Towards Exoscope Automation in Neurosurgery: A Markerless Visual-Servoing Approach

Elisa Iovene*, Alessandro Casella*, Alice Valeria Iordache, Junling Fu, *Graduate Student Member, IEEE*, Federico Pessina, Marco Riva, Giancarlo Ferrigno, *Senior Member, IEEE* and Elena De Momi, *Senior Member, IEEE*

***Abstract*—Exoscopes are a promising tool for neurosurgeons, offering improved visualisation and ergonomics compared with traditional surgical microscopes. They consist of an external scope that projects the surgical field onto a 2D or 3D monitor, providing a wider field of view and better access to the surgical site. Despite the advantages, exoscopes present some limitations, such as the need for manual or foot joystick repositioning, which can disrupt the flow of the procedure and increase the risk of user error. In this study, a markerless visual-servoing approach for autonomous exoscope control is proposed to address these limitations and enhance the ergonomics and reduce the physical and cognitive load compared with traditional joystick control. The system uses visual information from the operating field to control the exoscope, eliminating the need for markers or additional tracking devices. The proposed approach was validated using a 7-DOF robotic manipulator with a stereo camera in an eye-in-hand configuration. Results showed that the system achieved 89 % accuracy in detecting the target and tracking its movement with a tracking error ranging from 0.50 ± 0.17 cm for low-speed movements to 1.38 ± 0.73 cm for high-speed movements. The proposed system also demonstrated improved efficiency, with a shorter execution time of 72.07 ± 19.36 s compared with 106.52 ± 18.50 s for the foot-joystick control. Additionally, the time out of the FoV was significantly higher in the joystick control mode and the frequency of appearance of the instrument in the centre of the image was higher when using the proposed system. The NASA TLX results indicated lower physical and cognitive load compared with the joystick control-based modality.**

***Index Terms*—Visual servoing, Autonomous Camera Control, Exoscope placement, Surgical Robotics, Neurosurgery, Instrument Tracking**

I. Introduction

THE introduction of the exoscope in neurosurgery has led to a more flexible and ergonomic working environment for the neurosurgeon, as viewing images through the microscope eyepiece is no longer required, resulting in a more comfortable operating position [1]. The exoscope is a surgical telescope with high-quality magnification that can be mounted above the surgical scene by means of holding arms or robotic manipulators. Additionally, the system includes a 3D screen with high-definition (or 4K) resolution through which the surgeon can observe the images captured by the exoscope, and a wireless foot switch as shown in Fig. 1. Compared with the standard microscope, the exoscope provides comparable magnification, lighting, and high-definition images in all kinds of cranial and spine surgery, both for surgeons and operating room (OR) staff, as well as more manageability and higher surgical comfort [2].

Fig. 1. Operating room setting for neurosurgery with exoscope system. On the right is the holding arm with the 3D exoscope. A foot pedal is placed close to the surgeon to operate the system. Images from the exoscope are projected on external monitors that the surgeon watches while performing surgery.

Although the surgeon is provided with an enhanced view and extended working area, the surgeon is required to manually control it whenever the exoscope system has to be re-positioned for better vision capture, causing interruptions and distractions during surgical procedures. Exoscope systems such as the VITOM (Karl Storz, Tuttlingen, Germany) use a pneumatic holder that relies on compressed air to create a mechanical movement, a 3D wheel and four programmable function keys to control the camera [3]. The surgeon can position the exoscope as desired and then lock its position once again. However, this inevitably results in disrupted operational workflow, distractions, longer operating time, and increased surgeon's mental workload [4]. Alternatives to manual repositioning have been proposed. The most popular one, currently adopted by AESCULAP Aeos (Braun, Melsungen, Germany) and ORBEYE (Olympus, Tokyo, Japan), is a foot-operated joystick controller that, together with a pilot unit, controls the movement of the exoscope mounted on a robotic arm.

E. Iovene, J. Fu, G. Ferrigno, and E. De Momi are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy. (e-mail: elisa.iovene@polimi.it; junling.fu@polimi.it; giancarlo.ferrigno@polimi.it; elena.demomi@polimi.it) (*These authors equally contributed. Corresponding author: Junling Fu)

A. Casella is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano and also with the Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy (e-mail: alessandro.casella@polimi.it)

F. Pessina and M. Riva are with the Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy and also with the IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy (e-mail: federico.pessina@hunimed.eu, marco.riva@hunimed.eu)

This allows the surgeon to remain ambidextrous throughout the procedure, but the complexity of its use has emerged as a limiting factor [5]. The inconvenience of repositioning is likely to reduce the benefits of using the exoscope in the operating room and emphasises the importance of hands-free camera movement.

One promising solution to reduce the workload for the surgeon, and improve ergonomics and efficiency is to increase the level of autonomy (LoA) of the surgical robotic system to help the surgeon obtain the optimal viewpoint without the need for repetitive readjustments [6]. Considering the superiority of spatial positioning accuracy, dexterity, and non-fatigability of robotics, robot-assisted autonomous navigation techniques have been integrated into neurosurgery applications [7]. The first autonomous camera control was introduced by Synaptive Medical with the release of *ModusV* in 2017 [8]. This system enabled hands-free manipulation using optical tracking systems together with passive markers attached to the suction cannula. The main superiority of this system was the robotic movement of the camera. However, obstructions between the instrument and the tracking system could prevent the operation of hands-free control. In addition, modification of the tracked instrument was required to attach passive markers, which limited the instrument's manoeuvrability. Hence, determining how to enhance the LoA of robotic systems while ensuring safety and efficiency in neurosurgical scenarios remains a critical issue.

This work proposes an autonomous vision-guided exoscope robotic holder that aims to provide the surgeon with greater ergonomics during neurosurgery, reduce physical and cognitive load, and shorten operation time compared with the current foot-joystick camera control. The system is based on markerless visual-servoing techniques, which allow surgeons to operate without using their hands to control camera movements and focus entirely on the surgical procedure. Moreover, the surgeon is able to turn the autonomous camera movement on and off at any time, via a foot pedal, as well as to adapt it to specific surgical steps that require special viewpoints.

II. Related Work

Several strategies have been proposed to automate camera movement and reduce the need for the surgeon to perform secondary tasks during the surgical procedure. In this section, a detailed analysis of the main approaches is described.

A. Eye Gaze and Voice Control

Gaze and voice control are two of the most popular approaches to camera automation. Gaze control involves using gaze information and the user's eye movement to control the camera view. [9] proposed a system for laparoscopic surgery that uses gaze tracking to control the movement of a flexible robotic gastroscope. Although users reported positive "satisfaction" scores and acknowledged the usefulness of the system, they showed significantly faster performance using conventional endoscopy. In addition, these control methods rely on tracking technologies that are still relatively unreliable [10]. In voice control, the surgeon's voice commands are converted by the control unit and then sent to the motor controller, which moves the camera accordingly. Voice controls typically allow for zoom in/out, up/down, and right/left movements, but precise positioning can be complex. [11] presented a new voice-controlled endoscope that achieves operating times comparable to those of conventional endoscopy. However, it still has a non-negligible number of errors that could pose a risk during surgical procedures. Moreover, both strategies require the direct intervention of the surgeon to control the camera.

B. Instrument Tracking

An autonomous camera system that can understand the surgeon's intentions without instructions is one way to reduce the workload of the surgeon during the repositioning of the scope. In the field of surgery, monitoring the way instruments are manipulated by the surgeon is one of the most widely used methods for understanding the surgeon's intentions, reducing the cost of human assistants, and helping realise autonomous navigation [12]. Many methods have been proposed to track instruments. [13] and [14] have used the kinematic chain of the robotic manipulator to track the position of instruments and employ it in the control of the camera. However, this method cannot be used in some traditional surgeries, such as in neurosurgery where surgeons manipulate instruments free-hand without relying on robotic assistance.

Optical tracking systems represent another alternative. Instrument tracking can be done by attaching active or passive markers on the instruments as the reference point for the target location. However, these techniques have many drawbacks, such as limited working space, and poor performance in the presence of light variations and occlusions [15]. Moreover, optical tracking systems require modification of the tracked instrument to which markers can be attached. This reduces the manageability of the instrument and can result in higher costs [16]. A marker-based approach for the automation of the da Vinci endoscope is proposed by [17]. In this case, the detection of the tool is achieved via ArUco codes which may fail in a real scenario when smoke and blood or other fluids are present.

Computer vision methods eliminate the need for external sensors or additional markers and can overcome the limitations described above. Using information from vision sensors without the need for additional markers is a very convenient strategy, as a rich spectrum of information can be extracted from image data. In addition, machine learning and deep learning have become popular in object recognition and pose estimation [18]. Moreover, the extraction of visual features can be used as input to the control law of the robotic manipulator, through a visual servo loop. Therefore, the integration of advanced computer vision techniques that retrieve important information from a highly complex and unstructured surgical scenario can become a promising approach to improve the LoA of the robot-assisted exoscope system during neurosurgical procedures.

C. Visual Servoing

Visual-servoing control is a technique that relies on computer vision data to control the motion of robots. It has been widely integrated into various scenarios, such as automotive and industrial robotics but has only gained more attention in the medical field in the last decade. Two different control approaches can be considered, namely Position-Based Visual Servoing (PBVS) and Image-Based Visual Servoing (IBVS) [19]. IBVS control schemes use the error between current and desired visual features on the image plane and do not involve any estimation of the target pose while PBVS control schemes use the camera pose with respect to some reference coordinate frame to define the error. Again, image features are extracted but they are used to estimate the 3D position of the object in Cartesian space. The reconstructed position of the point of interest can eventually be used as feedback in the control loop of a robotic manipulator. This approach was used in the laparoscopy field by [20]. Here, a tooltip localisation method based on surgical tool segmentation and a visual-servoing approach was proposed. And, in the field of microsurgery, [21] proposed a markerless PBVS technique to improve the accuracy of surgical procedures. In this study, a stereo microscope was used to track the tip of a handheld micro-manipulator to help the surgeon reach a target with high accuracy and avoid collision with anatomical structures that could lead to complications. The identification of the tip on the image plane was achieved by marking it with colored paint, which is not representative of an actual surgical scenario. Here, the potential of visual servoing in the medical field was demonstrated despite the simplified tracking techniques used. In the field of endoscopy, an autonomous system was proposed by [22] where only the endoscope was used as a vision sensor to segment and track surgical instruments, and a visual servo approach ensured smooth and appropriate movements of the endoscope.

III. Materials and Methods

The method proposed in this study uses an eye-in-hand markerless visual-servoing approach that can recognise and track a selected surgical instrument and consists of three main steps. First, a convolutional neural network (CNN) detects and estimates the center of the distal region (CDR) of the surgical tool in the 2D image space. In this way, the information provided by the camera was employed in the control loop of the robotic holder, and no additional external sensors or markers are needed. Second, the 3D position of the target in the robot space was retrieved using pose reconstruction algorithms. Finally, the desired pose of the target was sent to the visual-servoing controller, which was responsible for zeroing the error between the desired and actual position. The aim of the visual-servoing algorithm was to keep the camera in a fixed relative position to the instrument so that it always remains in the field of view (FoV), ideally in the center of the image.

Note that, in the neurosurgical scenario, only the movements of the camera in 2D space (along the X-Y plane) need to be considered and implemented since the movements along the depth direction (Z-axis) are only used for adjusting the focus of the camera. Furthermore, movements along the Z-axis would cause the manipulator to occupy the surgeons' workspace, limiting their dexterity and causing possible interference during surgery. In this regard, in our framework, the Z-axis movement was controlled by an autonomous zoom control module.

The schematic of the overall system can be divided into two modules as shown in Fig. 2. From the bottom up, the vision module includes the detection of the CDR of the tool, the reconstruction of its 3D position and the zoom system and the control module of the robotic manipulator.

A. Vision Module

The vision module of the proposed framework was composed of an object detection neural network (i.e., Yolov3 [23])

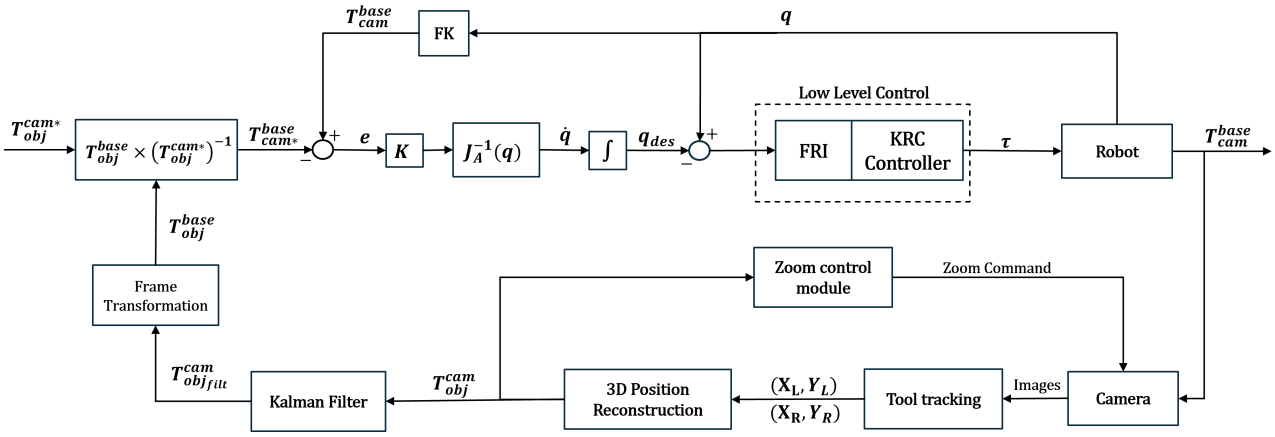

Fig. 2. System overview: from the bottom, the instrument position $(\mathbf{x_R}, \mathbf{y_R})$ and $(\mathbf{x_L}, \mathbf{y_L})$ in the image space is identified, the cartesian position $P(X, Y, Z) = \boldsymbol{T_{obj}^{cam}}$ is estimated. The Z position of the tool is sent to the Zoom control module to adjust the zoom level of the camera. The position $\boldsymbol{T_{obj}^{cam}}$ filtered by the Kalman Filter $\boldsymbol{T_{obj_{filt}}^{cam}}$ together with the desired one $\boldsymbol{T_{obj}^{cam*}}$ gives the desired position in the robot frame $\boldsymbol{T_{cam*}^{base}}$. The error $\boldsymbol{e}$, given by the desired position and the actual position $\boldsymbol{T_{base}^{cam}}$ retrieved using the forward kinematics, is sent to the controller that uses the pseudoinverse of the manipulator's Jacobian matrix $\boldsymbol{J_{Apseudo}^{-1}}$ to compute the required joint positions $\boldsymbol{q_d}$ that compensate for the error. This, together with the actual joint position $\boldsymbol{q}$, is sent to the low level control stage that outputs the required torque $\boldsymbol{\tau}$ to move the camera $\boldsymbol{T_{cam}^{base}}$.

trained to identify the distal region of the tip of laparoscopic forceps. Images were acquired from a Full HD stereo camera mounted on the end-effector (EE) of the robotic manipulator. Left and right RGB frames, with a resolution of 960x1080 pixels each, were concatenated horizontally in a single RGB image of 1920x1080 to process them simultaneously, thus minimising the delay between instrument detection and position transmission to the controller as much as possible. After the concatenation, the images were downsampled to 416x416 using bilinear interpolation and fed to the CNN. The CNN predicts the position of the distal region of the tip through a bounding box (one for each stereo image) and provides the confidence of the prediction. After the identification of the instrument, the coordinates of the CDR of the tool were considered to be the center of the bounding box (X_R, Y_R), (X_L, Y_L) as shown in Fig. 3. A pre-trained model of the object

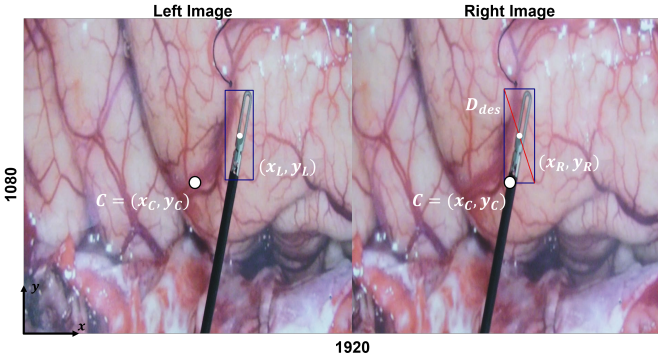

Fig. 3. The images coming from right and left camera were put together in a single image 1920 x 1080. $(\mathbf{x_R}, \mathbf{y_R})$ and $(\mathbf{x_L}, \mathbf{y_L})$ represent the coordinates of the center of the bounding boxes in the image plane. $(\mathbf{x_C}, \mathbf{y_C})$ is the center of the image plane. $\mathbf{D_{des}}$ is the desired size of the bounding box diagonal.

detection CNN on the Common Objects in Context (COCO) dataset is available [24]. However, this dataset does not include images of medical instruments; therefore, to train the CNN in identifying surgical tools, a fine-tuning was performed with a custom dataset for instrument tool detection.

Once the 2D coordinates of the CDR of the instrument were extracted in both images, the 3D position was computed by triangulation with respect to (w.r.t.) the right camera reference frame, using the direct linear transform (DLT). The DLT algorithm considered a linear relationship between the 2D coordinates in the image plane and the 3D coordinates in the camera reference frame $\boldsymbol{X}$:

$$\boldsymbol{A}\boldsymbol{X} = \mathbf{0} \tag{1}$$

where $\boldsymbol{A}$ is a matrix containing the 2D camera space coordinates and the projection components of the projection matrix. Solving for $\boldsymbol{A}$ by means of singular value decomposition (SVD) algorithm, the 3D coordinate $\boldsymbol{X}$ was reconstructed. As a result, the current cartesian position of the tool $\boldsymbol{P(X,Y,Z) = T_{obj}^{cam}}$ w.r.t. the camera frame was estimated.

In addition, a Kalman Filter [25] was adopted to minimise the reconstruction noise caused by noisy image coordinates estimation. The motion of the instrument was modeled with a 3D constant velocity model and was used to predict the position of the tool at the next time step $n+1$:

$$\hat{\boldsymbol{x}}_{n+1,n} = \boldsymbol{F}\hat{\boldsymbol{x}}_{n,n} + \boldsymbol{w_n} \tag{2}$$

where $\boldsymbol{F}$ is the state transition matrix, $\hat{\boldsymbol{x}}$ is the system state vector represented by the coordinates X, Y, Z and the velocities $\dot{X}$, $\dot{Y}$, $\dot{Z}$ of the instrument tip and, $\boldsymbol{w_n}$ is the process noise modeled as discrete white noise. $\hat{\boldsymbol{x}}_{n,n}$ is the system state vector estimated at time step n from the state update equation:

$$\hat{\boldsymbol{x}}_{n,n} = \hat{\boldsymbol{x}}_{n,n-1} + \boldsymbol{K_n}(z_n - \boldsymbol{H}\hat{\boldsymbol{x}}_{n,n-1}) \tag{3}$$

where $\boldsymbol{H}$ is the observation matrix, $\boldsymbol{z_n}$ is the noisy measurement, and $\boldsymbol{K_n}$ is the Kalman gain that weights the prediction given by the state equation and the measurement in the new position estimate, $\hat{\boldsymbol{x}}_{n,n}$. The filter computes the position uncertainty as the variance between all estimations. The measurement uncertainty and the process noise represent the filter's hyperparameters, which were empirically selected and listed in Table I.

TABLE I
KALMAN FILTER'S HYPERPARAMETERS

Hyperparameter	Value $\left(\text{mm}^2\right)$
Measurement uncertainity X (σ^2)	9
Measurement uncertainity Y (σ^2)	9
Measurement uncertainity Z (σ^2)	100
Process noise (WN) (σ^2)	25

The bounding boxes detected by the CNN were forwarded to the Zoom Controller, which, based on the detection, generated zoom commands which were sent to the camera to adjust the zoom level [26]. The main idea was to keep the size of the bounding box of the instrument constant as long as the tool was within a region of interest (ROI) defined as a circle of radius $r = 7$ cm, with the centre of the ROI matching the centre of the image plane, $C = (x_c, y_c)$. Once the instrument was outside the ROI (i.e. the borders of the image), the camera zoomed out to keep the instrument visible. Therefore, two states could be distinguished:

State 1 - the tool inside the ROI: when the distal region of the tool was inside a predefined region, the zoom module tried to keep the size of the diagonal of the bounding box, D_{des}, constant and equal to 350 pixels. In particular, the condition:

$$D_{des} - \Delta < D_{des} < D_{des} + \Delta \tag{4}$$

is satisfied such that when the diagonal of the bounding box, D_{des}, is lower than the desired range, $D_{des} + \Delta$ (with $\Delta = 50$ pixels), the camera zooms in and vice versa.

State 2 - the tool outside the ROI: when the distal region of the tool was in a forbidden region, the zoom module sent a zoom out command to preserve the tool within the camera FoV.

B. Controller Design

For the control strategy, the first step was to use the visual information to estimate the homogeneous transform of the object w.r.t. the camera $\boldsymbol{T_{obj}^{cam}}$. Then, a desired relative coordinate transformation between the object and the camera, $\boldsymbol{T_{obj}^{cam*}}$, was defined. Since the goal was to maintain the surgical instrument tool at the center of the image, the desired position was defined as follows:

$$\boldsymbol{T_{obj}^{cam*}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

The x and y positions of the camera had to be identical to the one of the object while no constraint was considered along the Z-axis. The actual position of the camera w.r.t. the robot reference frame was retrieved as:

$$\boldsymbol{T_{cam}^{base}} = \boldsymbol{T_{LL}^{base}} * \boldsymbol{T_{cam}^{LL}} \tag{6}$$

where $\boldsymbol{T_{LL}^{base}}$ is the position of the last (seventh) link w.r.t. the robot base obtained from the kinematic chain and $\boldsymbol{T_{cam}^{LL}}$ is the position of the camera w.r.t. the last link of the manipulator calculated from the eye-hand calibration procedure. The calibration involved capturing eight images of a static checkerboard model (13x9 squares with a side length of 20 mm) with the camera to determine its position using the camera's intrinsic parameters. At the same time, the position of the last link (LL) was recorded using the manipulator's kinematic chain. The fixed position of the checkerboard in the eight different poses allowed the calculation of the camera position relative to the last link to be simplified to solve the equation which was accomplished using the OpenCV library [27]:

$$\boldsymbol{AX = XB} \tag{7}$$

where X is the robot hand-to-eye transformation and A and B are the transformations between two selected camera and end effector positions, respectively. To estimate the desired position of the camera, the position of the object needed to be expressed w.r.t. robot's base reference frame:

$$\boldsymbol{T_{obj}^{base}} = \boldsymbol{T_{LL}^{base}} * \boldsymbol{T_{cam}^{LL}} * \boldsymbol{T_{obj}^{cam}} \tag{8}$$

The new desired position of the camera w.r.t. the robot reference frame was then defined as:

$$\boldsymbol{T_{cam*}^{base}} = \boldsymbol{T_{obj}^{base}} * (\boldsymbol{T_{obj}^{cam*}})^{-1} \tag{9}$$

All the defined transformations are illustrated in Fig. 4.

The operational space error was then formulated as:

$$\boldsymbol{e = T_{cam*}^{base} - T_{cam}^{base}} \tag{10}$$

where $\boldsymbol{T_{cam*}^{base}}$ represents the desired position of the camera while $\boldsymbol{T_{cam}^{base}}$ is the actual one w.r.t the manipulator's base. The control goal of reaching the desired position was expressed as an exponential decrease in error in the operating space:

$$\boldsymbol{\dot{e} = -Ke} \tag{11}$$

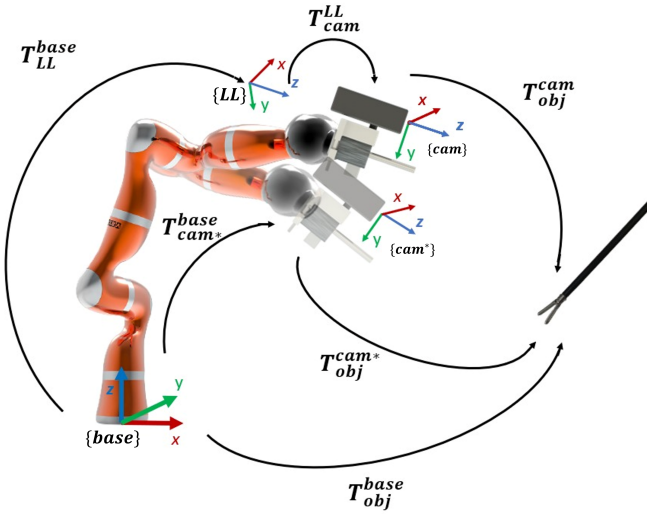

Fig. 4. Coordinate transformation to calculate the desired camera pose in the manipulator's reference frame.

where $\boldsymbol{K}$ is the gain that leads to the convergence of the error. Applying differential kinematics, the relation between motion in joint space and task space was defined as:

$$\boldsymbol{\dot{e} = -J_A(q)\dot{q}} \tag{12}$$

where $\mathbf{J_A}(q) \in \mathbb{R}^{6\times7}$ is the analytical Jacobian matrix of the robotic manipulator. By combining (10) and (11), the appropriate change in joint space was computed, given an error in space coordinates, as follows:

$$\boldsymbol{\dot{q} = J_{Apseudo}^{-1}(q)Ke} \tag{13}$$

where $\boldsymbol{J_{Apseudo}^{-1}}$ is the pseudoinverse of the Jacobian matrix $\boldsymbol{J_A(q)}$. The final desired joint values, $\boldsymbol{q_{des}}$, were computed by integration. These, together with the actual joint positions $\boldsymbol{q}$, were commanded through the FRI library to a low-level control that outputted the desired torque $\boldsymbol{\tau}$ to move the manipulator to the desired position.

IV. Experimental Setup

A 7-DoFs redundant lightweight robot (LWR4+, KUKA, Germany) was employed as a camera holder. The redundant DoF of the robotic manipulator allowed for increased manipulability and obstacle and singularities avoidance capability. The robot was used with a stereo camera mounted in an eye-in-hand configuration through a 3D printed mount. The JVC GS-TD1 Full HD 3D Camcorder was used as a vision sensor in this study. The camera was equipped with two 1/4.1" cm OS sensors and Twin HD GT lenses with an intra-axial distance of 35 mm, which allowed for the capture of two different points of view of the scene being recorded, thus offering 3D capabilities to the camera. Additionally, it offered full 1920x1080p high definition capabilities and various zoom modes in both formats (2D and 3D). It had a focal length ranging from 3.76 to 18.8 mm, providing a maximum of 5x zoom in 3D mode. In this paper, a maximum zoom level of 2x was used, considering the operational space, which corresponds to a displacement of 15 cm along the vertical

direction (Z-axis) of the robotic manipulator. An infrared transmitter connected to an Arduino Uno board was used to send the zoom command to the camera.

The dataset for training the CNN contained a total number of 5900 images. Among them, 4100 were recorded and manually annotated, while the remaining were extracted from the 2017 EndoVis challenge [28]. The final dataset was split into 5300 images for the training and 600 images for the validation test. Data augmentation was performed on the training dataset as rotations, translations, changes in brightness, and left-right flips. The network was trained on an Intel Xeon with a 12Gb Nvidia Titan X GPU for 400 epochs with a learning rate set to 0.001. A mini-batch size of eight images was used. The hyper-parameter Intersection over Union (IoU) representing how much the predicted bounding box overlapped with the ground truth was set to 0.5, meaning that a predicted bounding box overlap of more than 50 % with the ground truth was considered a true positive (TP). Otherwise, a false positive (FP) was considered. Finally, the confidence threshold was set to 0.4 representing the value above which a prediction was considered to be valid.

As the system was able to continuously track the CDR of the surgical instrument, a foot pedal was introduced such that the autonomous tracking was activated only when the pedal was pressed. This was done to prevent instrument tracking from becoming inconvenient for the surgeon in situations where tracking was not necessary, such as during instrument replacements. It also provided a safety precaution against uncontrolled movements by being activated only when the pedal was pressed.

A. System Characterisation

Each constitutive module of the proposed system (i.e. Detection Module, Position Reconstruction, and Control Strategy) was evaluated separately. Specific metrics were defined for each module to measure its respective performance.

Detection Module: Both the accuracy and inference time of the detection were examined in order to assess the performance of the detection module. The accuracy was evaluated using the average precision (AP) on the validation set. The AP was defined as the area under the precision-recall curve $p(r)$:

$$AP = \int_0^1 p(r)\, dr \tag{14}$$

As for the inference time, it was computed as the mean of the inference times on the validation set.

Position Reconstruction: Both the performance of the triangulation algorithm and the calibration procedure were considered in the evaluation of the 3D position reconstruction, as both dictate its precision. Eight images of a 13x9 calibration chessboard were acquired, and the corners of the chessboard were taken as reference points. The root mean square (RMS) re-projection error, $RMS_{re-projection}$, was used to evaluate the quality of the camera calibration and is defined as:

$$RMS_{re-projection} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2} \tag{15}$$

where ϵ_i represents the error on the image plane between the actual coordinates of the control points on the calibration target (chessboard corners) and the image coordinates reprojected using the calibration parameters. The performance of the triangulation algorithm was evaluated in terms of reconstruction error, E_{rec}, defined as the mean Euclidian distance between the reconstructed object points and the ground truth:

$$E_{rec} = \frac{1}{n}\sum_{i=1}^{n} |d_a - d_r| \tag{16}$$

where the ground truth, d_a, is defined as the real distance between two chessboard corners, equal to 20 mm, and d_r is the distance between the corners reconstructed by triangulation.

Control Strategy: To assess whether the system was able to track the CDR of the instrument under conditions of continuous change of direction and speed, the behavior of

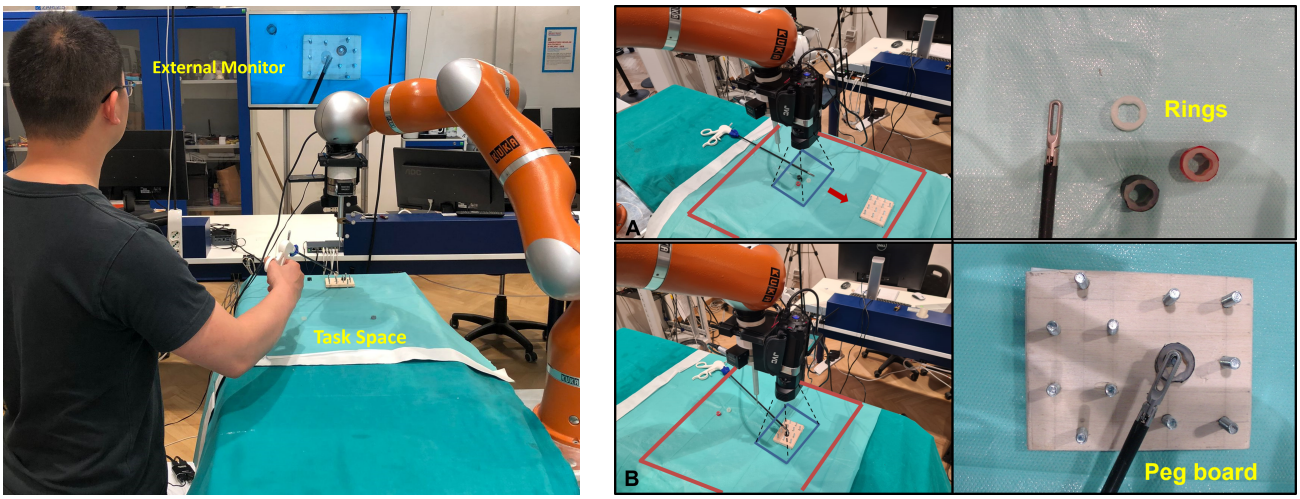

Fig. 5. Task scene (on the left): the user must execute the task by observing only the limited task space projected on the external monitor. Task description (on the right): (A) the task starts with the camera positioned in such a way that rings were seen. (B) shows the camera motion required to provide the view of the destination board. The red delimited area represents the work space.

the visual-servoing control was analysed. The instrument was moved in the XY plane along a circular trajectory with a diameter $d = 7.4$ cm at three different increasing velocities defined as slow, medium, and fast. Since the instrument was moved by a human user, maintaining a constant speed throughout the experiment proved difficult. As a result, the following velocity ranges were considered: velocities from 0 to 1 cm/s were defined as slow motion, from 1 to 1.5 cm/s as medium motion, and from 1.5 to 2.5 cm/s as fast motion. The experiment was performed in triplicate for each velocity level to evaluate the repeatability of the control strategy. To quantify the functionality of the system, the tracking error was used as a performance metric and was calculated as:

$$T_{err} = \sqrt{(X_d - X_a)^2 + (Y_d - Y_a)^2} \tag{17}$$

where (X_d, Y_d) is the CDR position of the tool and (X_a, Y_a) is the actual position of the EE in 3D space. The tracking error was considered negligible when the operational space error computed from the reconstructed position of the tool was less than 5 mm.

B. System Usability

A user study was carried out including both autonomous and joystick control modalities. The task designed for validation was a pick-and-place in which users were asked to use a surgical instrument to pick up four randomly distributed rings, one at a time, in a designed workspace and placed them on a target pegboard, as shown in Fig. 5(A). Users were asked to perform this task by observing the scene on an external monitor where the FoV of the camera was displayed. The task began with one of the four rings in sight, while the other three had to be found by the user by controlling the camera as described in Fig. 5 (B). However, at the beginning of the task, the user was allowed to observe the scene to get a general idea of the position of the objects within the task space. The distance in the Z direction between the camera and the task space was kept fixed at 0.2 m to provide a reduced FoV and force the user to move the camera to complete the task. The task space was limited to an area of 40x50 cm. The task was performed in two different modalities:

1) *Autonomous camera control* (*ACC*): the exoscope moved autonomously to try to keep the instrument inside the FoV. The user could activate and deactivate the motion of the holder by pressing a foot pedal.
2) *Joystick control* (*JC*): the user moved the exoscope using a foot-controlled joystick every time a different viewpoint was needed.

For each modality (autonomous or joystick control), each user performed three repetitions with each repetition ending when all four rings were placed on the target board. The user study was conducted on $S = 8$ non-medical subjects (aged between 23 and 29, five males and three females, and all right-handed). The experiments were made according to authorization number 16/2020. Informed consent was obtained from all participants in the study. The exoscopic system with the two control modalities was first shown to the participants who were then given three minutes to become familiar with the system.

The following objective metrics were considered:

- Mean execution time [s] defined as the sum of the execution time reported by all users for the three repetitions:

$$\bar{t}_{ex}(j) = \frac{1}{N_u} \sum_{i=1}^{N_u} t_{ex}(j)_i \tag{18}$$

  where j is the repetition number, $t_{ex}(j)_i$ represents the time required by the user i to complete the task in repetition j, and N_u is the number of the users.
- Total time outside the FoV [s] defined as the time the instrument was outside the camera FoV (hence, not visible to the user) normalised by the total execution time, t_{exi}:

$$t_{fov_{out}} = \frac{t_{fov_{\text{out}}\, i}}{t_{\text{exi}}} \tag{19}$$

- Instrument's position frequency representing the frequency distribution of the instrument in the image plane:

$$f_i = \frac{counts_i}{\max_i(counts_i)} \tag{20}$$

  For the computation of the frequency, the pixels of the image were partitioned into a grid of 10x10 cells (100 bins in total). $counts_i$ represented the number of times the distal region of the instrument fell inside the bin i of the image plane. The frequency within the bins was normalised by the maximum bin density.

Finally, a qualitative analysis was carried out using the NASA-TLX [29] survey. Users were asked to rank on a scale from 0 to 100, with steps of 5, the perceived workload while performing the task in either mode. A total of six scores were given to mental, physical, and temporal demand, and performance, effort, and frustration. The final score was calculated as the average of all scores. The Wilcoxon signed-rank test for paired samples was used to compare all the metrics, with statistical significance assessed at $p < 0.05$.

V. Results and Discussion

A. System Characterisation Results

The results obtained for each module are presented below. AP on the validation set reached a maximum value of 0.89 with a number of epoch $n = 360$. The inference time was 80 ± 10 ms resulting in a processing of 12.5 fps. The overall accuracy of the calibration in terms of RMS re-projection error resulted in a mean value equal to 0.71 pixels. As for the accuracy of the triangulation method, the reconstruction error resulted in a value of 1.31 ± 0.5 mm.

In the case of fast motion (2.01 ± 1.84 cm/s), the tracking error on the x- and y-axis, along with the position and velocity of the joints are shown in Fig. 6. A root mean square tracking error of 1.15 cm and 1.03 cm on the x- and y-axis, respectively, can be observed, with an overall tracking error of 1.54 ± 1.08 cm. Additionally, the proposed scheme was able to keep joint angles within the admissible range ($\pm 170°$ for joint θ_1, θ_3, θ_5 and θ_7 and $\pm 120°$ for the remaining joints [30]), and joint velocities within the limits (180°/s for θ_5 and 112.5°/s for the remaining joints) [31]. The tracking error computed in the three repetitions for the three different speed conditions

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2023.3258524

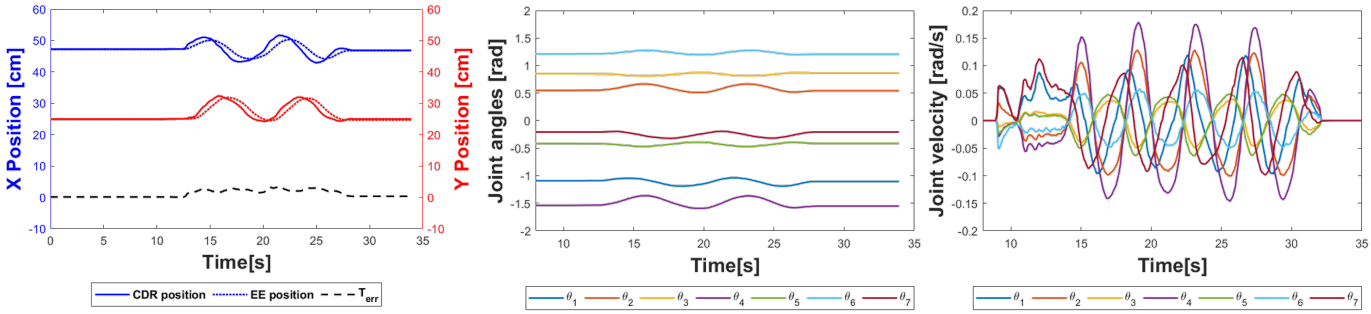

Fig. 6. Left: tracking error (dashed line). The solid line defines the position of the surgical instrument, the dotted line the EE position. In the center, joint angles; in the right, joint velocities during a fast instrument movement.

is shown in Fig. 7. A mean error with the standard deviation (SD) of 0.50 ± 0.17 cm, 0.90 ± 0.37 cm, and 1.38 ± 0.73 cm, representing a percentage error of 0.78 %, 1.4 % and 2.15 % in relation to the total available workspace, was found for slow, medium, and fast motion, respectively. For each of the three velocity conditions, the mean tracking error remains relatively constant, and SD is comparatively small. This shows that the system is repeatable and functional in keeping the CDR of the instrument within the FoV. However, a significant increase in the tracking error can be observed for high-velocity values (fast motion).

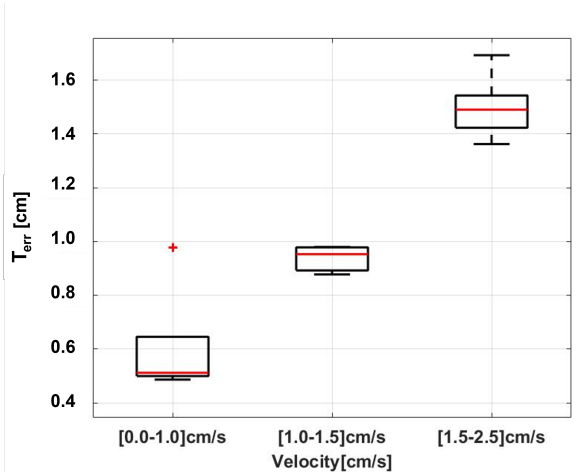

Fig. 7. Tracking error for the 3 repetitions of the experiment for 3 increasing velocities[cm/s]. From left to right $0 < v < 1$, $1 < v < 1.5$, $1.5 < v < 2.5$

B. System Usability Experiment Results

The mean and the SD of the execution time across all users for the three repetitions are illustrated in Fig. 8. A significant difference (p-value < 0.05) was found for each repetition between modalities. A mean value of 72.074 ± 19.36 s and 106.52 ± 18.50 s was observed in the third repetition for the autonomous and joystick modality, respectively. A trend can be observed for both modalities with the mean execution time decreasing from 111.5 s to 106 s and 83.6 s to 72 s in the *JC* and *ACC* control, respectively. However, the completion time for the joystick control remained constant between the second and third repetition, while it decreased in the autonomous control. This might indicate a steeper learning curve for the autonomous modality. With more repetitions of the experiment, a stronger conclusion can be drawn. The lower execution time that occurred with the autonomous control mode was due to the fact that the user was relieved of secondary tasks, such as repositioning the camera. In this way, the user could focus only on the main task and, as a result, achieve lower execution times. The time out of the FoV normalised by the

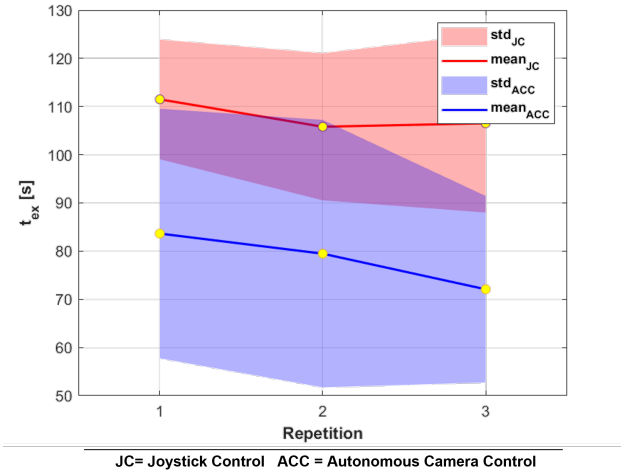

Fig. 8. Mean execution time and standard deviation for the three repetitions for the two tested modalities.

total execution time is shown in Fig. 9. For each repetition, the two distributions were found statistically different with p-value < 0.05. The time out of the FoV was significantly higher in the joystick-controlled camera compared with the autonomous one, with a mean value in the third repetition of 18.8 ± 13.50 % and 3.6 ± 4.5 %, respectively. This could be due to users' need to divide their attention between the execution of the task and the repositioning of the camera when using the joystick control mode. Each time the camera had to be moved, the user focused on the repositioning task, which caused the instrument to be outside the FoV. In a real surgical scenario, the surgeon's inability to see the instrument during surgery poses a risk to the patient's safety as unwanted contact between the instrument and delicate structures could occur.

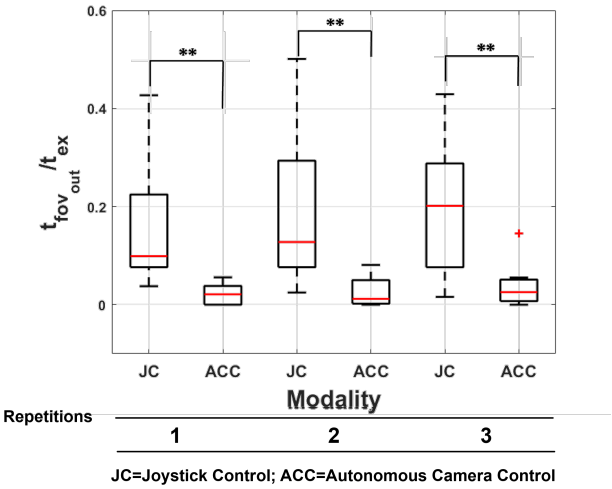

Fig. 9. Normalized time out of FoV for the three repetitions for the two tested modalities. (**, $p < 0.01$).

The frequency of the distal region of the tool in the right image for the three repetitions for all users is shown in Fig. 10. The frequency was normalised by the total number of pixels recorded in each distribution to have comparable values in the colormap. For autonomous control, a high tool frequency was observed in the centre of the image with a median value for all users during the three repetitions of 453 and 511 pixels on the x- and y-axis, respectively, while with the joystick-controlled exoscope, higher values were observed around the left corner of the image with a median of 232 and 695 pixels for the x- and y-axis, respectively. This means that the *ACC* is more effective in keeping the CDR of the instrument in the center of the image, providing a better view of the scene.

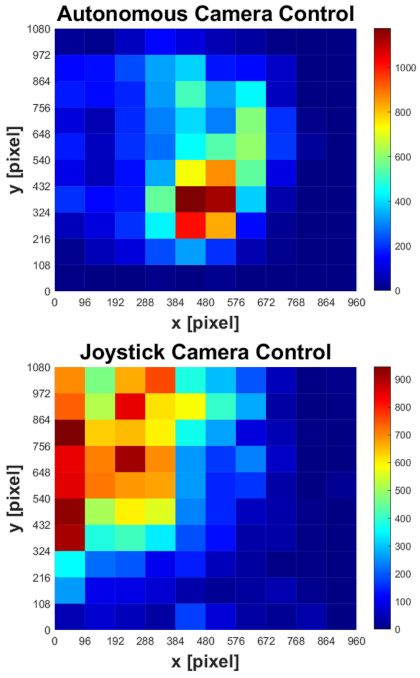

Fig. 10. Instrument's tip frequency inside the image plane normalized by the total number of pixels for the autonomous (upper) and joystick (lower) control modality. The color map ranging from blue to red shows the low and high frequency, respectively.

As for the qualitative analysis, subjective metrics have been extracted from the NASA-TLX survey. The users rated the task load perceived with the autonomous camera control modality much lower than the joystick control. The two groups resulted statistically different (p-value < 0.05) with an overall mean value of 32.5 ± 13.93 and 54.05 ± 12.84 for the autonomous and joystick control, respectively.

VI. Conclusion

This work proposes a framework for an autonomous vision-guided exoscope holder based on a visual-servoing technique. The main objective of the work was to test whether the proposed system could support the surgeon more effectively than traditional joystick control. From the initial analysis, the system has demonstrated its ability to detect and track the surgical instrument, ensuring that it remains in the centre of the FoV. The results of the user study showed that the system can improve the user experience by reducing completion time and physical and cognitive loads during tasks. In addition, it was demonstrated that the autonomous exoscope holder decreases the amount of time the instrument was outside the FoV which may decrease the risk of complications during surgery.

The proposed system could pave the way toward exoscope automation in neurosurgery. However, in its current form, some limitations are still present. It must be mentioned that the training data set for the object detection CNN is not representative of a clinical scenario, but it is used to conceptually demonstrate the current method. In addition, although the proposed markerless visual-servoing approach is capable of detecting multiple surgical instruments simultaneously, potential risks exist when sending multiple instrument positions to the robot controller. Consequently, future work will focus on the design of an optimal control strategy for multiple existing surgical instrument scenarios. Additionally, a dataset containing real clinical images comprising the specific artifacts present in neurosurgery should be utilised for training the neural network to obtain robust training results. The processing speed is far from being real-time causing a delay between the movement of the instrument and that of the exoscope holder. This is due to the limited hardware resources. Therefore, to exploit the potential of CNN and minimise the response delay of the control strategy as much as possible, it would be necessary to use high computing power in the future. In addition, it would be important to consider a task representative of the neurosurgical scenario, as the one proposed was performed to verify the functionality of the system. It would also be necessary to have medical subjects test the system. Finally, the introduction of multiple degrees of freedom (pan/tilt) in the camera motion should be considered, as this is a common motion in neurosurgery.

References

[1] N. Montemurro, A. Scerrati, L. Ricciardi, and G. Trevisi, "The Exoscope in Neurosurgery: An Overview of the Current Literature of Intraoperative Use in Brain and Spine Surgery," *Journal of Clinical Medicine*, vol. 11, no. 1, p. 223, 12 2021.

[2] L. Ricciardi, K. L. Chaichana, A. Cardia, V. Stifano, Z. Rossini, A. Olivi, and C. L. Sturiale, "The exoscope in neurosurgery: an innovative "point of view". a systematic review of the technical, surgical, and educational aspects," *World neurosurgery*, vol. 124, pp. 136–144, 2019.

[3] P. Kullar, R. Tanna, M. Ally, A. Vijendren, and G. Mochloulis, "Vitom 4k 3d exoscope: a preliminary experience in thyroid surgery," *Cureus*, vol. 13, no. 1, 2021.

[4] R. Berguer, J. Chen, and W. D. Smith, "A Comparison of the Physical Effort Required for Laparoscopic and Open Surgical Techniques," *Archives of Surgery*, vol. 138, no. 9, pp. 967–970, 09 2003.

[5] B. Fiani, R. Jarrah, F. Griepp, and J. Adukuzhiyil, "The role of 3d exoscope systems in neurosurgery: An optical innovation." *Cureus*, vol. 13, no. 6, 06 2021.

[6] P. Fiorini, K. Y. Goldberg, Y. Liu, and R. H. Taylor, "Concepts and trends in autonomy for robot-assisted surgery," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 993–1011, 2022.

[7] S. I. Ahmed, G. Javed, B. Mubeen, S. B. Bareeqa, H. Rasheed, A. Rehman, M. M. Phulpoto, S. S. Samar, and K. Aziz, "Robotics in neurosurgery: a literature review," *JPMA. The Journal of the Pakistan Medical Association*, vol. 68, no. 2, p. 258, 2018.

[8] D. J. Langer, T. G. White, M. Schulder, J. A. Boockvar, M. Labib, and M. T. Lawton, "Advances in intraoperative optics: a brief review of current exoscope platforms," *Operative Neurosurgery*, vol. 19, no. 1, pp. 84–93, 2020.

[9] A. Sivananthan, A. Kogkas, B. Glover, A. Darzi, G. Mylonas, and N. Patel, "A novel gaze-controlled flexible robotized endoscope; preliminary trial and report," *Surgical Endoscopy*, vol. 35, no. 8, pp. 4890–4899, 08 2021.

[10] A. Pandya, L. A. Reisner, B. King, N. Lucas, A. Composto, M. Klein, and R. D. Ellis, "A review of camera viewpoint automation in robotic and laparoscopic surgery," *Robotics*, vol. 3, no. 3, pp. 310–329, 2014.

[11] M. Takahashi, M. Takahashi, N. Nishinari, J. Matsuya, T. Tosha, Y. Minagawa, O. Shimooki, and T. Abe, "Clinical evaluation of complete solo surgery with the "viky®" robotic laparoscope manipulator," *Surgical Endoscopy*, vol. 31, no. 2, pp. 981–986, 2016.

[12] L. Qiu, C. Li, and H. Ren, "Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 159–164, 12 2019.

[13] T. Da Col, A. Mariani, A. Deguet, A. Menciassi, P. Kazanzides, and E. De Momi, "Scan: System for camera autonomous navigation in robotic-assisted surgery," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 2996–3002.

[14] T. Da Col, G. Caccianiga, M. Catellani, A. Mariani, M. Ferro, G. Cordima, E. De Momi, G. Ferrigno, and O. de Cobelli, "Automating endoscope motion in robotic surgery: A usability study on da vinci-assisted ex vivo neobladder reconstruction," *Frontiers in Robotics and AI*, vol. 8, 2021.

[15] M. Chmarra, C. Grimbergen, and J. Dankelman, "Systems for tracking minimally invasive surgical instruments," *Minimally Invasive Therapy & Allied Technologies*, vol. 16, pp. 328–40, 02 2007.

[16] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten, "Robotic endoscope control via autonomous instrument tracking," *Frontiers in Robotics and AI*, vol. 9, 2022.

[17] C. Molnár, T. D. Nagy, R. N. Elek, and T. Haidegger, "Visual servoing-based camera control for the da vinci surgical system," in *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 2020, pp. 107–112.

[18] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[19] F. Chaumette and S. Hutchinson, "Visual servoing and visual tracking," 2008.

[20] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten, "Robotic endoscope control via autonomous instrument tracking," *Frontiers in Robotics and AI*, vol. 9, 2022.

[21] B. C. Becker, V. Sandrine, R. A. MacLachlan, G. D. Hager, and C. N. Riviere, "Active guidance of a handheld micromanipulator using visual servoing," *IEEE International Conference on Robotics and Automation*, pp. 339–344, 2009.

[22] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten, "Robotic endoscope control via autonomous instrument tracking," *Frontiers in Robotics and AI*, vol. 9, 2022.

[23] R. Joseph and F. Ali, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[24] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[25] J. Fu, M. Poletti, Q. Liu, E. Iovene, H. Su, G. Ferrigno, and E. De Momi, "Teleoperation control of an underactuated bionic hand: Comparison between wearable and vision-tracking-based methods," *Robotics*, vol. 11, no. 3, p. 61, 2022.

[26] T. Sutjaritvorakul, A. Nejadfard, A. Vierling, and K. Berns, "Adaptive zoom control approach of load-view crane camera for worker detection," in *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC)*. Dubai, UAE: International Association for Automation and Robotics in Construction (IAARC), November 2021, pp. 553–560.

[27] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[28] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, L. Herrera, W. Li, V. Iglovikov, H. Luo, J. Yang, D. Stoyanov, L. Maier-Hein, S. Speidel, and M. Azizian, "2017 robotic instrument segmentation challenge," 2019.

[29] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," *Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, 2006.

[30] C. Gaz, F. Flacco, and A. De Luca, "Identifying the dynamic model used by the kuka lwr: A reverse engineering approach," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1386–1392.

[31] Y. Zhang and S. Li, "A neural controller for image-based visual servoing of manipulators with physical constraints," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5419–5429, 2018.

Elisa Iovene received her Master Degree in Biomedical Engineering - Technologies for Electronics from Politecnico di Milano in April 2021. She is currently a PhD candidate at the Department of Electronics, Information and Bioengineering (DEIB) of the Politecnico di Milano.

Alessandro Casella received the M.Sc. degree in Biomedical Engineering at Politecnico di Milano, Milan, Italy. He is currently pursuing a PhD in Computer Vision and Artificial Intelligence at the Advanced Robotic Laboratory of Istituto Italiano di Tecnologia, Italy.

Alice Valeria Iordache received her M.Sc. Degree in Biomedical Engineering - Technologies for Electronics from Politecnico di Milano. She is currently a Software Engineer at Asensus Surgical.

Junling Fu received the B.E. degree in Mechanical Engineering from Guangzhou University and M.S. degree in Mechanical Engineering from South China University of Technology, Guangzhou, China, in 2017 and 2020, respectively. Currently, he is pursuing a Ph.D. degree in the Department of Electronics, Information and Bioengineering (DEIB) of Politecnico di Milano, Italy. He is a member of the Neuroengineering and Medical Robotics Laboratory (NearLab).

Prof. Federico Pessina, M.D: received his specialisation in Neurosurgery at the University of Insubria in Varese, Italy, From 2007 to 2008, he was Clinical Fellow at the Hopital Universitaire in Lausanne, Switzerland. In 2018 he was Skull Base fellow at the Hopital Lariboisiere in Paris. Since 2022 he has been a Full Professor in Neurosurgery at Humanitas University, Italy.

Marco Riva, M.D: has perfected his expertise at the National Institute of Neurological Disorders and Stroke (NINDS) and the Department of Neurological Survey at the University of California San Francisco (UCSF). He is currently a surgeon specialising in Neurosurgery and a tutor at the graduate school of neurosurgery at the University of Milan, Italy.

Prof. Giancarlo Ferrigno, Ph.D. received the M.Sc. degree in electrical engineering and the Ph.D. degree in bioengineering from the Politecnico di Milano, Milan, Italy. He is a Full Professor of Electronics and Information Bioengineering and the Founder of the Neuroengineering and Medical Robotics Laboratory with the Department of Electronics, Information and Bioengineering, Politecnico di Milano.

Prof. Elena De Momi, Ph.D. received her M.Sc. and Ph.D. degrees in biomedical engineering from the Politecnico di Milano, Milan, Italy. She is currently an Assistant Professor in the Department of Electronics, Information, and Bioengineering, Politecnico di Milano.