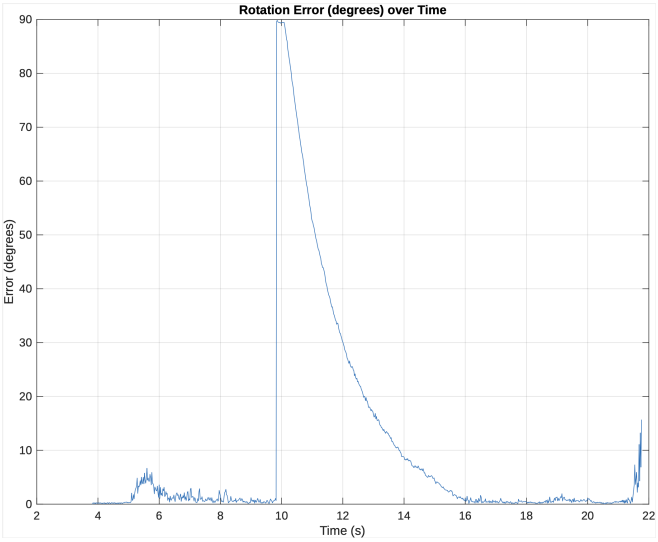
This paper introduced a hybrid visual servoing framework for robotic assistance in minimally invasive middle ear surgery. By integrating PBVS and IBVS strategies with dual-camera feedback and a constraint-aware QP controller, the system achieves sub-millimetric accuracy while preserving safety constraints. The combination of MegaPose-based 3D pose estimation and blob tracking for fine targeting demonstrated robustness against anatomical variability and motion disturbances.

Simulation and real-robot experiments validated the approach, confirming precise and stable tool guidance throughout the surgical workflow. The modular design facilitates future extensions, including complete automation, integration of force sensing, and adaptation to other microsurgical domains. These results highlight the potential of hybrid visual servoing to enhance surgical precision and reduce surgeon workload in ENT procedures. Although the system demonstrated robust performance throughout most of the operational range, a degradation in MegaPose accuracy was observed when the camera approached the ear closer to 10 cm. This limitation can be mitigated by designing the camera–tool mount so that the end-effector reaches the target before such proximity is reached. Moreover, the QP-based visibility constraint provides an additional safeguard by preventing the camera from moving too close, thus maintaining a reliable pose estimation.

In future work, realistic anatomical or cadaver specimens will be evaluated to assess performance under conditions closer to surgery, including variable lighting, partial occlusions, and fluid reflections. This validation will be essential to confirm the robustness and clinical transferability of the proposed hybrid visual servoing approach.



(a)



(b)

Fig. 6. PBVS Error Convergence. (a) Temporal evolution of translation error during PBVS control. (b) Temporal evolution of orientation error.

REFERENCES

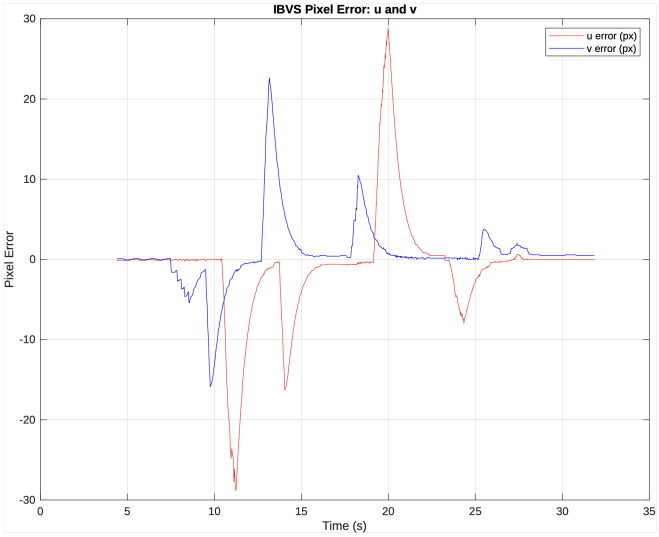
[1] A. Saadoun *et al.*, "Minimally invasive ossiculoplasty via an endoscopic transtympanic approach," *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 141, no. 2, pp. 93–97, 2024, doi: 10.1016/j.anorl.2023.08.002.

[2] W. Zhang *et al.*, "State of the art in movement around a remote point: A review of Remote Center of Motion in robotics," *Frontiers of Mechanical Engineering*, vol. 19, no. 2, 2024, doi: 10.1007/s11465-024-0785-3.

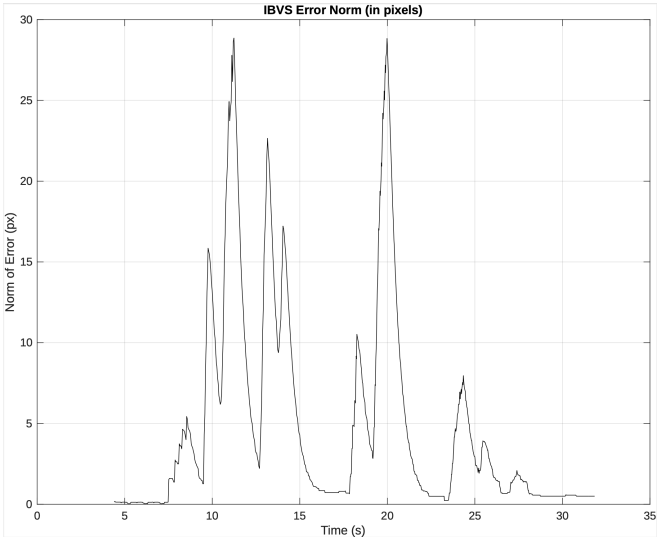
[3] S. Vittoria, G. Lahlou, R. Torres, H. Daoudi, I. Mosnier, S. Mazalaigue, E. Ferrary, Y. Nguyen, and O. Sterkers, "Robot-based assistance in middle ear surgery and cochlear implantation: first clinical report," *European Archives of Otorhinolaryngology*, vol. 278, no. 1, pp. 77–85, Jan. 2021, doi: 10.1007/s00405-020-06070-z.

[4] M. Miroir *et al.*, "RobOtol: from design to evaluation of a robot for middle ear surgery," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, 2010, pp. 850–856, doi: 10.1109/IROS.2010.5650390.

[5] A. Rosales, T. Heikkilä and M. Suomalainen, "Visual Servoing Based on 3D Features: Design and Implementation for Robotic Insertion



(a)



(b)

Fig. 7. IBVS Tracking Performance. (a) Pixel errors in u and v coordinates during penetration point tracking. (b) Euclidean norm of pixel error over time.

Tasks," 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 2024, pp. 1-6.

[6] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, Jun. 1992, doi: 10.1109/70.143350.

[7] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, Oct. 1996, doi: 10.1109/70.538972.

[8] F. Chaumette and S. Hutchinson, "Visual servo control. II. Advanced approaches," *IEEE Robot. Autom. Mag.*, vol. 13, no. 1, pp. 109–118, Mar. 2006.

[9] O. Tahri and F. Chaumette, "Point-based and region-based image moments for visual servoing of planar objects," *IEEE Transactions on Robotics*, vol. 21, no. 6, pp. 1116–1127, Dec. 2005, doi: 10.1109/TRO.2005.853500.

[10] J. Lee and J. Song, "Deep reinforcement learning for adaptive visual servoing under occlusions," *Robotics and Autonomous Systems*, vol. 163, p. 104354, 2023.

[11] A. Krupa *et al.*, "Autonomous 3-D positioning of surgical instruments in robotized laparoscopic surgery using visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 842–853, Oct. 2003, doi: 10.1109/TRA.2003.817086.

[12] S. Trinh, F. Spindler, E. Marchand, and F. Chaumette, "A modular framework for model-based visual tracking using edge, texture and depth features," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 89–96, doi: 10.1109/IROS.2018.8594003.

[13] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, Mar. 2004, doi: 10.1023/B:VISI.0000011205.11775.fd.

[14] A. Comport, E. Marchand, M. Pressigout, and F. Chaumette, "Real-time markerless tracking for augmented reality: The virtual visual servoing framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 615–628, Jul.-Aug. 2006, doi: 10.1109/TVCG.2006.78.

[15] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 17868–17879, doi: 10.1109/CVPR52733.2024.01692.

[16] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "MegaPose: 6D pose estimation of novel objects via render & compare," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2022, pp. [page numbers if available].

[17] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 9903–9913, doi: 10.1109/CVPR52733.2024.00945.

[18] E. Marchand, F. Spindler, F. Chaumette, "ViSP for visual servoing: a generic software platform with a wide class of robot control skills" in *IEEE Robotics & Automation Magazine*, 12(4):40–52, 2005, doi:10.1109/MRA.2005.1577023

[19] Z. Zhang, "A flexible new technique for camera calibration" in *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2002, doi:10.1109/34.888718