DOI:
10.3389/frobt.2022.832208

Document Version
Peer reviewed version

frontiers

Robotic Endoscope Control via Autonomous Instrument Tracking

Caspar Gruijthuijsen [1†], **Luis C. Garcia-Peraza-Herrera** [2,4*†], **Gianni Borghesan** [1], **Dominiek Reynaerts** [1], **Jan Deprest** [3], **Sebastien Ourselin** [4], **Tom Vercauteren** [4], **and Emmanuel Vander Poorten** [1]

[1] *Department of Mechanical Engineering, KU Leuven, Leuven, Belgium*
[2] *Department of Medical Physics and Biomedical Engineering, University College London, London, United Kingdom*
[3] *Department of Development and Regeneration, Division Woman and Child, KU Leuven, Leuven, Belgium*
[4] *Department of Surgical & Interventional Engineering, King's College London, London, United Kingdom*

Correspondence*:
9th Floor, Becket House, 1 Lambeth Palace Road, London, SE1 7EU
luis_c.garcia_peraza_herrera@kcl.ac.uk

ABSTRACT

Many keyhole interventions rely on bi-manual handling of surgical instruments, forcing the main surgeon to rely on a second surgeon to act as a camera assistant. In addition to the burden of excessively involving surgical staff, this may lead to reduced image stability, increased task completion time and sometimes errors due to the monotony of the task. Robotic endoscope holders, controlled by a set of basic instructions, have been proposed as an alternative, but their unnatural handling may increase the cognitive load of the (solo) surgeon, which hinders their clinical acceptance. More seamless integration in the surgical workflow would be achieved if robotic endoscope holders collaborated with the operating surgeon via semantically rich instructions that closely resemble instructions that would otherwise be issued to a human camera assistant, such as "focus on my right-hand instrument". As a proof of concept, this paper presents a novel system that paves the way towards a synergistic interaction between surgeons and robotic endoscope holders. The proposed platform allows the surgeon to perform a bimanual coordination and navigation task, while a robotic arm autonomously performs the endoscope positioning tasks. Within our system, we propose a novel tooltip localization method based on surgical tool segmentation and a novel visual servoing approach that ensures smooth and appropriate motion of the endoscope camera. We validate our vision pipeline and run a user study of this system. The clinical relevance of the study is ensured through the use of a laparoscopic exercise validated by the European Academy of Gynaecological Surgery which involves bi-manual coordination and navigation. Successful application of our proposed system provides a promising starting point towards broader clinical adoption of robotic endoscope holders.

Keywords: Minimally Invasive Surgery, Endoscope Holders, Endoscope Robots, Endoscope Control, Visual Servoing

† These authors have contributed equally to this work and share first authorship.

1 INTRODUCTION

In recent years, many surgical procedures shifted from open surgery to minimally invasive surgery (MIS). Although MIS offers excellent advantages for the patient, including reduced scarring and faster recovery, it comes with challenges for the surgical team. Most notable is the loss of direct view onto the surgical site. In keyhole surgery, the surgeon manipulates long and slender instruments introduced into the patient through small incisions or keyholes. The surgeon relies on endoscopes, also long and slender instruments equipped with a camera and light source, to obtain visual feedback on the scene and the relative pose of the other instruments. The limited field of view (FoV) and depth of field of the endoscope urge an efficient endoscope manipulation method that allows covering all the important features and hereto optimizes the view at all times.

In typical MIS, surgeons cannot manipulate the endoscope themselves as their hands are occupied with other instruments. Therefore, a camera assistant, typically another surgeon takes charge of handling the endoscope. Human camera assistants have a number of shortcomings. An important drawback relates to the cost of the human camera assistant (Stott et al., 2017). Arguably, highly trained clinicians could better be assigned to other surgical duties that require the full extent of their skill set (as opposed to mainly manipulating the endoscope). If made widely feasible, solo MIS surgery would improve cost-effectiveness and staffing efficiency. An additional source of weakness related to human camera assistants is the ergonomic burden associated with assisting in MIS (Wauben et al., 2006; Lee et al., 2009). This may lead to reduced image stability, fatigue, distractions, increased task completion times, and erroneous involuntary movements (Goodell et al., 2006; Platte et al., 2019; Rodrigues Armijo et al., 2020). This problem aggravates for long interventions or when the assistant has to adopt particularly uncomfortable postures. Besides the ergonomic challenges, miscommunication between the surgeon and the assistant may lead to sub-optimal views (Amin et al., 2020).

In order to help or bypass the human camera assistant and to optimize image stability, numerous endoscope holders have been designed in the past (Jaspers et al., 2004; Bihlmaier, 2016; Takahashi, 2020). One can distinguish passive endoscope holders and active or robotic endoscope holders. Passive endoscope holders are mechanical devices that lock the endoscope in a given position until manually unlocked and adjusted. A problem common to passive endoscope holders is that they result in an intermittent operation that interferes with the manipulation task (Jaspers et al., 2004). When surgeons want to adjust the endoscopic view themselves, they will have to free one or both hands to reposition the endoscope. To counter this problem, robotic endoscope holders have been developed. These motorized devices offer the surgeon a dedicated interface to control the endoscope pose. Well-designed robotic endoscope holders do not cause additional fatigue, improve image stability, and increase ergonomics (Fujii et al., 2018). Also, hand-eye coordination issues may be avoided. Overall such robotic endoscope holders may lower the cognitive load of the surgeon and reduce operating room (OR) staff time and intervention cost (Ali et al., 2018). However, despite these advantages and the number of systems available, robotic endoscope holders have not found widespread clinical acceptance (Bihlmaier, 2016). This has been linked to the suboptimal nature of the human interface and consequently the discomfort caused to the surgeon by the increased cognitive load needed to control the camera. Popular robotic endoscope holders use foot pedals, joysticks, voice control, gaze control, and head movements (Kommu et al., 2007; Holländer et al., 2014; Fujii et al., 2018). The context switching between surgical manipulation and these camera control mechanisms seems to hinder the ability of the surgeon to concentrate on the main surgical task (Bihlmaier, 2016).

1.1 Contributions

In this work, we introduce the framework of *semantically rich endoscope control*, which is our proposal on how robotic endoscope control could be implemented to mitigate interruptions and maximize the clinical acceptance of robotic endoscope holders. We claim that *semantically rich instructions* that relate to the instruments such as "focus on the right/left instrument" and "focus on a point between the instruments" are a priority, as they are shared among a large number of surgical procedures. Therefore, we present a novel system that paves the way towards a synergistic interaction between surgeons and robotic endoscope holders. To the best of our knowledge, we are the first to report how to construct an autonomous instrument tracking system that allows for solo-surgery using only the endoscope as a sensor to track the surgical tools. The proposed platform allows the surgeon to perform a bi-manual coordination and navigation tasks while the robotic arm autonomously performs the endoscope positioning.

Within our proposed platform, we introduce a novel tooltip localization method based on a hybrid mixture of deep learning and classical computer vision. In contrast to other tool localization methods in the literature, the proposed approach does not require manual annotations of the tooltips, but relies on tool segmentation, which is advantageous as the manual annotation effort could be trivially waived employing methods such as that recently proposed in Garcia-Peraza-Herrera et al. (2021). This vision pipeline was individually validated and the proposed tooltip localization method was able to detect tips in 84.46% of the frames. This performance proved sufficient to allow for a successful autonomous guidance of the endoscope (per user study of the whole robotic system).

We propose a novel visual servoing method for a generalized endoscope model with support for both remote center of motion and endoscope bending. We show that a hybrid of position-based visual servoing (PBVS) and 3D image-based visual-servoing (IBVS) is preferred for robotic endoscope control.

We run a user study of the whole robotic system on a standardized bi-manual coordination and navigation laparoscopic task accredited for surgical training (European Academy of Gynaecological Surgery, 2020). In this study we show that the combination of novel tool localization and visual servoing proposed is robust enough to allow for the successful autonomous control of the endoscope. During the user study experiments (8 people, 5 trials), participants were able to complete the bi-manual coordination surgical task without the aid of a camera assistant and in a reasonable time ($172\,$s on average).

1.2 Towards semantically rich robotic endoscope control

While solo surgery has been demonstrated with simple robotic endoscope control approaches (Takahashi, 2020), we argue that to overcome the usability issues that impede broad clinical adoption of robotic endoscope holders and move towards solo surgery, robotic endoscope control should be performed at the task autonomy level. To efficiently operate in this setting, a robotic endoscope holder should accept a set of *semantically rich instructions*. These instructions correspond to the commands that a surgeon would normally issue to a human camera assistant. This contrasts with earlier approaches, where the very limited instruction sets (up, down, left, right, zoom in, zoom out) lead to a semantic gap between the robotic endoscopic holder and the surgeon (Kunze et al., 2011). With semantically rich instructions, it would be possible to bridge this gap and restore the familiar relationship between the surgeon and the (now tireless and precise) camera assistant.

A semantically rich instruction set should contain commands that induce context-aware actions. Examples of such are "zoom in on the last suture", "hold the camera stationary above the liver", and "focus the camera on my right instrument". When these instructions are autonomously executed by a robotic endoscope holder,

we refer to the control as *semantically rich robotic endoscope control*. We believe that semantically rich robotic endoscope control can effectively overcome the problem of intermittent operation with endoscopic holders, does not disrupt the established surgical workflow, ensures minimal overhead for the surgeon, and overall maximizes the usability and efficiency of the intervention.

Although instructions that have the camera track an anatomical feature are relevant, autonomous instrument tracking instructions (e.g. "focus the camera between the instruments") play a prominent role, as they are common to a large number of laparoscopic procedures and form a fundamental step towards solo surgery. Therefore, in this work we focus on semantically rich instructions related to the autonomous instrument tracking (AIT) of a maximum of two endoscopic instruments (one per hand of the operating surgeon, see Fig. 1). Particularly, the proposed method implements the instructions "focus on the right/left instrument" and "focus on a point between the instruments". User interface methods to translate requests expressed by the surgeon (e.g. voice control) to these AIT instructions fall outside the scope of this work.

The remainder of the paper is organized as follows. After describing the related work, the AIT problem is stated in Sec. 3. The quality of the AIT depends on robust methods to localize one or more surgical instruments in the endoscopic view. Sec. 4 describes a novel image-processing pipeline that was developed to tackle this problem. Visual servoing methods are described in Sec. 5. These methods provide the robotic endoscope control with the ability to track the detected instruments autonomously. An experimental user study campaign is set up and described in Sec. 6 to demonstrate the value of AIT in a validated surgical training task. Sec. 7 discusses the obtained results and Sec. 8 draws conclusions regarding the implementation of the AIT instructions proposed in this work.

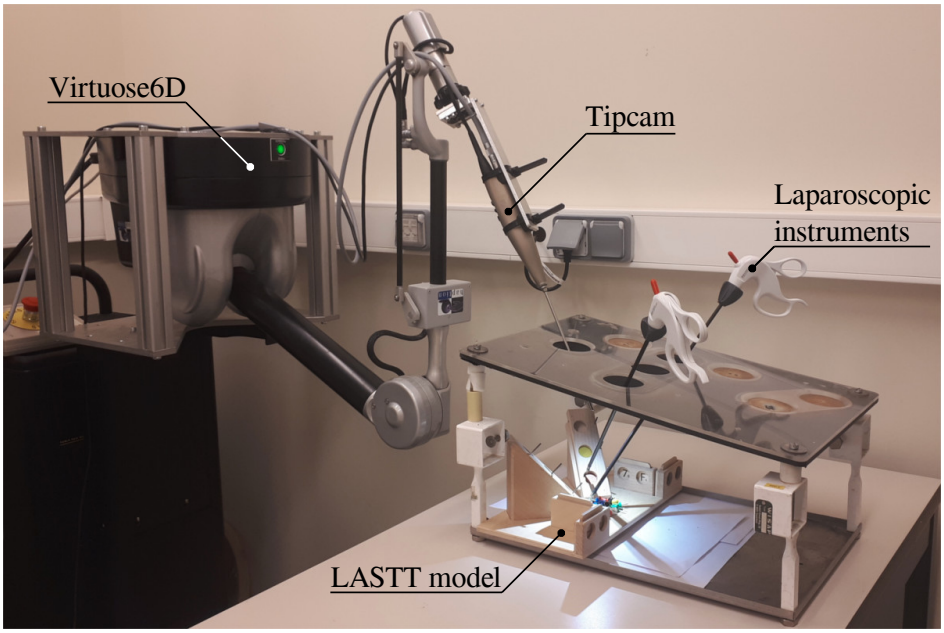

Figure 1. Proposed experimental setup for autonomous endoscope control in a laparoscopic setting. The LASTT model (European Academy of Gynaecological Surgery, 2020) showcased within the box trainer is designed for the simulation of simple surgical tasks with a focus on gynaecological procedures. It is common ground for the practice and evaluation of hand-eye and bi-manual coordination skills.

2 RELATED WORK

Robotic endoscope control (REC) allows the surgeon to control the endoscope without having to free their hands. A wide variety of dedicated control interfaces have been developed and commercialized for this purpose, including joystick control, voice control, gaze control and head gesture control (Taniguchi et al., 2010). Despite the apparent differences in these interfaces, established approaches offer the surgeon a basic instruction set to control the endoscope. This set typically consists of six instructions: zoom in/out, move up/down, and move left/right. The basic nature of these instructions makes detailed positioning cumbersome, usually resorting to a lengthy list of instructions. This shifts the surgeon's focus from handling the surgical instruments towards the positioning of the endoscope, as concurrent execution of those actions is often strenuous and confusing (Jaspers et al., 2004). Moreover, the increased mental workload stemming from simultaneous control of the instruments and the endoscope might adversely affect intervention outcomes (Bihlmaier, 2016). Similar to passive endoscope holders, robotic endoscope holders that offer a basic set of instructions prevent fluid operation and lead to intermittent action.

A number of REC approaches that pursue fully autonomous operation have been proposed as well. A starting point for an autonomous REC strategy is to reposition the endoscope so as to keep the surgical instrument tip centered in the view. Such an approach could work already when only the 2D position of the instrument in the endoscopic image is available (Uecker et al., 1995; Osa et al., 2010; Agustinos et al., 2014; Zinchenko and Song, 2021). In this case, the endoscope zoom level (depth) is left uncontrolled. Some of these methods also require a 3D geometrical instrument model (Agustinos et al., 2014), limiting the flexibility of the system. Approaches such as proposed by Zinchenko and Song (2021) have also suggested to replace the endoscope screen with a head-mounted virtual reality device that facilitates the estimation of the surgeon's attention focus from the headset's gyroscope. In this scenario, the autonomous REC strategy aims to reposition the endoscope with the aim of maintaining the weighted center of mass between the instruments' contour centers and the point of focus in the center of the view. However, it has been shown in works such as (Hanna et al., 1997; Nishikawa et al., 2008) that the zoom level is important for effective endoscope positioning. Other authors tried to circumvent the lack of depth information in 2D endoscopic images by relating the inter-instrument distance to the zoom level (Song et al., 2012; King et al., 2013). This approach is obviously limited to situations where at least two instruments are visible.

When the 3D instrument tip position is available, smarter autonomous REC strategies are possible. In the context of fully robotic surgery, kinematic-based tooltip position information has been used to provide autonomously guided ultrasound imaging with corresponding augmented reality display for the surgeon (Samei et al., 2020). Kinematics have also been employed by Mariani and Da Col *et al.* (Mariani et al., 2020; Da Col et al., 2021) for autonomous endoscope guidance in a user study on *ex vivo* bladder reconstruction with the da Vinci Surgical System. In their experimental setup, the system could track either a single instrument or the midpoint between two tools. Similarly, Avellino et al. (2020) have also employed kinematics for autonomous endoscope guidance in a co-manipulation scenario [1]. In (Casals et al., 1996; Mudunuri, 2010), rule-based strategies switch the control mode between single-instrument tracking or tracking points that aggregate locations of all visible instruments. Pandya *et al.* argued that such schemes are reactive and that better results can be obtained with predictive schemes, which incorporate knowledge of the surgery and the surgical phase (Pandya et al., 2014). Examples of such knowledge-based methods are (Wang et al., 1998; Kwon et al., 2008; Weede et al., 2011; Rivas-Blanco et al., 2014; Bihlmaier, 2016; Wagner et al., 2021). While promising in theory, in practice, the effort to create complete and reliable

[1] https://www.youtube.com/watch?v=R1qwKAWFOIk

models for an entire surgery is excessive for current surgical data science systems. In addition, accurate and highly robust surgical phase recognition algorithms are required, increasing the complexity of this solution considerably.

With regards to the levels of autonomy in robotic surgery, Yang et al. (2017) have recently highlighted that the above strategies aim for very high autonomy levels but take no advantage of the surgeon's presence. In essence, the surgeon is left with an empty instruction set to direct the endoscope holder. Besides being hard to implement given the current state of the art, such high autonomy levels may be impractical and hard to transfer to clinical practice. Effectively, an ideal camera assistant only functions at the task autonomy level. This is also in line with the recent study by Col et al. (2020), who concluded that it is important for endoscope control tasks to find the right trade-off between user control and autonomy.

To facilitate the autonomous endoscope guidance for laparoscopic applications when the 3D instrument tip position is not available, some authors have proposed to attach different types of markers to the instruments (e.g. optical, electromagnetic). This modification often comes with extra sensing equipment that needs to be added to the operating room.

In Song and Chen (2012), authors proposed to use a monocular webcam mounted on a robotic pan-tilt platform to track two laparoscopic instruments with two colored rings attached to each instrument. They employed the estimated 2D image coordinates of the fiducial markers to control all the degrees of freedom of the robotic platform. However, this image-based visual servoing is not able to attain a desired constant depth to the target tissue (as also shown in our simulation of image-based visual servoing in Sec. 5.3). In addition, the choice of fiducial markers is also an issue. Over the years, going back at least as far as to (Uenohara and Kanade, 1995), many types of markers have been proposed by the community for tool tracking purposes. For example, straight lines (Casals et al., 1996), black stripes (Zhang and Payandeh, 2002), cyan rings (Tonet et al., 2007), green stripes (Reiter et al., 2011), multiple colour rings for multiple instruments (blue-orange, blue-yellow) (Seong-Young Ko et al., 2005), and multiple colour (red, yellow, cyan and green) bio-compatible markers (Bouarfa et al., 2012). However, although fiducial markers such as colored rings ease the tracking of surgical instruments, attaching or coating surgical instruments with fiducial markers presents serious sterilization, legal and installation challenges (Stoyanov, 2012; Bouget et al., 2017). First, the vision system requires specific tools to work or a modification of the current ones, which introduces a challenge for clinical translation. At the same time, computational methods designed to work with fiducials cannot easily be trained with standard previously recorded interventions. Additionally, to be used in human experiments, the markers need to be robust to the sterilisation process (e.g. autoclave). This poses a manufacturing challenge and increases the cost of the instruments. The positioning of the markers is also challenging. If they are too close to the tip, they might be occluded by the tissue being manipulated. If they are placed back in the shaft, they might be hidden to the camera, as surgeons tend to place the endoscope close to the operating point. Even if they are optimally positioned, fiducials may be easily covered by blood, smoke, or pieces of tissue. In addition to occlusions, illumination (reflections, shadows) and viewpoint changes still remain a challenge for the detection of the fiducial markers.

In contrast to using colored markers, Sandoval et al. (2021) used a motion capture system (MoCap) in the operating room to help the autonomous instrument tracking. The MoCap consisted of an exteroceptive sensor composed of 8 high resolution infrared cameras. This system was able to provide the position of the reflective markers placed at the instruments (4 markers per instrument) in real time. However, the MoCap increases considerably the cost of the proposed system and complicates the surgical workflow. Instruments need to be modified to add the markers, and the MoCap needs to be installed in the operating room. As any other optical tracking system, it also possesses the risk of occlusions in the line of sight, making it

impossible for the system to track the instruments when such occlusions occur. As opposed to all these different markers, the endoscope is necessary to perform the surgery, and the surgeon needs to be able to see the instruments to carry out the intervention, so using the endoscope and its video frames without any instrument modifications to help track the tools is a solution that stems naturally from the existing surgical workflow. This has also been the path followed in the devise of AUTOLAP™ (Medical Surgery Technologies, Yokneam, Israel) (Wijsman et al., 2018, 2022), which is, to the best of our knowledge, the only robotic laparoscopic camera holder that claims to have incorporated image-based laparoscopic camera steering within its features. However, no technical details are provided in these publications on how it is achieved.

3 AUTONOMOUS INSTRUMENT TRACKING

In a typical surgical scenario, a surgeon manipulates two instruments: one in the dominant and one in the non-dominant hand. In such a case, the surgeon might want to focus the camera on one specific instrument, or center the view on a point in between the instruments, depending on their relative importance. AIT strives to automate these tasks, as explained next.

3.1 Centering instrument tips in FoV

With one instrument present, the proposed AIT objective is to center the instrument tip position $\boldsymbol{s}$ in the FoV, as is illustrated in Fig. 2 (top). With two visible instruments, a relative dominance factor $w_d \in [0, 1]$ can be assigned to the instruments (adjustable via a semantically rich instruction "change dominance factor X% to the right/left"). The AIT controller can then track the virtual average tip position according to

$$\boldsymbol{s} = (1 - w_d)\boldsymbol{s}_l + w_d\boldsymbol{s}_r, \tag{1}$$

where $\boldsymbol{s}_l$ and $\boldsymbol{s}_r$ are the respective tip positions of the left and right instrument as visualized in Fig. 2, bottom.

If the AIT were implemented to continuously track the virtual tip position $\boldsymbol{s}$, the view would never come to a standstill, which would be disturbing for the surgeon. As a solution, also suggested in (Bihlmaier, 2016) and Eslamian et al. (2020), a position hysteresis behaviour can be implemented. In this work, a combination of instructions with position hysteresis is implemented based on three zones in the endoscope FoV. As illustrated in Fig. 2, target zone A captures the ideal location of the tooltip, and transition zone B represents a tolerated region. Entering a violation zone C triggers re-positioning of the endoscope. Whenever $\boldsymbol{s}$ moves from zone B to zone C, the AIT will be activated. It will then stay active until $\boldsymbol{s}$ reaches zone A. Afterwards, the FoV will be kept stable, until $\boldsymbol{s}$ again crosses the border between zone B and zone C.

This implementation of AIT offers the surgeon instructions to track either instrument, to change the dominance factor, or to stop the tracking by disabling the AIT. Note that this implementation of AIT only specifies two degrees of freedom (DoFs) out of the four available DoFs in typical laparoscopy. The depth DoF is controlled by an additional instruction for the zoom level, i.e., the distance between the camera and the instrument tip. The DoF that rolls the endoscope around its viewing axis is controlled to always enforce an intuitive horizontal orientation of the camera horizon. If desired, a semantically rich instruction could be added to alter this behaviour.

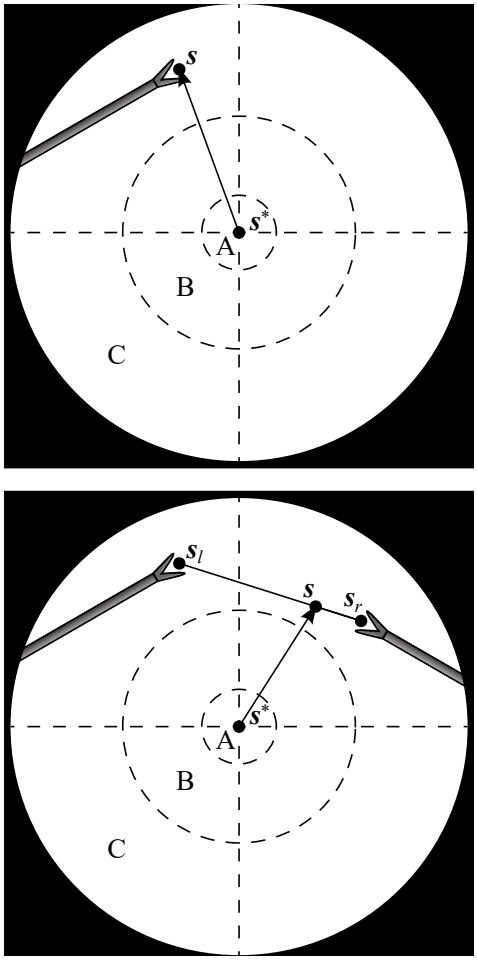

Figure 2. Endoscopic view with one instrument tip at position $\boldsymbol{s}$ (top), and two instrument tips at positions $\boldsymbol{s}_l$ and $\boldsymbol{s}_r$, combined to a virtual tip position $\boldsymbol{s}$ (bottom). The AIT controller aims to make $\boldsymbol{s}$ coincide with a desired position $\boldsymbol{s}^*$. A, B and C are respectively the target, transition and violation zones of a programmed position hysteresis approach.

3.2 Comanipulation fallback

As neither the set of AIT instructions nor any other set of instructions can realistically cover all instructions relevant for semantically rich REC, situations can arise in surgical practice where the capabilities of the robotic endoscope holder are insufficient. In such a case, it is necessary to switch to a comanipulation mode. This kind of switching is already the clinical reality for commercial robotic endoscope holders (Gillen et al., 2014; Holländer et al., 2014) and is particularly relevant when the system is used to support rather than replace the surgical assistant.

This work proposes to embed an easy switching functionality as a system feature. A natural transition from REC to comanipulation mode can be made possible through the use of a mechanically backdrivable robotic endoscope holder. This way, no extra hardware components are needed for switching, neither is it necessary to release the endoscope from the robotic endoscope holder. Instead, the surgeon can simply release one instrument, grab the endoscope and comanipulate it jointly with the robotic endoscope holder. During comanipulation, the human provides the intelligence behind the endoscope motions, while still experiencing support in the form of tremor-eliminating damping and fatigue-reducing gravity compensation. Such an approach broadens the scope of interventions where REC can be realistically applied.

4 MARKERLESS INSTRUMENT LOCALIZATION

REC based on semantically rich instructions requires the robotic endoscope holder to autonomously execute context-aware tasks. This implies a need to autonomously collect contextual information. The AIT instruction relies on knowledge of the tip position $\boldsymbol{s}$ of the surgical instruments in the endoscopic view. To obtain this information, without the need to alter the employed instruments or surgical workflow, a markerless instrument localization pipeline is developed in this section. Note that the term *localization* is employed here, instead of the commonly used term *tracking*, as for the sake of clarity this work reserves *tracking* for the robotic servoing approaches needed for AIT.

4.1 Instrument localization approaches

If, in addition to the endoscope, the instruments are also mounted on a robotic system (Eslamian et al., 2016; Weede et al., 2011) or if they are monitored by an external measurement system (Nishikawa et al., 2008; Polski et al., 2009), the position of the instruments can be directly obtained, provided that all involved systems are correctly registered and calibrated. However, in this work, manual handling of off-the-shelf laparoscopic instruments precludes access to such external localization information.

An alternative, which we use in this work, is to exploit the endoscope itself as the sensor. A review on this topic has been published relatively recently by Bouget et al. (2017). In their work Bouget et al. present a comprehensive survey of the last years of research in tool detection and tracking with a particular focus on methods proposed prior to the advent of the deep learning approaches. Recent instrument localization techniques based on Convolutional Neural Networks (CNN) (González et al., 2020; Pakhomov et al., 2020) are currently recognized as the state-of-the-art approaches (Allan et al., 2019; Roß et al., 2021) for such problems. In this work, we leverage our previous experience with CNN-based real-time tool segmentation networks (García-Peraza-Herrera et al., 2016; Garcia-Peraza-Herrera et al., 2017) and embed the segmentation in a stereo pipeline to estimate the location of the tooltips in 3D.

4.2 Instrument localization pipeline

A multi-step image processing pipeline was developed for markerless image-based instrument localization (see Fig. 3). As input, the pipeline takes the raw images from a stereo endoscope. As output, it provides the 3D tip positions of the visible instruments. The maximum number of instruments and tips per instrument are required as inputs. In the task performed in our user study, presented in Sec. 6), a maximum of two instruments with two tips may be present.

The 2D tooltip localization in image coordinates is a key intermediate step in this pipeline. Training a supervised bounding box detector for the tips could be a possible approach to perform the detection. However, to implement the semantically rich AIT presented in Sec. 3 and Fig. 2 we would still need to know whether the detected tips belong to the same or different instruments, and more precisely whether they belong to the instrument handled by the dominant or non-dominant hand. Therefore, we opted for estimating the more informative tool-background semantic segmentation instead. Via processing the segmentation prediction, we estimate how many instruments are in the image, localize the tips, and associate each tip with either the left or right-hand instrument. A downside of using semantic segmentation in comparison to a detector is the increased annotation time required to build a suitable training set. However, recent advances to reduce the number of contour annotations needed to achieve the segmentation such as (Vardazaryan et al., 2018; Fuentes-Hurtado et al., 2019; Garcia-Peraza-Herrera et al., 2021) greatly mitigate this drawback.

In the remaining of this section we first discuss the assumptions made, imaging hardware, and preprocessing steps. Then, we proceed to describe the localization pipeline. The localization method consists of the following steps: binary tool-background segmentation (Sec. 4.2.3), skeletonization of the segmentation mask (Sec. 4.2.4), graph extraction from the pixel-wide skeleton (Sec. 4.2.4), entrynode detection on the graph (Sec. 4.2.5), leaf node detection on the graph (Sec. 4.2.6), leaf node to entry node matching (Sec. 4.2.6), and left/right instrument identification (Sec. 4.2.8). After matching leaf nodes to entry nodes we have a subgraph for each instrument, and we distinguish between the left/right instrument using the estimated location of each instrument's entry node (Sec. 4.2.8).

The implementation of the whole localization pipeline was done in Python, reading the video feed from the framegrabber V4L2 device with OpenCV, and performing the deep learning inference with Caffe (Jia et al., 2014) on an NVIDIA GeForce GTX Titan X GPU.

4.2.1 Assumptions of proposed instrument localization pipeline

In our instrument localization pipeline, we assume that the instruments are not completely occluded. Partial occlusions are supported, as long as there is a visible path from the *entrypoint* to the tip of the instrument. Note that with *entrypoint* we refer to the point located at the edge of the endoscopic content area where the instrument enters the image. This point is not to be confused with the *incision point* which is the point on the body wall where the incision is made through which the instrument enters the patient's body. Now, if the tip is occluded, the tooltip will be estimated on the furthermost point of the shaft. When the entrypoint is completely covered, the instrument will not be detected in the current approach. Methods that exploit knowledge of the incision point could help in such a case (and could be explored in future work as they do not form the core of this work). The current limitations are illustrated in Fig. 4. The assumption that instruments have to enter from the edge serves two purposes, 1) as a noise reduction technique for the segmentation, because false positive *islands* of pixels can be easily discarded, and 2) to detect whether the instrument is held by the right/left hand of the surgeon (as explained in Sec. 4.2.8). In most cases, the entrypoint of at least one of the instruments will be visible. Therefore, the benefits of the assumption that instruments will not be completely occluded largely outweigh its limitations. The proposal of surgical tool segmentation models that are robust to entrypoint occlusions (Fig. 4, right) or complete occlusions is out of the scope of this work.

4.2.2 Imaging hardware and preprocessing procedure

The stereo camera and endoscopy module of choice for this work were the TIPCAM1 S 3D ORL (30° view on a 4 mm outer diameter shaft) and IMAGE1 S D3-LINK respectively (both from KARL STORZ, Germany). The DVI output of the endoscopy module is plugged into a DVI2PCIE DUO framegrabber (EPIPHAN, Canada). The endoscopy module produces images at 60 fps and at a resolution of 1920×1080 pixels, which turns into 1920×540 as each grabbed image contains a stereo pair with the left frame on even rows and the right frame on odd ones. These images are upsampled in the y-axis so that two images of 1920×1080 pixels are obtained. Upscaling is chosen (as opposed to downsampling to 960×540 pixels) to avoid degrading the depth resolution based on x-axis disparity. The left-right cameras are calibrated using a chessboard pattern of 1.1 mm-wide squares (Cognex Glass Calibration Plate Set 320-0015R, APPLIED IMAGE INC., NY, USA). Both frames, left and right, are rectified in real-time. Then, the black background of the images is cropped out, keeping just a square crop of the endoscopic circle content area (as shown in Fig. 2), which results in an image of 495×495 pixels. Finally, the image where the 2D tooltip localization is going to be performed (either the left or right frame can be chosen without loss of generality) is downsampled to 256×256 pixels to speed up the subsequent processing steps (i.e. segmentation, graph

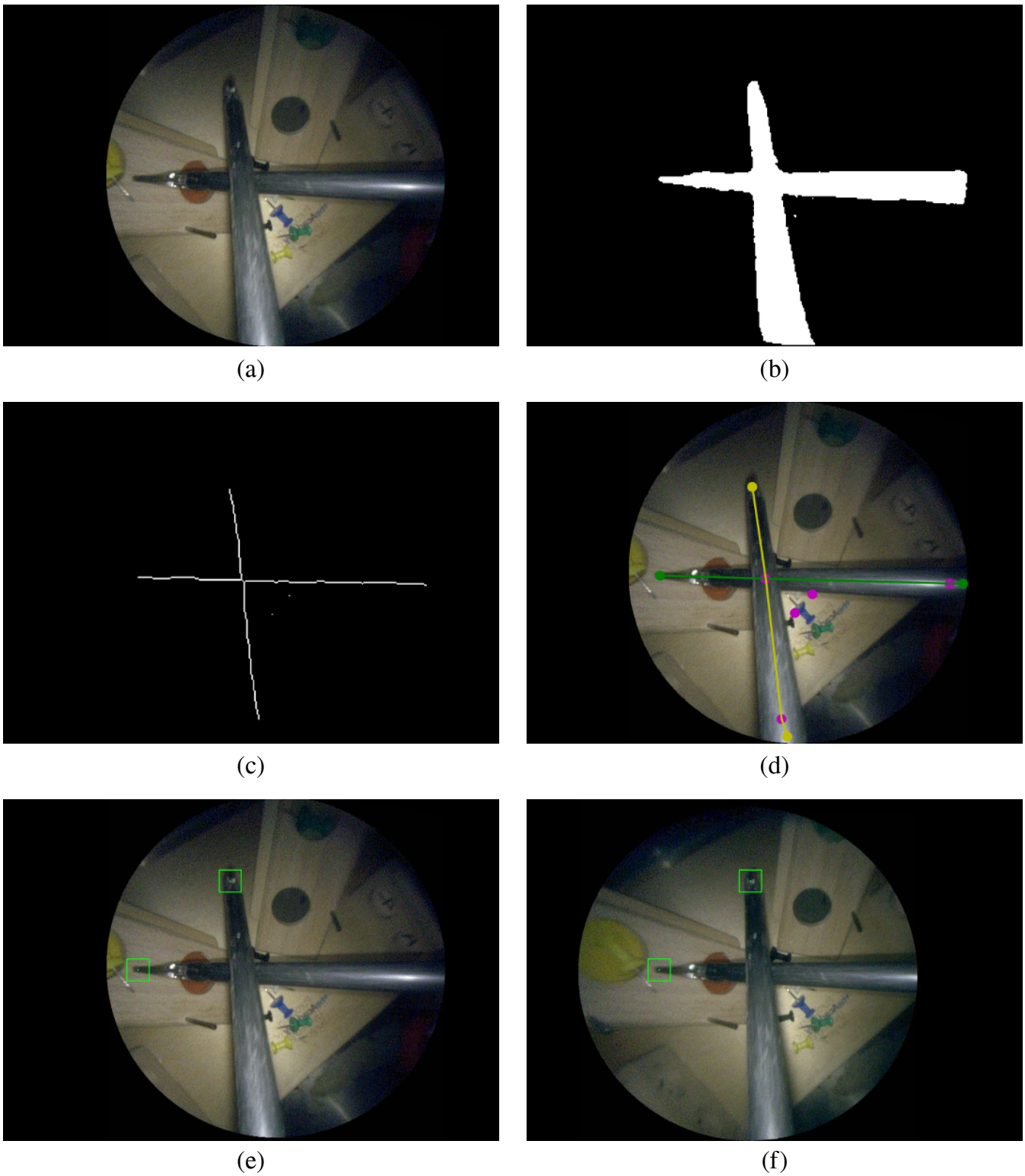

Figure 3. Instrument localization pipeline:(a) stereo-rectified right camera input image; (b) predicted tool segmentation; (c) skeletonisation; (d) graph extraction, 2D detection of entrypoints and tips, right and left instrument labelled in green and yellow; (e) left and (f) right stereo-matched tooltips in 2D (bottom row). The pink dots in (d) are graph nodes extracted from the skeleton in (c) but represent neither entrypoints nor tooltips.

extraction and 2D tooltip localization). Once the 2D tooltips have been estimated, they are extrapolated to the original image size and the disparity estimation and 3D tooltip reconstruction in Sec. 4.2.9 is performed on the original upsampled images of 1920×1080 pixels.

4.2.3 Instrument segmentation

In this work, we trained a CNN to segment instruments in our experimental setup (see Fig. 1). While having the necessary characteristics for a bimanual laparoscopic task, the visual appearance of the surgical training model we use is not representative of a real clinical environment. Therefore, we do not propose a new image segmentation approach but rather focus on the downstream computational questions. In order to translate our pipeline to the clinic, a newly annotated dataset containing real clinical images would need to be curated, and the images would need to contain artifacts typical of endoscopic procedures such as blood, partial occlusions, smoke, and blurring. Alternatively, an existing surgical dataset could be used. We have compiled a list of public datasets for tool segmentation [2] where the data available includes surgical scenes such as retinal microsurgery, laparoscopic adrenalectomy, pancreatic resection, neurosurgery, colorectal surgery, nephrectomy, proctocolectomy, and cholecystectomy amongst others. The compiled list also includes datasets for similar endoscopic tasks such as tool presence, instrument classification, tool-tissue action detection, skill assessment and workflow recognition, and laparoscopic image-to-image translation. The unlabelled data in these other datasets could also be potentially helpful for tool segmentation.

Next, we provide the details on how we built the segmentation model for our particular proof-of-concept of the robotic endoscope control. The dataset we curated consists of 1110 image-annotation pairs used for training, and 70 image-annotation pairs employed for validation (hyperparameter tuning). These $1110 + 70$ image-annotation pairs were manually selected by the first co-authors so that the chosen images represent well the variety of scenes in the task. They have been extracted from the recording of a surgical exercise in the lab, prior to the user study, and in a different location. There is no testing set at this point because the segmentation is an intermediary step. In Sec. 7.1, we give more details about our testing set, which is used to evaluate the whole tooltip localization pipeline (as opposed to just the segmentation). The images in each stereo pair do not look the same: there is an observable difference in colour tones between them.

[2] https://github.com/luiscarlosgph/list-of-surgical-tool-datasets

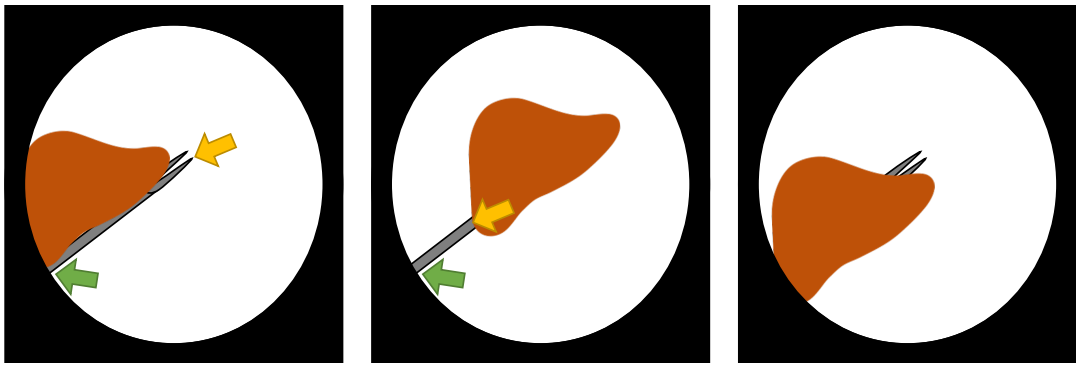

Figure 4. Behaviour and limitations of the instrument localization pipeline in the presence of occlusions. The detected entrypoint and tooltip are indicated by the green and yellow arrow, respectively. In the partially occluded instrument (left), there is a visible path from entrypoint to tip, therefore the instrument is correctly detected. However, when the tip is occluded (center), the tooltip is detected to be on the shaft. If the entrypoint is occluded (right), the instrument is not detected in this stage of the research as tools are expected to enter the scene from the boundary of the content area.

Therefore, the data set has an even number of left and right frames such that either of them could be used as input for the surgical tool segmentation. In the training set, 470 images (42%) do not contain any tool. In them, the endoscope just observes the task setting under different viewpoints and lighting conditions (diverse intensities of the light source). The remaining 640 images of the training set, and all images of the validation set, have been manually labelled with delineations of the laparoscopic tools. The U-Net (Ronneberger et al., 2015) architecture showed superior performance in the tool segmentation EndoVis MICCAI challenge (Allan et al., 2019). Therefore, this was the architecture of choice employed for segmentation (32 neurons in the first layer and convolutional blocks composed of Conv + ReLU + BN). A minibatch of 4 images is used. Default conventional values and common practice was followed for setting the hyperparameters as detailled hereafter. The batch normalization (Ioffe and Szegedy, 2015) momentum was set to 0.1 (default value in PyTorch). Following the U-Net implementation in (Ronneberger et al., 2015), Dropout (Srivastava et al., 2014) was used. In our implementation, Dropout was employed in layers with ≥ 512 neurons (p=0.5), as in (Garcia-Peraza-Herrera et al., 2017). Following Bengio (2012), the initial learning rate (LR) of choice was set to $1e-2$. The network was trained for a maximum of 100 epochs. As is common practice, LR decay was employed during training, multiplying the LR by 0.5 every 10 epochs. Data augmentation was limited to on-the-fly left-right flips. As we evaluate our segmentation using the intersection over union (IoU), our loss function $\mathcal{L}_{\text{IoU}}$ is a continuous approximation to the intersection over union (Rahman and Wang, 2016) averaged over classes:

$$
\begin{aligned}
I(\hat{\mathbf{y}}, \mathbf{y}, k) &= \sum_{i=1}^{P} \hat{y}_{i,k} \cdot y_{i,k}, \\
U(\hat{\mathbf{y}}, \mathbf{y}, k) &= \sum_{i=1}^{P} \hat{y}_{i,k} + \sum_{i=1}^{P} y_{i,k} - \sum_{i=1}^{P} \hat{y}_{i,k} \cdot y_{i,k}, \\
\mathcal{L}_{\text{IoU}}(\hat{\mathbf{y}}, \mathbf{y}) &= 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{I(\hat{\mathbf{y}}, \mathbf{y}, k) + \epsilon}{U(\hat{\mathbf{y}}, \mathbf{y}, k) + \epsilon},
\end{aligned}
\tag{2}
$$

where P is the number of pixels, $K = 2$ is the number of classes (instrument and background), $\hat{\mathbf{y}}$ represents the estimated probability maps, $\mathbf{y}$ represents the ground truth probability maps, $\hat{y}_{i,k}$ is the estimated probability of the pixel i belonging to the class k, and $y_{i,k}$ is the ground truth probability of the pixel i belonging to class k. A machine epsilon ϵ is added to prevent divisions by zero (e.g., in case that both prediction and ground truth are all background).

Once we have obtained a segmentation prediction from the trained convolutional model, we proceed to convert the segmentation into a graph, which is a stepping stone towards the tooltip detection.

4.2.4 Instrument graph construction

The instrument segmentation prediction is skeletonized via medial surface axis thinning (Lee et al., 1994). The resulting skeleton is converted via the Image-Py skeleton network framework (Xiaolong, 2019) into a pixel skeleton graph $G = (V, E)$ (see Fig. 5e), where V is a set of vertices and $E \subseteq \{\{x, y]\} : x, y \in V \wedge x \neq y\}$ is a set of edges. The nodes $v_i \in V$ are defined as a tuple $v_i = (i, \boldsymbol{p_i})$ where i and $\boldsymbol{p_i} = \{x_i, y_i\}$ represent the node index and $2D$ point image coordinates, respectively.

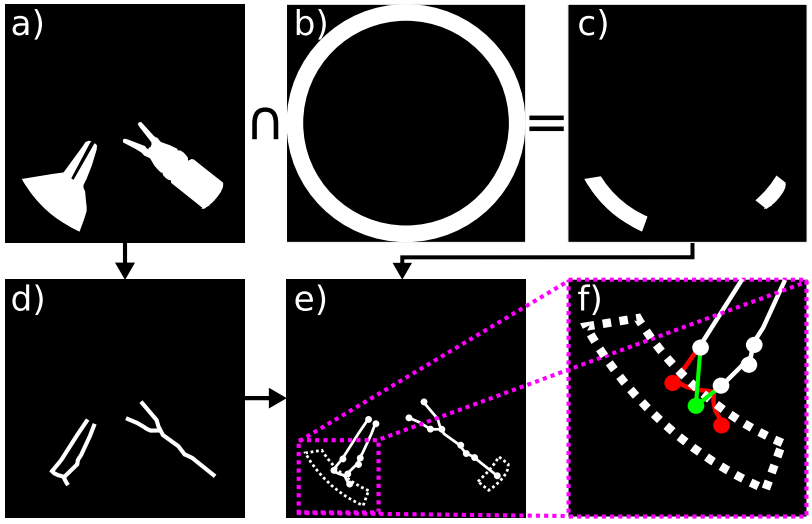

Figure 5. Surgical instrument graph construction and entry node extraction: a) segmentation mask; b) circle mask used to detect entrypoints; c) intersection of segmentation mask and circle mask; d) segmentation mask skeleton obtained according to (Lee et al., 1994); e) graph obtained from skeleton by means of (Xiaolong, 2019); f) entrypoint detection. If several graph nodes lie inside the entrypoint mask (in red), they are merged into a new single entry node (in green) whose position attribute is set to the centroid of all the graph nodes inside the dotted area.

4.2.5 Instrument entry node extraction

As the size of the image is known, a circular segmentation mask (see Fig.5b) is used to detect the graph nodes that could potentially correspond to instrument entrypoints. That is, given G, we populate a set R containing those graph nodes that represent tool entrypoints into the endoscopic image. Those graph nodes contained within the intersection of the circle mask and the tool segmentation mask are collapsed into a single new *entry* node $v_c = (n, \boldsymbol{p_c})$ per instrument, where $\boldsymbol{p_c} = \{x_c, y_c\}$ is set to the centroid of all nodes captured within the aforementioned intersection. See Fig. 5b-5f for an example of entry node extraction.

A depth-first search is launched from each entry node to determine all the graph nodes that can be reached from entry nodes. Those that cannot be reached are pruned from the graph.

4.2.6 Instrument leaf node to entry node matching

Let $L = \{v \in V : d_G(v) = 1 \wedge v \notin R\}$ be the set containing all *leaf* nodes, where $d_G(v) = |\{u \in V : \{u, v\} \in E\}|$. In this part of the instrument localization pipeline each *leaf* node in L is paired to an entrypoint node in R. This is solved by recursively traversing G, starting from each *leaf*. The criteria to decide which node to traverse next is coined in this work as *dot product recursive traversal*. It is based on the assumption that the correct path from a tip to a corresponding entrypoint is the one with minimal direction changes. The stopping condition is reaching an entry node.

Dot product recursive traversal operates as follows. Let $v_i, v_j \in V$ be two arbitrary nodes, and $\{v_i, v_j\}$ the undirected edge connecting them. Assuming v_i is previously visited and v_j being traversed, the next node $v*$ to visit is chosen following:

$$v^* = \underset{(k,\boldsymbol{p_k}) \in N(v_j) - A}{\operatorname{argmax}} \left(\frac{\boldsymbol{p_j} - \boldsymbol{p_i}}{||\boldsymbol{p_j} - \boldsymbol{p_i}||_2} \cdot \frac{\boldsymbol{p_k} - \boldsymbol{p_j}}{||\boldsymbol{p_k} - \boldsymbol{p_j}||_2} \right), \tag{3}$$

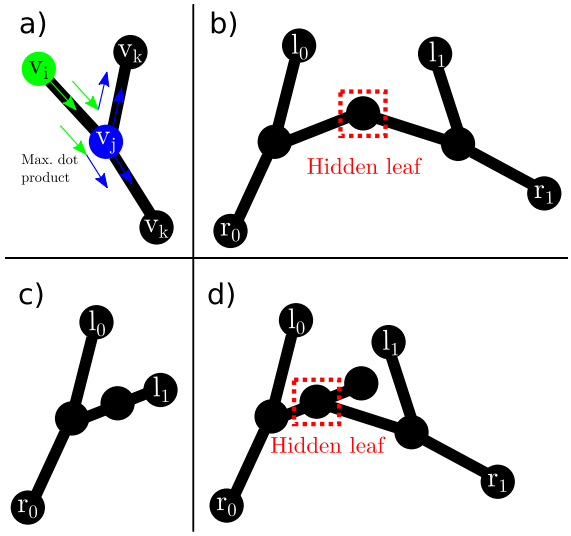

Figure 6. Leaf traversal and hidden leaves: a) graph traversal step. v_i (green) represents the previous node. v_j (blue) is the current node. v_k are possible next nodes. Following (3), the next node will be the one that maximizes the dot product; b) possible *hidden leaf* (dashed red box) connected to two nodes; c) node with two neighbours that does not represent a *hidden leaf* as both connecting edges are labelled after dot product traversal from l_1; d) possible *hidden leaf* (dashed red box) with three neighbours.

where $N(v_i) = \{w \in V : \{v_i, w\} \in E\}$, and $A = \{v_i\}$ is the set of nodes previously traversed. The idea behind (3) is to perform a greedy minimization of direction changes along the path from tooltip to entrypoint (Fig. 6a).

In the case of two tools present in the image, it is possible to have *hidden leaves* (Fig. 6b and 6d), defined as graph nodes that represent an overlap between the tip of an instrument and another instrument. This situation can easily occur (Fig. 6) in surgical tasks, including the task presented in the experiments from Sec. 6. There are two possible graph arrangements that can lead to *hidden leaves*. A node with exactly two (Fig. 6b) or three neighbours (Fig. 6d). Nonetheless, the number of neighbours alone does not facilitate the discovery of such *hidden leaves* (and subsequent disentanglement of tools), as it is also possible for a node with exactly two (could be a chain instead) or three (could be a bifurcation instead) neighbours to not be a *hidden leaf* (see Fig. 6c). Hence, extra information is needed. Aiming in this direction, after each successful traversal from a normal *leaf* to an entry node, all the edges along the path are labelled with the index of the entry node. In addition, all the edges directly connected to an entry node are also labelled.

A node with exactly two or three neighbours whose edges are all labelled with different entry nodes is a *hidden leaf*. Labelling helps to solve some of the *hidden leaf* cases. Such leaves can be duplicated, effectively splitting the graph into two, and disentangling the overlapped instruments. After disentanglement, they become normal leaves which can be assigned to an entry node by dot product traversal (3). Although not a *hidden leaf*, a node with exactly four neighbours whose edges are labelled represents an overlap which can be trivially disentangled. *Hidden leaves* such as the ones presented in Fig. 6b and 6d cannot be classified with certainty as such just with the graph/skeleton information. As shown in Fig. 6, different tool configurations/occlusions could lead to the same graph configuration. As not all the tips can be unambiguously detected, entry nodes that are unmatched after dot product traversal (i.e., they were not reached after launching a traversal from each leaf node to a possible entry node) are paired to the furthest opposing node connected to them.

Although the traversal from tips to entrypoints has been illustrated in this section with one or two instruments (as it is the case in our setup, see Fig. 1), the dot product traversal generalizes to setups with more instruments as the assumption that the path from tip to entrypoint is the one with less direction changes still holds.

4.2.7 Instrument graph pruning

Noisy skeletons can lead to inaccurate graphs containing more than two leaves matched to the same entry node, or more than two entry nodes connected to leaves. In our framework, unmatched entry nodes are deleted. In addition, due to the nature of our experimental setup, a maximum of two tools with two tips each can be found. Therefore, when more than two leaves are matched to the same entry node, only the two furthest are kept. Analogously, when more than two entry nodes are found and matched to leaves, the two kept are those with the longest chain (from entry node to leaves). That is, a maximum of two disentangled instrument graphs remain after pruning.

4.2.8 Left/right instrument identification

In the presence of a single instrument, left/right vertical semi-circles determine whether the instrument is left/right (see Fig. 7, right), i.e. if the entrypoint of the tool is located in the right half of the image, it is assumed that the subgraph connected to this entrypoint is the right instrument, and viceversa. Note that this simple method is also generalizable to scenarios with three to five instruments, which are different from the two-instrument solo surgery setting examined in this work (see Fig. 11), but still worth mentioning as there are some endoscopic procedures that may involve such number of tools (Abdi et al., 2016).

When two instruments are detected (i.e. two entrypoints with their corresponding subgraphs), a line segment connecting the entrypoints of both instruments is assumed to be the viewing horizon. A vertical line that is parallel to the vertical axis of the image and cuts through the central point of the viewing horizon defines whether the entrypoints (and subsequently the instruments) are left/right (see Fig. 7, left).

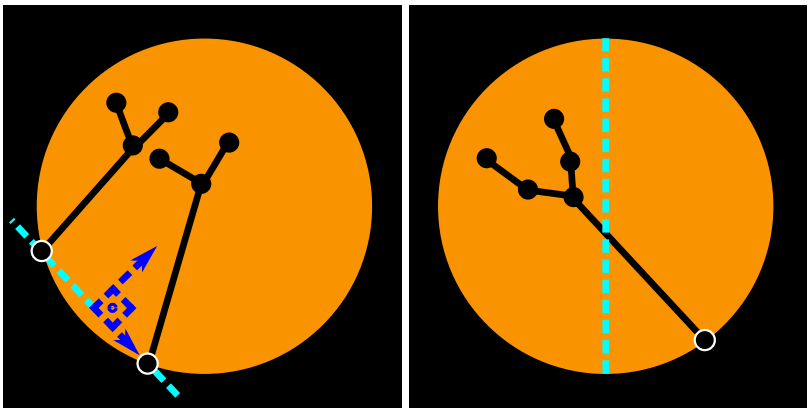

Figure 7. Left/right instrument identification. Two-instrument case (left). Single-instrument case (right). The location of the entrypoints is used to identify whether the instruments are left/right. When two instruments are visible, an imaginary vertical line (parallel to the vertical axis of the image) that crosses over the central point of the segment connecting both entrypoints is used to determine if the instrument is left/right. When there is only one instrument, the location of the entrypoint with regards to the vertical axis of the image determines which tool is visible. If the entrypoint resides in the right half, as in the figure above, this is considered to be the right instrument.

4.2.9 Tooltip stereo matching

Once the tips of the instruments have been detected and classified as right/left instrument, the disparity for each tooltip in the left and right stereo images is estimated using classical intensity-based template matching. As endoscope images are stereo-rectified, template matching with a sliding window of 64×64 pixels running over the same line (and only in one direction) suffices for the stereo matching. Normalized cross-correlation is the cost function of choice. Given the disparity measure for each tooltip, its depth can be reconstructed using the pinhole camera model and conventional epipolar geometry. The 3D reconstruction was performed with an extended Kalman filter (EKF). The EKF is valuable here, because of its capacity to bridge potential measurement gaps and to reduce the noise on the 3D position estimate, which is very sensitive to small disparity variations, as the lens separation of the TIPCAM is only 1.6 mm. The details of the EKF are specified in Sec. 5.2.2.

Although in our proposed experimental setup we use stereovision because we have an stereo-endoscope, many centers still use monoscopic endoscopes. In this case, a method such as that presented by Liu et al. (Liu et al., 2020) could be used to estimate the 3D tip location directly from the 2D endoscopy.

5 VISUAL SERVOING FOR ROBOTIC ENDOSCOPE CONTROL

A visual servoing controller determines the relative motion between a camera and an observed target in order to produce the desired camera view upon the target. In the case of AIT, the target is the (virtual) instrument tip position $\boldsymbol{s}$, defined by (1), and the camera is the endoscope held by the robotic endoscope holder. When working with endoscopes, the visual servoing controller needs to take into account a number of aspects specific for endoscopy, including the presence of the incision point which imposes a geometric constraint and the endoscope geometry. For the online estimation of the incision point, automated routines exist, such as (Gruijthuijsen et al., 2018; Dong and Morel, 2016). This section formalizes visual servoing approaches for REC in MIS.

5.1 Visual servoing approaches

Two classical approaches exist for visual servoing problems: image-based visual servoing (IBVS) and position-based visual servoing (PBVS) (Chaumette and Hutchinson, 2008). For REC, an extension to these methods is necessary as the camera motion is constrained by the presence of the incision point. In IBVS, this can be done by modifying the interaction matrix, such that it incorporates the kinematic constraint of the incision point (Osa et al., 2010) or such that it describes the mapping between the image space and the joint space of the robotic endoscope holder (Uecker et al., 1995; Zhao, 2014). As these IBVS approaches only act in the image plane, the zoom level can be controlled by a decoupled depth controller (Chen et al., 2018). PBVS approaches can incorporate the incision constraint in an inverse kinematics algorithm that computes the desired robot pose, given the desired endoscopic view (Yu et al., 2013; Eslamian et al., 2016).

Implementations of the above approaches, that the authors are aware of, lack generality: they are formulated for a specific robotic endoscope holder and do not cover oblique-viewing endoscopes, while such endoscopes are commonly used in MIS procedures. Yet, oblique-viewing endoscopes are the most challenging to handle for clinicians (Pandya et al., 2014), and could thus reap most benefits of REC. Generic constraint-based control frameworks, such as eTaSL (Aertbeliën and De Schutter, 2014), could be applied with a generalized endoscope model, like presented below, although they are slower than explicit visual servoing methods.

5.2 Visual servoing with generalized endoscope model

This section introduces a novel generalized endoscope model for visual servoing that incorporates the incision constraint, as well as the endoscope geometry. Such a model is presented here, along with the ensuing modifications to the classical IBVS and PBVS approaches.

5.2.1 Generalized endoscope model

Endoscopes come in different forms and sizes. Rigid endoscopes are typically straight, but can also be pre-bent. The camera can be oriented collinear with the longitudinal axis of the scope or can be positioned at an oblique angle. Some scopes are flexible over their entire length, others contain a proximal rigid straight portion with a distal bendable portion.

Fig. 8 visualizes a general endoscope geometry that encompasses all the above configurations, along with the frames of interest. The incision frame $\{i\}$ is defined at the incision point and is common for all robotic endoscope holders. The z-axis of $\{i\}$ is the inward-pointing normal of the body wall. A frame $\{t\}$ is connected to the distal tip of the straight portion of the endoscope shaft, with its z-axis pointing along the shaft axis. In the most general case, the camera frame $\{c\}$ is located at an offset and rotated with respect to $\{t\}$. The offset can reproduce for tip articulation, for stereo camera lens separation. The rotation can account for oblique-viewing endoscopes. As such, this endoscope model can describe complex endoscopes, such as the articulating 3D video endoscope ENDOEYE FLEX 3D (OLYMPUS, Japan) (Fig. 9).

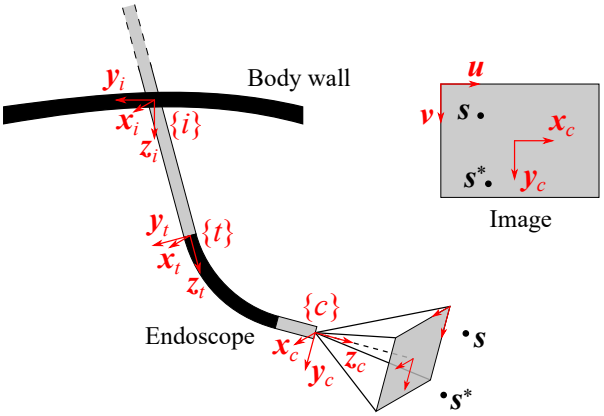

Figure 8. Definition of reference frames in a generalized endoscope model. The incision frame $\{i\}$ is located at the incision point in the body wall, the distal tip frame $\{t\}$ at the end of the straight endoscope shaft and the camera frame $\{c\}$ at the end of the endoscope, in the optical center of the camera. The image produced by the endoscope is also shown (upper right insert), along with the projections of the detected feature of interest $\boldsymbol{s}$ and its desired position $\boldsymbol{s}^*$, and with the image coordinate vectors $(\boldsymbol{u}, \boldsymbol{v})$.

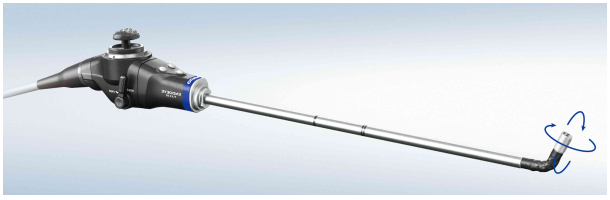

Figure 9. ENDOEYE FLEX 3D ©OLYMPUS CORPORATION (Tokyo, Japan).

Starting from this general endoscope model, different visual servoing approaches for REC will be detailed next. The visual servoing approaches strive to determine the endoscope motion that is needed to match the position $\boldsymbol{s} = \begin{bmatrix} x & y & z \end{bmatrix}^T$ of the feature of interest, expressed in the camera frame, with its desired position $\boldsymbol{s}^* = \begin{bmatrix} x^* & y^* & z^* \end{bmatrix}^T$, while taking into account the presence of the incision point. The visual servoing approaches assume that the robot endoscope holder has three active DoFs that can produce any linear velocity of the endoscope tip. In order to obtain a fully determined endoscope motion, it is further assumed that the remaining rolling DoF about the endoscope axis is not controlled by the visual servoing controller, but by an external controller. Note that, as was pointed out in Sec. 3, this DoF could be employed to control the camera horizon.

The following notation will be used in the subsequent sections: a rotation of angle ξ about the axis i will be denoted by $\boldsymbol{R}_i(\xi)$. For a transformation from a frame $\{j\}$ to a frame $\{i\}$, the notation $\boldsymbol{T}^i_j$ will be used, consisting of a rotation $\boldsymbol{R}^i_j$ and a translation $\boldsymbol{P}^i_j$. Further, the twist vector $\boldsymbol{t} = \begin{bmatrix} \boldsymbol{v}^T & \boldsymbol{\omega}^T \end{bmatrix}^T$ is defined as the concatenation of a linear velocity $\boldsymbol{v}$ and an angular velocity $\boldsymbol{\omega}$. For all kinematic variables, the reference frames will be indicated with a trailing superscript. For the features $\boldsymbol{s}$ and the error $\boldsymbol{e}$ in the camera frame, the trailing superscript ^{c} is mostly omitted for brevity.

5.2.2 EKF for tooltip 3D position reconstruction

The instrument localization pipeline from Sec. 4.2 yields the tooltip image coordinates u_l, v_l and the disparity d_x. The 3D tooltip position, required for the visual servoing methods, is estimated from these measurement data, through an EKF. The state transition model describes a linear tooltip motion of exponentially decreasing velocity, partially expressed in frames $\{i\}$ and $\{c\}$ to limit the non-linearity, and the observation model implements the pinhole camera model:

$$\begin{aligned} \boldsymbol{x}_k &= g(\boldsymbol{x}_{k-1}, \boldsymbol{t}^c_{c,k}) + \boldsymbol{\epsilon}_k, \\ &= \begin{bmatrix} \boldsymbol{s}^c_{k-1} + \Delta T(\boldsymbol{R}^c_i \dot{\boldsymbol{s}}^i_{k-1} + \boldsymbol{L}_{3D}(\boldsymbol{s}^c_{k-1})\boldsymbol{t}^c_{c,k}) \\ \lambda_s \dot{\boldsymbol{s}}^i_{k-1} \end{bmatrix} + \boldsymbol{\epsilon}_k, \end{aligned} \tag{4}$$

$$\boldsymbol{z}_k = h(\boldsymbol{x}_k) + \boldsymbol{\delta}_k = \begin{bmatrix} \frac{x_k f_x}{z_k} + c_x \\ \frac{y_k f_y}{z_k} + c_y \\ \frac{b_c f_x}{z_k} \end{bmatrix} + \boldsymbol{\delta}_k, \tag{5}$$

where $\boldsymbol{x}_k = \begin{bmatrix} \boldsymbol{s}^{c^T}_k & \dot{\boldsymbol{s}}^{i^T}_k \end{bmatrix}^T$ is the state vector, $\boldsymbol{s}^c_k$ is the tooltip position in the camera frame $\{c\}$, $\dot{\boldsymbol{s}}^i_k$ is the tooltip velocity in the incision frame $\{i\}$, $\boldsymbol{t}^c_c$ is the camera twist, $\boldsymbol{L}_{3D}$ is the 3D interaction matrix from Eq. (20), λ_s is a reduction factor < 1 (governing the exponential decrease of $\dot{\boldsymbol{s}}^i$), $\boldsymbol{z}_k = \begin{bmatrix} u_{l,k} & v_{l,k} & d_{x,k} \end{bmatrix}^T$ is the observation vector, f_x, f_y, c_x, c_y are the intrinsic camera parameters, b_c the distance between the optical centres of the (left and right) cameras, and $\boldsymbol{\epsilon}_k$ and $\boldsymbol{\delta}_k$ are the usual process and observation noises. The velocity reduction factor λ_s is introduced to scale down the contribution of dead reckoning during measurement gaps.

5.2.3 Image-based visual servoing (IBVS)

IBVS aims to determine the camera motion to move the 2D projection of the 3D feature point $\boldsymbol{s}$ to its desired position in the image plane. Assuming a pinhole camera model, the 2D projection $\boldsymbol{s}_n$ is obtained

by expressing $\boldsymbol{s}$ in normalized camera coordinates:

$$\boldsymbol{s}_n = \begin{bmatrix} x_n & y_n \end{bmatrix}^T = \begin{bmatrix} x/z & y/z \end{bmatrix}^T. \tag{6}$$

Classically, the relation between the camera twist $\boldsymbol{t}_c^c$ and the 2D feature point velocity $\dot{\boldsymbol{s}}_n$ is expressed by the interaction matrix $\boldsymbol{L}_{2D}$ (Chaumette and Hutchinson, 2008):

$$\dot{\boldsymbol{s}}_n = \boldsymbol{L}_{2D}\boldsymbol{t}_c^c, \tag{7}$$

where

$$\boldsymbol{L}_{2D} = \begin{bmatrix} -1/z & 0 & x_n/z & x_n y_n & -(1+x_n^2) & y_n \\ 0 & -1/z & y_n/z & 1+y_n^2 & -x_n y_n & -x_n \end{bmatrix}. \tag{8}$$

In this equation, it is assumed that the camera has six DoFs, while only three are available for the endoscope control. To incorporate these constraints, the camera twist $\boldsymbol{t}_c^c$ needs to be mapped first to the twist of tip of the endoscope's straight portion $\boldsymbol{t}_t^t$:

$$\boldsymbol{t}_c^c = \boldsymbol{J}_t^c \boldsymbol{t}_t^t, \tag{9}$$

where $\boldsymbol{J}_t^c$ is the well-known expression for a twist transformation:

$$\boldsymbol{J}_t^c = \begin{bmatrix} \boldsymbol{R}_t^c & -\boldsymbol{R}_t^c[\boldsymbol{p}_c^t]_\times \\ \boldsymbol{0} & \boldsymbol{R}_t^c \end{bmatrix}. \tag{10}$$

The operator $[\,]_\times$ is the operator for the skew-symmetric matrix. The incision constraint introduces a coupling between the linear tip velocity $\boldsymbol{v}_t^t$ and angular tip velocity $\boldsymbol{\omega}_t^t$ and can be expressed as:

$$\boldsymbol{t}_t^t = \boldsymbol{J}_i \boldsymbol{v}_t^t, \tag{11}$$

with the incision transformation

$$\boldsymbol{J}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1/l & 0 \\ 1/l & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{12}$$

and the inserted endoscope length $l = \|\boldsymbol{p}_t^i\|$. Combining (7)-(12) yields the modified interaction matrix $\boldsymbol{L}'_{2D}$:

$$\boldsymbol{L}'_{2D} = \boldsymbol{L}_{2D}\boldsymbol{J}_t^c\boldsymbol{J}_i \tag{13}$$

which maps $\boldsymbol{v}_t^t$ to $\dot{\boldsymbol{s}}_n$. This matrix is a generalized form for the modified interaction matrix presented in (Osa et al., 2010).

As is customary in visual servoing, the error in the normalized image space is expressed as

$$\boldsymbol{e}_n = \boldsymbol{s}_n - \boldsymbol{s}_n^* \tag{14}$$

and the control law enforces an exponential decay of the error:

$$\dot{\boldsymbol{e}}_n = -\lambda \boldsymbol{e}_n, \tag{15}$$

characterized by the time constant $\tau = 1/\lambda$. For a constant $\boldsymbol{s}_n^*$, this yields the desired endoscope tip velocity:

$$\begin{aligned} \dot{\boldsymbol{e}}_n = \dot{\boldsymbol{s}}_n = \boldsymbol{L}'_{2D}\boldsymbol{v}_t^t = -\lambda \boldsymbol{e}_n \\ \Rightarrow \boldsymbol{v}_t^t = -\lambda \boldsymbol{L}'^{+}_{2D}\boldsymbol{e}_n. \end{aligned} \tag{16}$$

5.2.4 Image-based visual servoing with decoupled depth control (IBVS+DC)

IBVS only seeks to optimize the 2D projected position $\boldsymbol{s}_n$ of the target point $\boldsymbol{s}$ in the image plane. As such IBVS alone is insufficient to control the 3D position of the endoscope. A decoupled depth controller can be added to control the third DoF. This was proposed in (Chen et al., 2018) and will be generalized here.

The depth controller acts along the z-axis of the camera frame $\{c\}$ and uses the kinematic relation between the camera twist $\boldsymbol{t}_c^c$ and the change in the depth z of $\boldsymbol{s}$:

$$\dot{z} = \boldsymbol{L}_z \boldsymbol{t}_c^c, \tag{17}$$

where

$$\boldsymbol{L}_z = \begin{bmatrix} 0 & 0 & -1 & -y & x & 0 \end{bmatrix}. \tag{18}$$

To reduce the depth error $e_z = z - z^*$, concurrently with the image-space error $\boldsymbol{e}_n$, a similar reasoning as with IBVS can be followed, yielding:

$$\begin{aligned} \begin{bmatrix} \dot{\boldsymbol{e}}_n \\ \dot{e}_z \end{bmatrix} = \begin{bmatrix} \dot{\boldsymbol{s}}_n \\ \dot{z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{2D} \\ \boldsymbol{L}_z \end{bmatrix} \boldsymbol{J}_t^c \boldsymbol{J}_i \boldsymbol{v}_t^t = \begin{bmatrix} \boldsymbol{L}'_{2D} \\ \boldsymbol{L}'_z \end{bmatrix} \boldsymbol{v}_t^t = -\lambda \begin{bmatrix} \boldsymbol{e}_n \\ e_z \end{bmatrix} \\ \Rightarrow \boldsymbol{v}_t^t = -\lambda \begin{bmatrix} \boldsymbol{L}'_{2D} \\ \boldsymbol{L}'_z \end{bmatrix}^{+} \begin{bmatrix} \boldsymbol{e}_n \\ e_z \end{bmatrix}. \end{aligned} \tag{19}$$

To differentiate between directions, it is possible to define λ as a diagonal matrix, rather than as a scalar.

5.2.5 3D image-based visual servoing (3D IBVS)

Instead of decoupling the control in the image plane and the depth control, the 3D feature $\boldsymbol{s}$ can also be used directly to define the 3D motion of the endoscope. This requires a 3D interaction matrix $\boldsymbol{L}_{3D}$, which can be derived from the kinematic equations of motion for the stationary 3D point $\boldsymbol{s}$ in the moving camera frame $\{c\}$:

$$\dot{\boldsymbol{s}} = -\boldsymbol{v}_c^c - \boldsymbol{\omega}_c^c \times \boldsymbol{s} = \boldsymbol{L}_{3D} \boldsymbol{t}_c^c, \tag{20}$$

with

$$\boldsymbol{L}_{3D} = \begin{bmatrix} -\boldsymbol{I} & [\boldsymbol{s}]_\times \end{bmatrix}. \tag{21}$$

As before, the modified interaction matrix $\boldsymbol{L}'_{3D}$ can be obtained by including the offset of the tip frame with respect to the camera frame and the incision constraint. The desired endoscope velocity that ensures an exponential decay of the error $\boldsymbol{e} = \boldsymbol{s} - \boldsymbol{s}^*$ follows then from:

$$\begin{aligned} \dot{\boldsymbol{e}} = \dot{\boldsymbol{s}} = \boldsymbol{L}_{3D} \boldsymbol{J}_t^c \boldsymbol{J}_i \boldsymbol{v}_t^t = \boldsymbol{L}'_{3D} \boldsymbol{v}_t^t = -\lambda \boldsymbol{e} \\ \Rightarrow \boldsymbol{v}_t^t = -\lambda \boldsymbol{L}'^{+}_{3D} \boldsymbol{e}. \end{aligned} \tag{22}$$

5.2.6 Position-based visual servoing (PBVS)

PBVS identifies the camera pose, with respect to an external reference frame, that produces the desired view upon the 3D feature $\boldsymbol{s}$ and moves the camera towards this pose. As mentioned before, the camera pose is constrained to three DoFs due to the presence of the incision point and the separate horizon stabilization. Finding the desired camera pose, while taking into account its kinematic constraints, involves solving the inverse kinematics for the endoscope as defined in Fig. 8.

The forward kinematics of the endoscope can be described as a function of three joint variables (θ_1, θ_2, l). Based on these variables, any endoscope pose can be reached by applying successive operations in a forward kinematics chains. When these joint variables are set to zero, the endoscope system is in a configuration where the incision frame $\{i\}$ coincides with the distal tip frame $\{t\}$, while the camera frame is offset by $\boldsymbol{p}_c^t$ and rotated by $\boldsymbol{R}_c^t = \boldsymbol{R}_x(\alpha)$, with α the oblique viewing angle of the endoscope. Starting from this configuration, θ_1 rotates $\{t\}$ about its y-axis, then θ_2 rotates it about its x-axis and finally l translates it along its z-axis. This leads to the following forward kinematic equations, expressed in the reference frame $\{i\}$:

$$\begin{aligned}\boldsymbol{T}_c^i\tilde{\boldsymbol{s}} &= \boldsymbol{T}_c^{i^*}\tilde{\boldsymbol{s}}^* \\ &= \boldsymbol{T}_t^{i^*}\boldsymbol{T}_c^t\tilde{\boldsymbol{s}}^* \\ &= \begin{bmatrix}\boldsymbol{R}_y(\theta_1^*)\boldsymbol{R}_x(\theta_2^*) & \boldsymbol{0} \\ \boldsymbol{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & l^*\hat{\boldsymbol{e}}_3 \\ \boldsymbol{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_x(\alpha) & \boldsymbol{p}_c^t \\ \boldsymbol{0}^T & 1\end{bmatrix}\tilde{\boldsymbol{s}}^*,\end{aligned} \tag{23}$$

with $\hat{\boldsymbol{e}}_3 = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T$ the unit vector along the z-direction. The trailing * designates a desired value, different from the current value. The $\tilde{}$ signifies the homogeneous representation of a 3D vector. Equation (23) constitutes a system of three equations in the unknowns $(\theta_1^*, \theta_2^*, l^*)$. 1 elaborates the analytic solution to this inverse kinematics problem.

The solution of the inverse kinematics can be inserted in the forward kinematics equations to obtain the desired position of the distal endoscope tip $\boldsymbol{p}_t^{i^*}$:

$$\boldsymbol{p}_t^{i^*} = \begin{bmatrix}l^*\sin(\theta_1^*)\cos(\theta_2^*) \\ -l^*\sin(\theta_2^*) \\ l^*\cos(\theta_1^*)\cos(\theta_2^*)\end{bmatrix}, \tag{24}$$

which straightforwardly leads to the position error of the distal tip, expressed with respect to the incision frame $\{i\}$:

$$\boldsymbol{e}^i = \boldsymbol{p}_t^i - \boldsymbol{p}_t^{i^*}. \tag{25}$$

When an exponential decaying error is required, the desired endoscope velocity becomes:

$$\boldsymbol{v}_t^i = \dot{\boldsymbol{e}}^i = -\lambda\boldsymbol{e}^i \tag{26}$$

$$\tag{27}$$

and can be expressed in the frame $\{t\}$ as:

$$\boldsymbol{v}_t^t = -\lambda\boldsymbol{R}_i^t\boldsymbol{e}^i. \tag{28}$$

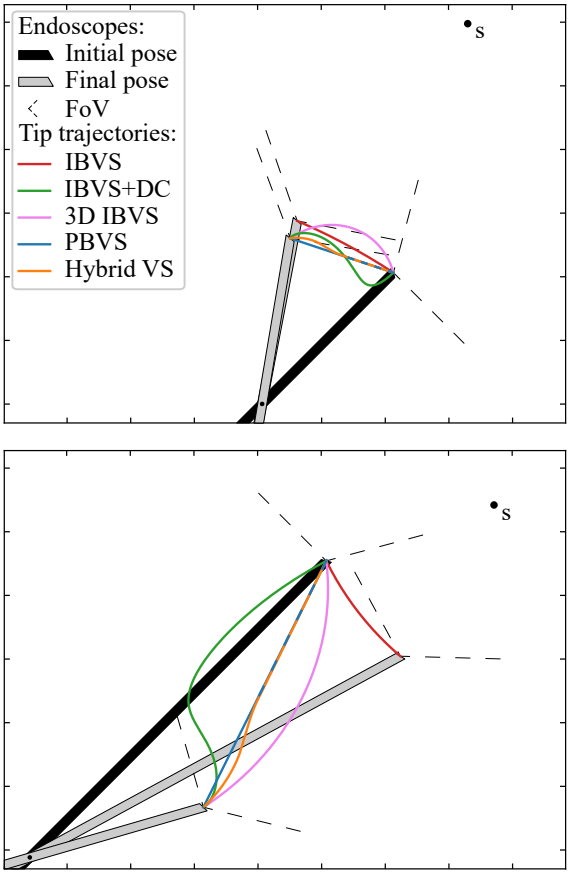

Figure 10. Comparison of the endoscope tip trajectories for different visual servoing approaches for REC. In this simulation, an oblique-viewing 30° endoscope with a 120° FOV was used. Its initial and final pose are drawn. In the final pose, the feature of interest $\boldsymbol{s}$ is in the desired position with respect to the camera. The trajectories are drawn for a case with a small initial depth error (top) and a large initial depth error (bottom).

5.3 Simulation of visual servoing methods

A simulation was implemented to validate all four visual servoing methods for REC: IBVS, IBVS+DC, 3D IBVS and PBVS. Fig. 10 presents a visual comparison between them, for a 30° oblique-viewing endoscope with a 120° FoV. In all simulations, $\boldsymbol{s}$ enters the FoV from a side. The visual serviong controller moves the endoscope to center $\boldsymbol{s}$ within its FoV at a given depth z^*, or $\boldsymbol{s}^* = \begin{bmatrix} 0 & 0 & z^* \end{bmatrix}^T$. The trajectories described by the endoscope tip are shown in the graphs, as well as the initial (marked in black) and final (marked in grey) endoscope poses.

5.3.1 Comparison of visual servoing methods

From the graphs, it is clear that IBVS differs from the other approaches in that, by construction, it does not attain the desired depth z^* in the final endoscope pose. Moreover, IBVS also doesn't guarantee a constant depth z. Consequently, z will drift towards undesired depths over time. In some configurations, this can be counteracted by separately controlling l to stay constant, but this does not hold for the general case. IBVS alone is thus unsuitable for REC and 3D information about $\boldsymbol{s}$ is a requirement.

Both IBVS+DC and 3D IBVS linearize the visual servoing problem in the camera frame. This enables a desired exponential decay of the targeted errors, but does not produce a well-controlled endoscope motion

in Cartesian space. It can be seen from Fig. 10 that the trajectories for these methods deviate from the straight trajectory that is accomplished by PBVS, and more so for large initial errors. As space is limited in REC and the environment delicate, the straight trajectory of PBVS appears favourable compared to its alternatives.

IBVS typically yields more accurate visual servoing results than PBVS, because the feedback loop in IBVS-based methods can mitigate camera calibration errors (excluding stereo calibration errors). However, the objective in REC often is to keep $\boldsymbol{s}$ inside a specific region of the endoscopic image (cf. position hysteresis), rather than at an exact image coordinate. The importance of the higher accuracy of IBVS is thus tempered by this region-based control objective: small calibration inaccuracies are acceptable. Therefore, and in contrast to the claims in (Osa et al., 2010), it can be argued that the predictability of a straight visual servoing trajectory outweighs the importance of the visual servoing accuracy. This argumentation points out why PBVS is the preferred approach for REC, especially when large initial errors exist.

5.3.2 Hybrid PBVS and 3D IBVS

If the accuracy of PBVS would need to be enhanced, e.g., when significant calibration errors exist, it is possible to apply a hybrid visual servoing method. PBVS can be used until the initial error drops below a certain threshold and from there, the visual servoing controller gradually switches to an IBVS-based approach for refinement, by applying a weighted combination of the desired tip velocities $\boldsymbol{v}_t^t$ computed by each visual servoing method. The curved shape of IBVS trajectories can thus be suppressed. In experiments that are not further documented here, it was observed that 3D IBVS, which ascertains an exponential Cartesian error decay, provided a more predictable and thus more desirable endoscope behaviour than IBVS+DC. To ensure robustness against potential calibration errors, the hybrid combination of PBVS and 3D IBVS was thus selected for the experiments in Sec. 6. Fig. 10 shows the simulated performance of the hybrid visual servoing approach, which gradually transitions from PBVS to 3D IBVS when the error $\|\boldsymbol{e}_n\|$ in the normalized image space goes from 0.6 to 0.3.

6 EXPERIMENTS

To determine the feasibility of the proposed autonomous endoscopy framework, an experimental setup was built (see Fig. 1). The mockup surgical setting consisted of a laparoscopic skills testing and training model (LASTT) placed within a laparoscopic box trainer (see Fig. 11). A *bi-manual coordination* exercise was chosen as the target surgical task for the experiments. In this task, a set of pushpins need to be passed between hands and placed in the right pockets. The choice of both laparoscopic trainer and surgical task was clinically motivated. The present study is largely inspired by the surgical scenario occurring during spina bifida Bruner (1999); Meuli and Moehrlen (2014); Kabagambe et al. (2018) surgical procedures (see Fig. 12). In this fetal treatment, a surgeon operates while another one guides the endoscope. The LASTT model along with the bi-manual coordination task have been developed by The European Academy for Gynaecological Surgery [3] as an initiative to improve quality control, training and education in gynaecological surgery (Campo et al., 2012). Therefore, they are ideal candidates for the feasibility study of the proposed autonomous endoscopy framework.

[3] https://esge.org/centre/the-european-academy-of-gynaecological-surgery

6.1 Bi-manual coordination task

This task starts by placing a set of coloured pushpins at the base of the LASTT model (see Fig. 11, left). There are two pins of each colour. The operator has to pick a pin with the non-dominant hand, pass it to the dominant hand (Fig. 11, right), and place it inside the pocket of the same colour. The LASTT task is successfully completed when a pushpin of each of the six colours has been placed in a corresponding pocket, within less than five minutes. If the pin is dropped during the procedure, a second pin of the same colour has to be picked up from the base. If the second pin is also dropped, the test is considered a failure.

Figure 11. LASTT (European Academy of Gynaecological Surgery, 2020) laparoscopic training model. Inital position of the pins before starting the bi-manual coordination task (left). Procedure to pass a pin from the non-dominant to the dominant hand (right).

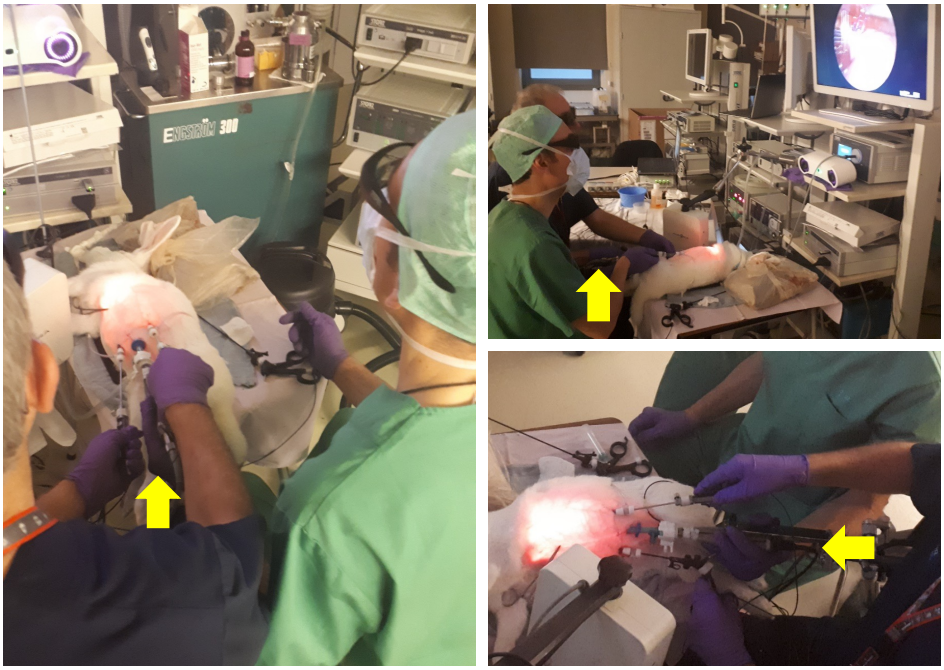

Figure 12. Spina bifida intervention performed on an animal model. The surgeon dressed in blue scrubs controls the instruments and manipulates the tissue. The colleague dressed in green guides and holds the endoscope camera during the intervention. The yellow arrows point to the hand of the assistant guiding the camera. As becomes evident in the pictures above, this operating arrangement is not ergonomic, leading to discomfort that increases with the duration of the intervention, and severely limiting the tasks that the surgeon controlling the camera can perform. Picture courtesy of Prof. Jan Deprest.

As shown in the demonstration video of the bi-manual coordination task with the LASTT model of the European Academy of Gynaecological Surgery[4], this exercise cannot be performed with a fixed immobile endoscope due to the reduced field of view of the endoscope, the limited space available for maneuvers within the operating cavity, and the small size of the pins (which resembles small tissue structures). All of these characteristics of the LASTT model mimic the real operating conditions, particularly for gynaecological interventions. Without a robotic endoscope holder, the bi-manual coordination task is performed with one trainee handling the laparoscopic graspers and another trainee acting as the (human) camera assistant. The assistant should hold the endoscope and keep the view centered on what the laparoscopic operator is doing. In our experiments, this human camera assistant is replaced by the VIRTUOSE6D[5] (HAPTION SA, Laval, France) robotic arm. As shown in (Avellino et al., 2020), the dimensions, workspace, and supported payload of this robotic arm are well suited for robotic endoscope control [6]. The operational workspace is defined as a cube of side $450\,\text{mm}$ and is located in the center of the workspace envelope. The extremities of the workspace envelope are bounded by a volume of $1330 \times 575 \times 1020\,\text{mm}^3$. The payload supported by the VIRTUOSE6D is $35\,\text{N}$ (peak)/ $10\,\text{N}$ (continuous). Additionally, the Virtuose6D features passive gravity compensation, which can be mechanically adjusted to carry up to $8\,\text{N}$. Therefore, although in our setup we are using a stereo-endoscope, this system is also able to hold laparoscopy cameras (e.g. those used in abdominal surgery). In our setup, the robotic arm was holding the KARL STORZ TIPCAM1, as shown in Fig. 1. The VIRTUOSE6D was programmed to respond to semantically rich AIT instructions (Sec. 3). Additionally, it featured a comanipulation fallback mode (Sec. 3.2), which it naturally supports owing to its mechanical backdrivability.

6.2 Study participants

A total of eight subjects participated in the study. Two *surgeons*, two *plateau novices*, and four *novices*. The *plateau novices* were authors of the study, who started out as novices, but familiarized themselves with the system and the task until they reached a plateau in the learning curve. Each participant performed the bi-manual coordination task five times. Before these trials, each participant practised 5–10 minutes to perform the task while assisted by the robotic endoscope holder.

6.3 Configuration of the autonomous endoscope for the study

The autonomous endoscope controller implemented the *Hybrid PBVS and 3D IBVS* method (Sec. 5.3.2), switching from PBVS to 3D IBVS when the error $\|e_n\|$ in the normalised image space decreased from 0.6 to 0.3. The target position of the endoscope tip was set to $\boldsymbol{s}^* = \begin{bmatrix}0 & 0 & z^*\end{bmatrix}^T$, where $z^* = 8\,\text{cm}$. The endoscope tip was controlled to track a trajectory towards its desired position with the tip velocity $\boldsymbol{v}_t^t$ limited to $2\,\text{cm/s}$. This trajectory was implemented as a soft virtual fixture, with stiffness of $0.3\,\text{N/mm}$. The aforementioned low speed and stiffness were found to provide smooth and predictable motions. They proved also helpful in avoiding sudden motions when one of the instruments is occluded and the remaining one is located in the violation zone. Low speed and stiffness were also necessary because of the $340\,\text{ms}$ delay on the measurements updates of $\boldsymbol{s}$. The framegrabber was responsible for $230\,\text{ms}$ of this delay. A framegrabber that supports NVIDIA GPUDirect, not available to us at the time of writing, could be used to mitigate this latency. The other $110\,\text{ms}$ came from the 3D tooltip localisation pipeline (Sec. 4.2). A delay that could be potentially reduced in future work using TensorRT. Measurements were available at $9\,\text{Hz}$.

[4] https://europeanacademy.org/training-tools/lastt/

[5] https://www.haption.com/en/products-en/virtuose-6d-en.html

[6] https://www.youtube.com/watch?v=R1qwKAWFOIk

The position hysteresis approach, which was illustrated in Fig. 2, was applied separately in the image plane and along the viewing direction. In the image plane, the target zone A occupied the first 40% of the endoscopic image radius, the transition zone B the next 20%, and the violation zone C the remaining 40%. Along the viewing axis, the target zone was set to $3\,$cm in both directions of z^*. The violation zone started at a distance of $5\,$cm with respect to z^*. The EKF for stereo reconstruction (Sec. 4.2.9) was used to fill missing data up to $1\,$s after the last received sample. When the instruments were lost from the view for more than 10 seconds, the REC switched from the AIT mode to the comanipulation fallback mode, waiting to be manually reset to a safe home position.

When designing the experiments, two preliminary observations were made: (1) the AIT instruction that fixes the tracking target on the tip of the instrument held by the dominant hand was most convenient, and (2) instructions to change the zoom level were not used. The latter observation is easily explained by the nature of the LASTT task, which requires overview rather than close-up inspections. The former observation points out that it is confusing to track a virtual instrument tip in between the real tooltips. While concentrated on the task, participants tend to forget their non-dominant hand and move it out of the view (or in and out of the view without a particular reason). This affected the position of the virtual instrument tip in unexpected ways. In those situations where tracking the non-dominant hand is relevant (e.g., when passing the pin from one hand to another), participants quickly learned to keep the tips together. Hence, tracking the dominant-hand tool was sufficient to provide a comfortable view, and this became the only operation mode that participants used. In fact, as an operator, it was convenient to know that the system is tracking your dominant hand: this is easy to understand and remember. Thus, during all the experiments, the only instruction that was issued was to make the camera track the dominant hand. As all participants were right-handed, $\boldsymbol{s}$ was assigned to the tip of the right-hand tool.

7 RESULTS AND DISCUSSION

In this section we provide quantitative results on the tooltip tracking accuracy, the responsiveness and usability of the visual servoing endoscopic guidance, and the learning curve of the user study participants.

7.1 Validation of instrument localization pipeline

Given an endoscopic video frame as input, the tooltip localization pipeline produces an estimate of the 2D location of the surgical instrument tips (see Fig. 3). The tooltip location in image coordinates is used for the later 3D position reconstruction of the tooltips. Therefore, we first validate the localization pipeline performance independently of the overall task. This includes the instrument segmentation (Sec. 4.2.3) together with the subsequent tooltip detection steps (Sec. 4.2.4–4.2.8).

For the selected bi-manual coordination task of Sec. 6, two laparoscopic instruments are used. Hence, a maximum of four tips may be encountered in any given endoscopic video frame. Two for the left and two for the right instrument. We define a bounding box around each detected tooltip. The chosen size for the bounding box is 200×200 pixels (cf. 1080p raw video frames). This corresponds to the size of the instrument distal part at a practical operation depth (Fig. 13). A comparison between the bounding box and the image size is shown in Fig. 13.

Following common practice in object detection (Everingham et al., 2015), a $\geq 50\%$ intersection over union (IoU) between the prediction and ground truth bounding boxes is considered a true positive. A predicted bounding box that does not surpass this threshold represents a false positive. The Hungarian method is employed to match predictions to ground truth bounding boxes. The number of unmatched or

missed bounding boxes from the ground truth represents the false negatives. In object detection, precision and recall at different confidence levels are commonly blended into a single performance metric, the average precision (AP) (Everingham et al., 2015). In the absence of a confidence level, we report precision and recall.

The testing set that we use to report results for the whole tooltip localization pipeline comprises 379 images. These images are evenly sampled video frames extracted at a constant frequency from the recording of the user study experiments, when participants operate the robot (i.e. they are not used during the training or validation of the segmentation model). Our tooltip tracking localization pipeline achieved a tooltip detection precision and recall of 72.45% and 61.89%, respectively. In 84.46% of the video frames, at least one of the present tips was correctly detected.

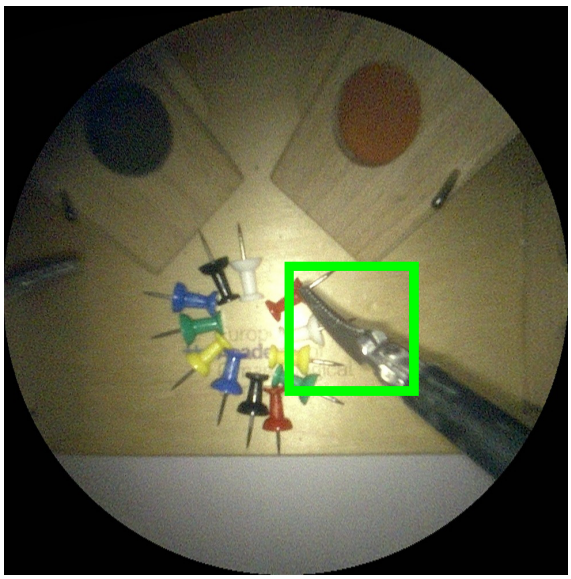

Figure 13. Image crop of visible area of 1080p endoscope video frame. The green square is the 200×200 pixel bounding box used to evaluate detection performance. An intersection over union $\geq$ **50%** between predicted and ground truth bounding boxes is considered a correct detection.

7.2 Responsiveness of the endoscopic guidance

The responsiveness of the proposed system was also evaluated. To navigate outside the view, participants have to place the tip of the instrument in the violation zone C (see Fig. 2 for a description of the zones). When this occurs, the AIT functionality is triggered until the instrument appears in zone A. Fig. 15 shows how long it took for the system to recover (entering zone A) after a violation (entering zone C) was detected. As shown in the figure, the control was responsive, taking an average of $\approx$ 3s ($\approx$ 2s in the viewing direction) to bring back the instrument tips to zone A. The slight difference between the correction time in the viewing direction and image plane is due to the difference in size of the zone A, which was relatively large along the viewing direction and therefore harder to violate.

In Fig. 15, a number of outliers are present. This occurred when the participants moved their hands too fast for the REC to follow, causing the instruments to entirely disappear from the FoV. On most of these occasions, the participants were able to put the instruments back inside FoV after some time, resuming normal navigation. However, in three instances (of the outliers $>$ 10s), the endoscope had to be manually brought back to a safe, centered home position, using its comanipulation fallback mode.

7.3 Usability of the endoscopic guidance

When a human trainee is operating the endoscope, it is important for the coordination and the overview of the surgeon that the view remains centred around the instrument. This is also the objective when a human trainee is operating the endoscope. To quantify this aspect, Fig. 16 shows the distribution of tip positions for the dominant-hand instrument across all the experiments. The REC indeed manages to keep the tooltip within the boundaries of the target zone A for most of the time. In the 2D image plane, the tip of the instrument was 46%, 23%, and 31% in target, transition, and violation zones, respectively. Similar behaviour was observed along the viewing direction, with a cumulative zone presence of 66%, 22%, and 12%, respectively.

7.4 Surgical skills assessment and learning curve on the bi-manual coordination task

The proposed system allowed the user study participants to perform the benchmark surgical task [7] with autonomous endoscope guidance within the allocated time. The completion time is shown in Fig. 14. The average completion time for the 40 trials was 172 s (only one outlier exceeding 300 s). As shown in the figure, the completion time for the *plateau novices* was relatively constant. This was not the case for *novices* and *surgeons*, where a learning curve can be appreciated despite the initial 5–10 minutes of practice. The average completion time across participants decreased from 209 s in the first attempt to 144 s in the last exercise. These results indicate that the system provided repeatable behaviour that participants were able to learn.

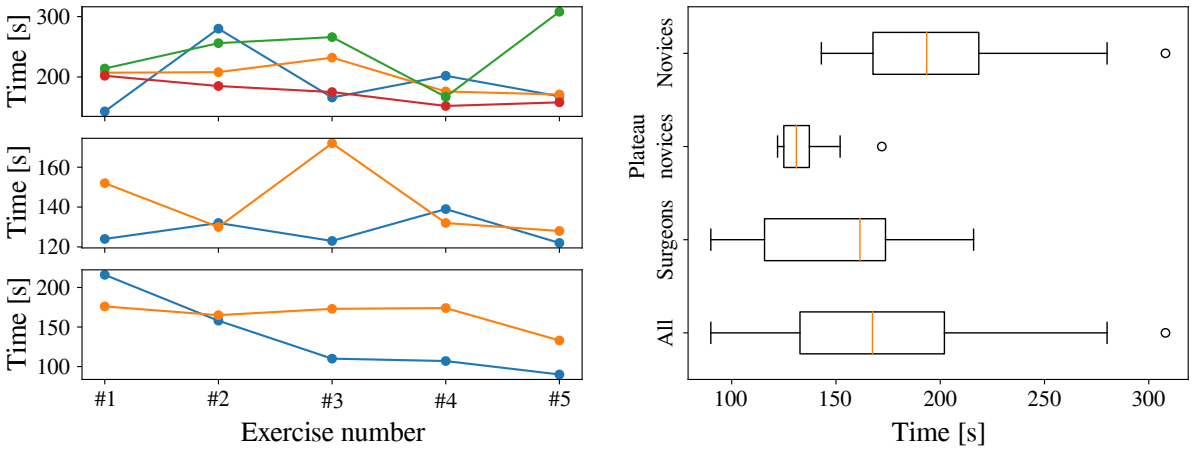

Figure 14. Completion time for all the participants in the bi-manual coordination task. (a) Completion time across attempts, with *novices* (top), *plateau novices* (centre), and *surgeons* (bottom). (b) Completion time per group across all trials.

8 CONCLUSION

In this work we proposed the use of *semantically rich instructions* to govern the interaction between a robotic autonomous endoscope holder and the operating surgeon. These are instructions such as "focus on the right tool" or "focus the camera between the instruments". This opposes previous endoscope holders handled via commands such as "move up" or "zoom in". *Semantically rich instructions* are similar to the

[7] An exemplary video is located in section "Exercise 3: Bi-manual Coordination" at `https://europeanacademy.org/training-tools/lastt/`.

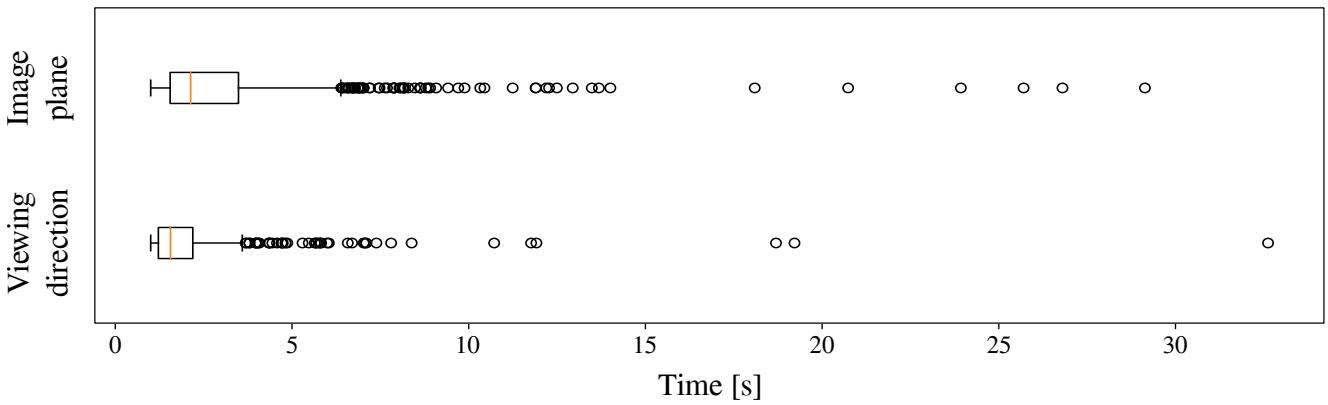

Figure 15. Time taken to correct the position of the endoscope after the dominant-hand tooltip entered the violation zone.

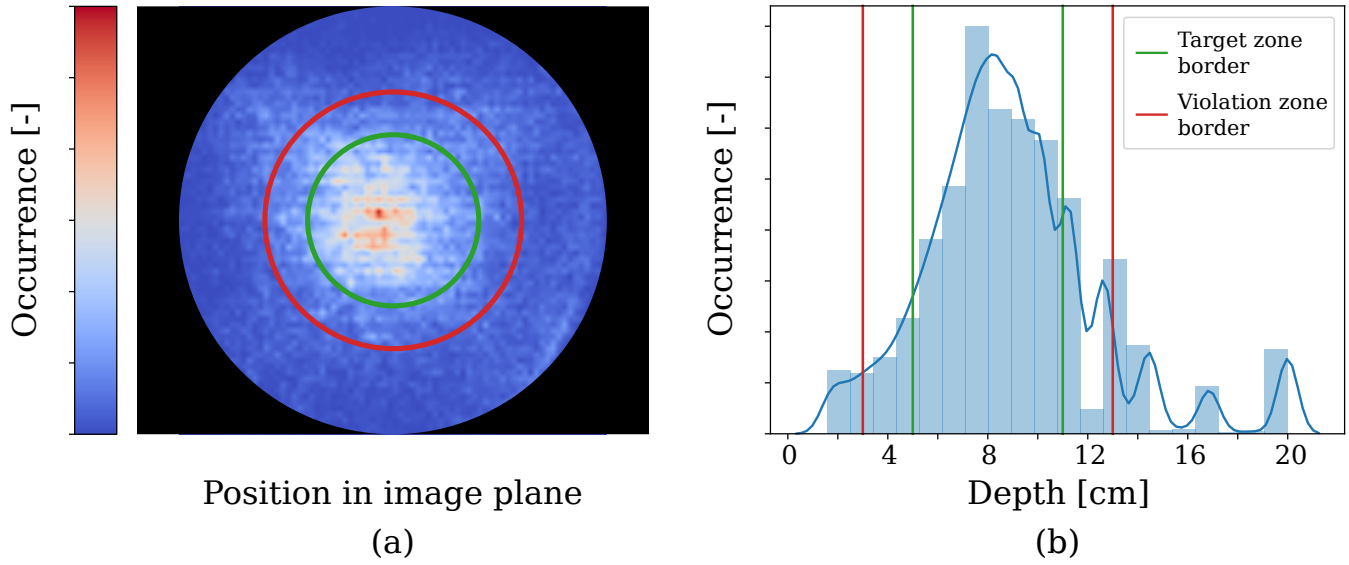

Figure 16. Distribution of dominant-hand tooltip presence across all the experiments in the 2D image and viewing direction.

instructions surgeons would issue to a human camera operator, and can therefore be naturally adopted in clinical practice. Thus, we believe that they may be a powerful tool to increase clinical acceptance.

As a first step towards implementing these instructions within a robotic endoscope holder, we concentrated our efforts on *semantically rich instructions* related to surgical instruments, which we called *autonomous instrument tracking (AIT)* instructions. To implement these instructions we built a robotic system capable of executing them without the need for additional sensors besides the endoscope. To the best of our knowledge, we are the first to report how to construct an autonomous instrument tracking system that allows for solo-surgery using only the endoscope as a sensor to track the instruments. Within the proposed system we included a novel tooltip detection method and a new visual servoing approach for a generalized endoscope model with support for remote center of motion and endoscope bending.

We found that our proposed localization method was able to detect tips in 84.46% of the frames, which in combination with our visual servoing approach allowed for a robust autonomous guidance of the endoscope. With regards to the visual servoing method, we found that a hybrid of position-based visual servoing (PBVS) and 3D image-based visual-servoing (IBVS) is preferred for robotic endoscope control.

During our experimental campaign we found that the REC-enabled AIT instructions yielded a predictable behaviour of the robotic endoscope holder that could be quickly understood and learned by the participants. The participants were able to execute a proven bi-manual coordination task within the prescribed completion time while assisted by the robotic endoscope holder. In three of the exercise runs, it was observed that the comanipulation fallback mode was required to solve for situations in which the instruments moved out of the view and the operator was unable to recover them in the view. This comanipulation mode thus ensures that failures in which the robotic endoscope holder has to be abandoned can be dealt with swiftly. An additional instruction to move back the robotic endoscope holder to a safe overview position could be considered as well. Such a safe location could for instance be close to the remote centre of motion (at the incision point). Although for the general case, when flexible instruments are used, care should be paid that such retraction does not cause the bending segment to hinge behind anatomic structures.

Besides the framework evaluation already performed, an in-depth comparison between human and robotic endoscope control remains as future work. Aspects such as time of completion, smoothness of motions, the stability of the image, number of corrections to the target zone, and average position of the instruments in the view remain to be compared. This contrast would quantify the difference in navigation quality between the proposed framework and a human-held endoscope.

While AIT instructions are necessary in most laparoscopic procedures, they are not the only instructions required for a semantic control of the endoscope holder, and it is a limitation of this study that it only focused on them. Therefore, we are positive that this work will pave the way for further developments to enlarge the set of *semantically rich instructions*.

ACKNOWLEDGMENTS

This work was supported by core and project funding from the Wellcome/EPSRC [WT203148/Z/16/Z; NS/A000049/1; WT101957; NS/A000027/1]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016985 (FAROS project). T. Vercauteren is supported by a Medtronic/Royal Academy of Engineering Research Chair [RCSRF1819\7\34].

APPENDIX

1 INVERSE KINEMATICS SOLUTION TO PBVS

The inverse kinematics problem (23) can be solved analytically to obtain $(\theta_1^*, \theta_2^*, l^*)$. This problem has four possible solutions. To select the appropriate solution, it is important that the z-axis of $\{i\}$ is defined as the inward-pointing normal of the body wall. As a first step, (23) should be rewritten as:

$$\boldsymbol{f}(\theta_1^*, \theta_2^*, l^*) = \begin{bmatrix} f_x(\theta_1^*, \theta_2^*, l^*) \\ f_y(\theta_1^*, \theta_2^*, l^*) \\ f_z(\theta_1^*, \theta_2^*, l^*) \end{bmatrix} = \boldsymbol{T}_c^{i^*} \tilde{\boldsymbol{s}}^* - \boldsymbol{T}_c^i \tilde{\boldsymbol{s}} = \boldsymbol{0}. \tag{29}$$

Next, l^* needs to be extracted from each expression in (f_x, f_y, f_z), yielding respective expressions (l_x^*, l_y^*, l_z^*). Equating $l_x^* = l_z^*$ and rewriting the result, eliminates θ_2^* and an expression of the form

$$a_1 \sin(\theta_1^*) + b_1 \cos(\theta_1^*) = c_1 \tag{30}$$

emerges, with a_1, b_1, c_1 constants. Solving this for θ_1^* yields two supplementary angles, of which the solution with the smallest absolute value should be retained. If the expression l_y^* is substituted in f_x and f_z, and both are squared and added according to:

$$f_x(\theta_1^*, \theta_2^*, l_y^*)^2 + f_z(\theta_1^*, \theta_2^*, l_y^*)^2 = 0, \tag{31}$$

the dependence on θ_1^* cancels out. Simplifying this equation leads to:

$$a_2 \cos^2(\theta_2^*) + b_2 \cos(\theta_2^*) + c_2 = 0, \tag{32}$$

with a_2, b_2, c_2 constants. This is a quadratic equation in $\cos(\theta_2^*)$. The solution with the smallest $|\theta_2^*|$ is to be retained, but the sign of θ_2^* still needs to be confirmed. It is now possible to determine l^*, by plugging the known θ_1^* and $|\theta_2^*|$ into one of the expressions (f_x, f_y, f_z). For numerical stability, f_y should be used if $|\sin(\theta_2^*)| > \frac{1}{2}$, f_x if $|\sin(\theta_1^*)| > \frac{1}{2}$, and f_z otherwise. As the final step, the two unused expressions within (f_x, f_y, f_z) need to be evaluated to determine the sign of θ_2^*. If they do not evaluate to 0, θ_2^* has to be negative and l^* needs to be recomputed.

REFERENCES

Abdi, E., Burdet, E., Bouri, M., Himidan, S., Bleuler, H., 2016. In a demanding task, three-handed manipulation is preferred to two-handed manipulation. Scientific Reports 6, 21758. URL: http://www.nature.com/articles/srep21758, doi:10.1038/srep21758.

Aertbeliën, E., De Schutter, J., 2014. etasl/etc: A constraint-based task specification language and robot controller using expression graphs, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1540–1546.

Agustinos, A., Wolf, R., Long, J.A., Cinquin, P., Voros, S., 2014. Visual servoing of a robotic endoscope holder based on surgical instrument tracking, in: 5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, IEEE. pp. 13–18. URL: https://ieeexplore.ieee.org/document/6913744, doi:10.1109/BIOROB.2014.6913744.

Ali, J.M., Lam, K., Coonar, A.S., 2018. Robotic Camera Assistance: The Future of Laparoscopic and Thoracoscopic Surgery? Surgical Innovation 25, 485–491. URL: http://journals.sagepub.com/doi/10.1177/1553350618784224, doi:10.1177/1553350618784224.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., Herrera, L., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M., 2019. 2017 Robotic Instrument Segmentation Challenge. Arxiv URL: http://arxiv.org/abs/1902.06426, arXiv:1902.06426.

Amin, M.S.A., Aydin, A., Abbud, N., Van Cleynenbreugel, B., Veneziano, D., Somani, B., Gözen, A.S., Redorta, J.P., Khan, M.S., Dasgupta, P., Makanjuoala, J., Ahmed, K., 2020. Evaluation of a remote-controlled laparoscopic camera holder for basic laparoscopic skills acquisition: a randomized controlled trial. Surgical Endoscopy URL: http://link.springer.com/10.1007/s00464-020-07899-5, doi:10.1007/s00464-020-07899-5.

Avellino, I., Bailly, G., Arico, M., Morel, G., Canlorbe, G., 2020. Multimodal and Mixed Control of Robotic Endoscopes, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing

Systems, ACM, New York, NY, USA. pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3313831.3376795, doi:10.1145/3313831.3376795.

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7700 LECTU, 437–478. URL: http://arxiv.org/abs/1206.5533, doi:10.1007/978-3-642-35289-8-26, arXiv:1206.5533.

Bihlmaier, A., 2016. Endoscope Robots and Automated Camera Guidance, in: Learning Dynamic Spatial Relations. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 23–102. URL: http://link.springer.com/10.1007/978-3-658-14914-7{_}2, doi:10.1007/978-3-658-14914-7_2.

Bouarfa, L., Akman, O., Schneider, A., Jonker, P.P., Dankelman, J., 2012. Real-Time Tracking of Surgical Instruments in Endoscopic Video. Minimally Invasive Therapy & Allied Technologies 21, 129–134. doi:10.3109/13645706.2011.580764.

Bouget, D., Allan, M., Stoyanov, D., Jannin, P., 2017. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. MedIA 35, 633–654. doi:10.1016/j.media.2016.09.003.

Bruner, J.P., 1999. Fetal Surgery for Myelomeningocele and the Incidence of Shunt-Dependent Hydrocephalus. JAMA 282, 1819. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.282.19.1819, doi:10.1001/jama.282.19.1819.

Campo, R., Molinas, C.R., De Wilde, R.L., Brolmann, H., Brucker, S., Mencaglia, L., Odonovan, P., Wallwiener, D., Wattiez, A., 2012. Are you good enough for your patients? The European certification model in laparoscopic surgery. Facts, views & vision in ObGyn 4, 95–101. URL: http://www.ncbi.nlm.nih.gov/pubmed/24753896http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3987500.

Casals, a., Amat, J., Laporte, E., 1996. Automatic guidance of an assistant robot in laparoscopic surgery, in: Proceedings of IEEE International Conference on Robotics and Automation, IEEE. pp. 895–900. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=503886, doi:10.1109/ROBOT.1996.503886.

Chaumette, F., Hutchinson, S., 2008. Visual servoing and visual tracking, in: Handbook of Robotics, pp. 563–583.

Chen, P.J., Lin, M.C., Lai, M.J., Lin, J.C., Lu, H.H.S., Tseng, V.S., 2018. Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. Gastroenterology 154, 568–575. URL: https://linkinghub.elsevier.com/retrieve/pii/S0016508517362510, doi:10.1053/j.gastro.2017.10.010.

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

Da Col, T., Caccianiga, G., Catellani, M., Mariani, A., Ferro, M., Cordima, G., De Momi, E., Ferrigno, G., de Cobelli, O., 2021. Automating Endoscope Motion in Robotic Surgery: A Usability Study on da Vinci-Assisted Ex Vivo Neobladder Reconstruction. Frontiers in Robotics and AI 8. URL: https://www.frontiersin.org/articles/10.3389/frobt.2021.707704/full, doi:10.3389/frobt.2021.707704.

Dong, L., Morel, G., 2016. Robust trocar detection and localization during robot-assisted endoscopic surgery, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 4109–4114.

Eslamian, S., Reisner, L.A., King, B.W., Pandya, A.K., 2016. Towards the Implementation of an Autonomous Camera Algorithm on the da Vinci Platform. Studies in health technology and informatics 220, 118–23. URL: http://www.ncbi.nlm.nih.gov/pubmed/27046563.

Eslamian, S., Reisner, L.A., Pandya, A.K., 2020. Development and evaluation of an autonomous camera control algorithm on the da vinci surgical system. The International Journal of Medical Robotics and Computer Assisted Surgery 16, e2036.

European Academy of Gynaecological Surgery, 2020. LASTT. URL: https://europeanacademy.org/training-tools/lastt/.

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2015. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision 111, 98–136. doi:10.1007/s11263-014-0733-5.

Fuentes-Hurtado, F., Kadkhodamohammadi, A., Flouty, E., Barbarisi, S., Luengo, I., Stoyanov, D., 2019. EasyLabels: weak labels for scene segmentation in laparoscopic videos. International Journal of Computer Assisted Radiology and Surgery 14, 1247–1257. doi:10.1007/s11548-019-02003-2.

Fujii, K., Gras, G., Salerno, A., Yang, G.Z., 2018. Gaze gesture based human robot interaction for laparoscopic surgery. Medical Image Analysis 44, 196–214. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841517301809, doi:10.1016/j.media.2017.11.011.

Garcia-Peraza-Herrera, L.C., Fidon, L., DrEttorre, C., Stoyanov, D., Vercauteren, T., Ourselin, S., 2021. Image Compositing for Segmentation of Surgical Tools without Manual Annotations. IEEE Transactions on Medical Imaging URL: https://ieeexplore.ieee.org/document/9350303/, doi:10.1109/TMI.2021.3057884.

Garcia-Peraza-Herrera, L.C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.V., Stoyanov, D., Vercauteren, T., Ourselin, S., 2017. ToolNet: Holistically-nested real-time segmentation of robotic surgical tools, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 5717–5722. URL: http://ieeexplore.ieee.org/document/8206462/, doi:10.1109/IROS.2017.8206462, arXiv:1706.08126.

García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-Time Segmentation of Non-rigid Surgical Tools Based on Deep Learning and Tracking, in: CARE workshop, held in conjunction with MICCAI 2016, pp. 84–95. doi:10.1007/978-3-319-54057-3_8.

Gillen, S., Pletzer, B., Heiligensetzer, A., Wolf, P., Kleeff, J., Feussner, H., Fürst, A., 2014. Solo-surgical laparoscopic cholecystectomy with a joystick-guided camera device: a case–control study. Surgical Endoscopy 28, 164–170. URL: http://link.springer.com/10.1007/s00464-013-3142-x, doi:10.1007/s00464-013-3142-x.

González, C., Bravo-Sánchez, L., Arbelaez, P., 2020. ISINet: An Instance-Based Approach for Surgical Instrument Segmentation, in: MICCAI, pp. 595–605. URL: http://arxiv.org/abs/2007.05533https://link.springer.com/10.1007/978-3-030-59716-0{_}57, doi:10.1007/978-3-030-59716-0_57, arXiv:2007.05533.

Goodell, K.H., Cao, C.G., Schwaitzberg, S.D., 2006. Effects of Cognitive Distraction on Performance of Laparoscopic Surgical Tasks. Journal of Laparoendoscopic & Advanced Surgical Techniques 16, 94–98. URL: http://www.liebertpub.com/doi/10.1089/lap.2006.16.94, doi:10.1089/lap.2006.16.94.

Gruijthuijsen, C., Dong, L., Morel, G., Vander Poorten, E., 2018. Leveraging the fulcrum point in robotic minimally invasive surgery. IEEE Robotics and Automation Letters 3, 2071–2078.

945 Hanna, G.B., Shimi, S., Cuschieri, A., 1997. Influence of direction of view, target-to-endoscope distance
946 and manipulation angle on endoscopic knot tying. The British journal of surgery 84, 1460–4. URL:
947 http://www.ncbi.nlm.nih.gov/pubmed/9361614.
948 Holländer, S.W., Klingen, H.J., Fritz, M., Djalali, P., Birk, D., 2014. Robotic Camera Assistance and Its
949 Benefit in 1033 Traditional Laparoscopic Procedures: Prospective Clinical Trial Using a Joystick-guided
950 Camera Holder. Surgical technology international 25, 19–23. URL: http://www.ncbi.nlm.nih.
951 gov/pubmed/25419950.
952 Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing
953 Internal Covariate Shift. Arxiv .
954 Jaspers, J.E.N., Breedveld, P., Herder, J.L., Grimbergen, C.A., 2004. Camera and instrument holders and
955 their clinical value in minimally invasive surgery. Surgical laparoscopy, endoscopy & percutaneous
956 techniques 14, 145–52. URL: http://www.ncbi.nlm.nih.gov/pubmed/15471021, doi:10.
957 1097/01.sle.0000129395.42501.5d.
958 Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014.
959 Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 .
960 Kabagambe, S.K., Jensen, G.W., Chen, Y.J., Vanover, M.A., Farmer, D.L., 2018. Fetal Surgery for
961 Myelomeningocele: A Systematic Review and Meta-Analysis of Outcomes in Fetoscopic versus Open
962 Repair. Fetal Diagnosis and Therapy 43, 161–174. URL: https://www.karger.com/Article/
963 FullText/479505, doi:10.1159/000479505.
964 King, B.W., Reisner, L.A., Pandya, A.K., Composto, A.M., Ellis, R.D., Klein, M.D., 2013. Towards
965 an autonomous robot for camera control during laparoscopic surgery. Journal of laparoendoscopic &
966 advanced surgical techniques. Part A 23, 1027–30. URL: http://www.ncbi.nlm.nih.gov/
967 pubmed/24195784, doi:10.1089/lap.2013.0304.
968 Kommu, S.S., Rimington, P., Anderson, C., Rané, A., 2007. Initial experience with the En-
969 doAssist camera-holding robot in laparoscopic urological surgery. Journal of Robotic Surgery 1,
970 133–137. URL: http://link.springer.com/10.1007/s11701-007-0010-5, doi:10.
971 1007/s11701-007-0010-5.
972 Kunze, L., Roehm, T., Beetz, M., 2011. Towards semantic robot description languages, in: 2011
973 IEEE International Conference on Robotics and Automation, IEEE. pp. 5589–5595. URL: http:
974 //ieeexplore.ieee.org/document/5980170/, doi:10.1109/ICRA.2011.5980170.
975 Kwon, D.S., Ko, S.Y., Kim, J., 2008. Intelligent Laparoscopic Assistant Robot through Surgery Task
976 Model: How to Give Intelligence to Medical Robots, in: Medical Robotics. I-Tech Education and Publish-
977 ing, pp. 197–218. URL: http://www.intechopen.com/books/medical{_}robotics/
978 intelligent{_}laparoscopic{_}assistant{_}robot{_}through{_}surgery{_}task{_
979 doi:10.5772/5249.
980 Lee, G., Lee, T., Dexter, D., Godinez, C., Meenaghan, N., Catania, R., Park, A., 2009. Er-
981 gonomic risk associated with assisting in minimally invasive surgery. Surgical Endoscopy 23,
982 182–188. URL: http://link.springer.com/10.1007/s00464-008-0141-4, doi:10.
983 1007/s00464-008-0141-4.
984 Lee, T., Kashyap, R., Chu, C., 1994. Building Skeleton Models via 3-D Medial Surface Axis Thinning
985 Algorithms. CVGIP: Graphical Models and Image Processing 56, 462–478.
986 Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M., 2020. Dense Depth
987 Estimation in Monocular Endoscopy With Self-Supervised Learning Methods. IEEE Transactions
988 on Medical Imaging 39, 1438–1447. URL: https://ieeexplore.ieee.org/document/
989 8889760/, doi:10.1109/TMI.2019.2950936.

Mariani, A., Colaci, G., Da Col, T., Sanna, N., Vendrame, E., Menciassi, A., De Momi, E., 2020. An Experimental Comparison Towards Autonomous Camera Navigation to Optimize Training in Robot Assisted Surgery. IEEE Robotics and Automation Letters 5, 1461–1467. URL: https://ieeexplore.ieee.org/document/8954796/, doi:10.1109/LRA.2020.2965067.

Meuli, M., Moehrlen, U., 2014. Fetal surgery for myelomeningocele is effective: a critical look at the whys. Pediatric Surgery International 30, 689–697. URL: http://link.springer.com/10.1007/s00383-014-3524-8, doi:10.1007/s00383-014-3524-8.

Mudunuri, A.V., 2010. Autonomous camera control system for surgical robots. Ph.D. thesis. Wayne State University.

Nishikawa, A., Nakagoe, H., Taniguchi, K., Yamada, Y., Sekimoto, M., Takiguchi, S., Monden, M., Miyazaki, F., 2008. How Does the Camera Assistant Decide the Zooming Ratio of Laparoscopic Images? Analysis and Implementation, in: MICCAI, pp. 611–618. URL: http://link.springer.com/10.1007/978-3-540-85990-1{_}73, doi:10.1007/978-3-540-85990-1_73.

Osa, T., Staub, C., Knoll, A., 2010. Framework of automatic robot surgery system using Visual servoing, in: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 1837–1842. URL: http://ieeexplore.ieee.org/document/5650301/, doi:10.1109/IROS.2010.5650301.

Pakhomov, D., Shen, W., Navab, N., 2020. Towards Unsupervised Learning for Instrument Segmentation in Robotic Surgery with Cycle-Consistent Adversarial Networks, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 8499–8504. URL: http://arxiv.org/abs/2007.04505https://ieeexplore.ieee.org/document/9340816/, doi:10.1109/IROS45743.2020.9340816, arXiv:2007.04505.

Pandya, A., Reisner, L., King, B., Lucas, N., Composto, A., Klein, M., Ellis, R., 2014. A Review of Camera Viewpoint Automation in Robotic and Laparoscopic Surgery. Robotics 3, 310–329. URL: http://www.mdpi.com/2218-6581/3/3/310, doi:10.3390/robotics3030310.

Platte, K., Alleblas, C.C., Inthout, J., Nieboer, T.E., 2019. Measuring fatigue and stress in laparoscopic surgery: validity and reliability of the star-track test. Minimally Invasive Therapy & Allied Technologies 28, 57–64. URL: https://www.tandfonline.com/doi/full/10.1080/13645706.2018.1470984, doi:10.1080/13645706.2018.1470984.

Polski, M., Fiolka, A., Can, S., Schneider, A., Feussner, H., 2009. A new partially autonomous camera control system, in: World Congress on Medical Physics and Biomedical Engineering, pp. 276–277. URL: http://link.springer.com/10.1007/978-3-642-03906-5{_}75, doi:10.1007/978-3-642-03906-5_75.

Rahman, M.A., Wang, Y., 2016. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation, in: Advances in Visual Computing, pp. 234–244. URL: http://link.springer.com/10.1007/978-3-319-50835-1{_}22, doi:10.1007/978-3-319-50835-1_22.

Reiter, A., Goldman, R.E., Bajo, A., Iliopoulos, K., Simaan, N., Allen, P.K., 2011. A learning algorithm for visual pose estimation of continuum robots, in: IROS, IEEE. pp. 2390–2396. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6048634, doi:10.1109/IROS.2011.6048634.

Rivas-Blanco, I., Estebanez, B., Cuevas-Rodriguez, M., Bauzano, E., Munoz, V.F., 2014. Towards a cognitive camera robotic assistant, in: 5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, IEEE. pp. 739–744. URL: https://ieeexplore.ieee.org/document/6913866, doi:10.1109/BIOROB.2014.6913866.

Rodrigues Armijo, P., Huang, C.K., Carlson, T., Oleynikov, D., Siu, K.C., 2020. Ergonomics Analysis for Subjective and Objective Fatigue Between Laparoscopic and Robotic Surgical Skills Practice Among Surgeons. Surgical Innovation 27, 81–87. URL: http://journals.sagepub.com/doi/10.1177/1553350619887861, doi:10.1177/1553350619887861.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: MICCAI, pp. 234–241. URL: http://link.springer.com/10.1007/978-3-319-24574-4{_}28, doi:10.1007/978-3-319-24574-4_28.

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.B., Bodenstedt, S., Bolmgren, J.L., Bravo-Sánchez, L., Chen, H.B., González, C., Guo, D., Halvorsen, P., Heng, P.A., Hosgor, E., Hou, Z.G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.L., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y.J., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B.P., Maier-Hein, L., 2021. Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge. Medical Image Analysis 70, 101920. doi:10.1016/j.media.2020.101920.

Samei, G., Tsang, K., Kesch, C., Lobo, J., Hor, S., Mohareri, O., Chang, S., Goldenberg, S.L., Black, P.C., Salcudean, S., 2020. A partial augmented reality system with live ultrasound and registered preoperative MRI for guiding robot-assisted radical prostatectomy. Medical Image Analysis 60, 101588. doi:10.1016/j.media.2019.101588.

Sandoval, J., med amine Laribi, Faure, J.P., Breque, C., Richer, J.P., Zeghloul, S., 2021. Towards an Autonomous Robot-Assistant for Laparoscopy Using Exteroceptive Sensors: Feasibility Study and Implementation. IEEE Robotics and Automation Letters 6, 6473–6480. URL: https://ieeexplore.ieee.org/document/9477029/, doi:10.1109/LRA.2021.3094644.

Seong-Young Ko, Kim, J., Dong-Soo Kwon, Woo-Jung Lee, 2005. Intelligent interaction between surgeon and laparoscopic assistant robot system, in: ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005., pp. 60–65. doi:10.1109/ROMAN.2005.1513757.

Song, C., Gehlbach, P.L., Kang, J.U., 2012. Active tremor cancellation by a "smart" handheld vitreoretinal microsurgical tool using swept source optical coherence tomography. Optics express 20, 23414–21. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3601638{&}tool=pmcentrez{&}rendertype=abstract.

Song, K.T., Chen, C.J., 2012. Autonomous and stable tracking of endoscope instrument tools with monocular camera, in: 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), IEEE. pp. 39–44. URL: http://ieeexplore.ieee.org/document/6266023/, doi:10.1109/AIM.2012.6266023.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15, 1929–1958.

Stott, M.C., Barrie, J., Sebastien, D., Hammill, C., Subar, D.A., 2017. Is the Use of a Robotic Camera Holder Economically Viable? A Cost Comparison of Surgical Assistant Versus the Use of a Robotic Camera Holder in Laparoscopic Liver Resections. Surgical Laparoscopy, Endoscopy & Percutaneous Techniques 27, 375–378. URL: http://journals.lww.com/00129689-201710000-00012, doi:10.1097/SLE.0000000000000452.

Stoyanov, D., 2012. Surgical Vision. Annals of Biomedical Engineering 40, 332–345. URL: http://link.springer.com/10.1007/s10439-011-0441-z, doi:10.1007/s10439-011-0441-z.

Takahashi, M., 2020. Solo surgery with VIKY: Safe, simple, and low-cost robotic surgery, in: Abedin-Nasab, M.H. (Ed.), Handbook of Robotic and Image-Guided Surgery. Elsevier, pp. 79–88. doi:10.1016/B978-0-12-814245-5.00005-0.

Taniguchi, K., Nishikawa, A., Sekimoto, M., Kobayashi, T., Kazuhara, K., Ichihara, T., Kurashita, N., Takiguchi, S., Doki, Y., Mori, M., Miyazaki, F., 2010. Classification, Design and Evaluation of Endoscope Robots, in: Robot Surgery. InTech, pp. 1–24. URL: http://www.intechopen.com/books/robot-surgery/classification-design-and-evaluation-of-endoscope-robots, doi:10.5772/6893.

Tonet, O., Thoranaghatte, R.U., Megali, G., Dario, P., 2007. Tracking endoscopic instruments without a localizer: a shape-analysis-based approach. Computer aided surgery : official journal of the International Society for Computer Aided Surgery 12, 35–42. doi:10.3109/10929080701210782.

Uecker, D.R., Lee, C., Wang, Y.F., Wang, Y., 1995. Automated instrument tracking in robotically assisted laparoscopic surgery. Journal of image guided surgery 1, 308–25. URL: http://www.ncbi.nlm.nih.gov/pubmed/9080352, doi:10.1002/(SICI)1522-712X(1995)1:6<308::AID-IGS3>3.0.CO;2-E.

Uenohara, M., Kanade, T., 1995. Vision-based object registration for real-time image overlay. Computers in Biology and Medicine 25, 249–260. doi:10.1016/0010-4825(94)00045-R.

Vardazaryan, A., Mutter, D., Marescaux, J., Padoy, N., 2018. Weakly-Supervised Learning for Tool Localization in Laparoscopic Videos, in: LABELS, pp. 169–179. doi:10.1007/978-3-030-01364-6_19.

Wagner, M., Bihlmaier, A., Kenngott, H.G., Mietkowski, P., Scheikl, P.M., Bodenstedt, S., Schiepe-Tiska, A., Vetter, J., Nickel, F., Speidel, S., et al., 2021. A learning robot for cognitive camera control in minimally invasive surgery. Surgical Endoscopy , 1–10.

Wang, Y.F., Uecker, D.R., Wang, Y., 1998. A new framework for vision-enabled and robotically assisted minimally invasive surgery. Computerized Medical Imaging and Graphics 22, 429–437. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895611198000524, doi:10.1016/S0895-6111(98)00052-4.

Wauben, L.S.G.L., van Veelen, M.A., Gossot, D., Goossens, R.H.M., 2006. Application of ergonomic guidelines during minimally invasive surgery: a questionnaire survey of 284 surgeons. Surgical Endoscopy And Other Interventional Techniques 20, 1268–1274. URL: http://link.springer.com/10.1007/s00464-005-0647-y, doi:10.1007/s00464-005-0647-y.

Weede, O., Monnich, H., Muller, B., Worn, H., 2011. An intelligent and autonomous endoscopic guidance system for minimally invasive surgery, in: 2011 IEEE International Conference on Robotics and Automation, IEEE. pp. 5762–5768. URL: http://ieeexplore.ieee.org/document/5980216/, doi:10.1109/ICRA.2011.5980216.

Wijsman, P.J.M., Broeders, I.A.M.J., Brenkman, H.J., Szold, A., Forgione, A., Schreuder, H.W.R., Consten, E.C.J., Draaisma, W.A., Verheijen, P.M., Ruurda, J.P., Kaufman, Y., 2018. First experience with THE AUTOLAP™ SYSTEM: an image-based robotic camera steering device. Surgical Endoscopy 32, 2560–2566. URL: http://link.springer.com/10.1007/s00464-017-5957-3, doi:10.1007/s00464-017-5957-3.

Wijsman, P.J.M., Molenaar, L., Voskens, F.J., van't Hullenaar, C.D.P., Broeders, I.A.M.J., 2022. Image-based laparoscopic camera steering versus conventional steering: a comparison study. Journal of Robotic Surgery URL: https://link.springer.com/10.1007/s11701-021-01342-0, doi:10.1007/s11701-021-01342-0.

Xiaolong, Y., 2019. Image-Py skeleton network module. URL: https://github.com/Image-Py/sknw.

Yang, G.Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P.E., Hata, N., Kazanzides, P., Martel, S., Patel, R.V., Santos, V.J., Taylor, R.H., 2017. Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy. Science Robotics 2, eaam8638. URL: https://robotics.sciencemag.org/lookup/doi/10.1126/scirobotics.aam8638, doi:10.1126/scirobotics.aam8638.

Yu, L., Wang, Z., Sun, L., Wang, W., Wang, T., 2013. A kinematics method of automatic visual window for Laparoscopic Minimally Invasive Surgical Robotic System, in: 2013 IEEE International Conference on Mechatronics and Automation, IEEE. pp. 997–1002. URL: http://ieeexplore.ieee.org/document/6618051/, doi:10.1109/ICMA.2013.6618051.

Zhang, X., Payandeh, S., 2002. Application of visual tracking for robot-assisted laparoscopic surgery. J. Robotic Systems 19, 315–328. doi:10.1002/rob.10043.

Zhao, Z., 2014. Real-time 3D visual tracking of laparoscopic instruments for robotized endoscope holder, in: Proceeding of the 11th World Congress on Intelligent Control and Automation, IEEE. pp. 6145–6150. URL: http://ieeexplore.ieee.org/document/7053773/, doi:10.1109/WCICA.2014.7053773.

Zinchenko, K., Song, K.T., 2021. Autonomous Endoscope Robot Positioning Using Instrument Segmentation With Virtual Reality Visualization. IEEE Access 9, 72614–72623. URL: https://ieeexplore.ieee.org/document/9429186/, doi:10.1109/ACCESS.2021.3079427.