UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF

Computer Science

GRADUATE SCHOOL OF

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Computer Science

XXXIV cycle

TITOLO DELLA TESI DI DOTTORATO

Medical SLAM in an autonomous robotic system

S.S.D. INF/01

Coordinator: Prof. Prof. Paolo Fiorini

Firma _____

Doctoral Student: Dott.Roberti Andrea

Firma _____

***** Per l'elenco dei Settori Scientifico-Disciplinari (SSD) si veda il D.M. del 4 Ottobre 2000, Allegato A *"Elenco dei Settori Scientifico –Disciplinari"* reperibile sul sito del Ministero dell'Università e della Ricerca al seguente indirizzo: http://www.miur.it/atti/2000/alladm001004_01.htm

One of the main challenges for computer-assisted surgery (CAS) is to determine the intra-operative morphology and motion of soft-tissues. This information is prerequisite to the registration of multi-modal patient-specific data for enhancing the surgeon's navigation capabilities by observing beyond exposed tissue surfaces and for providing intelligent control of robotic-assisted instruments. In minimally invasive surgery (MIS), optical techniques are an increasingly attractive approach for in vivo 3D reconstruction of the soft-tissue surface geometry. This thesis addresses the ambitious goal of achieving surgical autonomy, through the study of the anatomical environment by Initially studying the technology present and what is needed to analyze the scene: vision sensors.

A novel endoscope for autonomous surgical task execution is presented in the first part of this thesis. Which combines a standard stereo camera with a depth sensor. This solution introduces several key advantages, such as the possibility of reconstructing the 3D at a greater distance than traditional endoscopes. Then the problem of hand-eye calibration is tackled, which unites the vision system and the robot in a single reference system. Increasing the accuracy in the surgical work plan.

In the second part of the thesis the problem of the 3D reconstruction and the algorithms currently in use were addressed. In MIS, simultaneous localization and mapping (SLAM) can be used to localize the pose of the endoscopic camera and build ta 3D model of the tissue surface. Another key element for MIS is to have real-time knowledge of the pose of surgical tools with respect to the surgical camera and underlying anatomy. Starting from the ORB-SLAM algorithm we have modified the architecture to make it usable in an anatomical environment by adding the registration of the pre-operative information of the intervention to the map obtained from the SLAM. Once it has been proven that the slam algorithm is usable in an anatomical environment, it has been improved by adding semantic segmentation to be able to distinguish dynamic features from static ones.

All the results in this thesis are validated on training setups, which mimics some of the challenges of real surgery and on setups that simulate the human body within *Autonomous Robotic Surgery (ARS)* and *Smart Autonomous Robotic Assistant Surgeon (SARAS)* projects.

Contents

List of Figures

List of Tables

1

Introduction

An *autonomous system* is an artificially intelligent entity that makes decisions in response to input, independent of human interaction. *robotic systems* are physical entities that interact with the physical world. In this thesis, we consider an autonomous robotic system as a device that uses *artificial Intelligence (AI)* and has a physical presence in and interacts with the real world. These systems are complex, inherently hybrid, systems, combining both hardware and software; they often require in depth safety, legal, and ethical analysys. Autonomous robotics are increasingly being used in commonplace scenarios, such as driverless cars[47], pilotless aircraft[206], and domestic assistants[36, 205].

For many engineered systems, extensive tests, either through real deployment or via simulation, are deemed sufficient for the validation and evaluation of the developed algorithms. However, the unique challenges of autonomous robotics, their dependence on sophisticated software control and decision making, and their increasing deployment in safety-critical scenarios require a stronger form of verification.

Autonomous robotic systems lack formal specification and verification methods and often have safety-critical behaviors. A survey on safety-critical robotics[63] identified seven focus areas for the development of robots that can safely work alongside humans and other robots, in unstructured environments. These areas are:

- modeling and simulation of the physical environment to enable better safety analysis;
- formal verification of robot systems;
- models of human-robot interaction;
- controllers that are correct by construction;
- identification and monitoring of hazardous situations;
- online safety monitoring that adapts to the robot's context;
- certification evidence.

In this thesis, we focus more on the first three points by developing and using Simultaneous Localization and Mapping (SLAM) [42] [191], which is a technique for obtaining the 3D structure an unknown environment and for estimating sensor motion. The use of this method allows to compute precisely

the relative position of the robotic devices with respect to their environment and this guarantee the robotic mitions are safe by design.

1.1 Modeling the physical environment

We consider an autonomous robot as a device that implements AI techniques, has a physical presence, and interacts with the world. Thus, one of the most prominent challenges in controlling an autonomous robotic system is to verify its interaction with an unstructured environment. Interactions between a robot and its physical environment has a major influence on its behavior since the robot can react to changes in environmental conditions. Indeed, the behavior of adaptive systems is directly driven by environmental interactions, such as touch or object recognition. While it is accepted that formally modeling a robotic system within its physical environment is important[52, 199], it has had limited attention in the literature.

Apart from the difficulty of modeling the real world, a robot only has partial knowledge of its surroundings. To help robots learn more about their surroundings, they often come equipped with sensors that are used to estimate a robot's condition and its relation with the environment. Their signals are passed to a controller to enable appropriate behavior.

Sensors in robots replicate the functions of human sensory organs. Robots require extensive information about their environment in order to function effectively. But sensing limitations, caused by sensor blind-spots and interference between sensors, add extra complexity to the modeling process[12, 71, 147].

To address the challenge of combining discrete computations and a continuous environment, a robotic system is typically separated into several layers[2, 52]. At the bottom, the functional layer consists of control software for the robot's hardware. Then, the intermediate layer generally utilizes a middleware framework (such as ROS[154]) that provides an interface to the hardware components. The upper layer contains the decision-making components of the robotic system, which capture its autonomous behavior.

Some previous work has focused on a robot's decision making, ignoring its environment [85, 103]. Others assume that the environment is static and known, prior to the robot's deployment [56, 120, 205], which is often neither possible nor feasible[71].

For example, the environment may contain both fixed and mobile objects whose future behavior is unknown[12], or the robot's goal may be to map the environment, so the layout is unknown.

Other approaches abtract away from the environment and assume that a component has the ability to provide predicates that represent the continuous sensor data about the robot's environment [32, 33, 52, 189, 192].

While this encapsulates the high-level control, insulating it from the environment, the implicit assumption of a static, known environment often remains. Using models of the environment that are static and assume prior knowledge may limit their effectiveness. However, Sotiropoulos et al.[181] examine the ability of low-fidelity environmental simulations to reproduce bugs

that were found during field tests of the same robot. Of the 33 errors that occured during the field test, only one could not be reproduced in the low-fidelity simulation. Perhaps similar results might be obtained with low-fidelity formal models of a robot's environment.

1.1.1 Probabilistic models

Probabilistic models are a popular approach to capturing dynamic and uncertain environments. In[121], a *Probabilistic symbolic model checker* (PRISM) model captures the environment of a domestic assistant robot. Nonrobot actors in the environment are specified using probabilistic behaviors, to represent the uncertainty about their behavior. The robot model is also written in PRISM, so that it can be reasoned about within this probabilistic environment. This is a useful step that accepts the changing nature of the environment. However, for the model checker to explore outcomes based on a certain behavior, its probability must be encoded into the environment model. This still leaves the possibility of unforeseen situations having a detrimental effect on the robot's behavior.

Similarly, Hoffmann et al.[72] formalize a pilotless aircraft and its physical environment using an extension of Probabilistic Finite-State Machines (PFSMs). They use PRISM to verify properties about their model, with the pilotless aircraft tasked with foraging for objects, which it must return to a drop-off location. Their approach involves running small-scale simulations of the pilotless aircraft in its environment to determine the probabilities and timing values for their formal model.

1.1.2 Hybrid models

A hybrid system is composed of a robot and dynamic obstacles in the unknown environment. Robots make discrete control choices (e.g., compute the actuator set values for acceleration, braking, or steering), which in turn influence their actual physical behavior (e.g., slow down to a stop, move along a curve). Hybrid systems have been considered as joint models for both components, since verification of either component alone does not capture the full behavior of a robot and its environment.

Modeling the environment is particularly relevant for navigation, and tackling collision avoidance and safe robot navigation[12, 118, 147, 153] often feature in the literature. Many navigation algorithms have been proposed for autonomous mobile robots. Few of these algorithms, however, have been verified to ensure the safety of the robot[146]. One consequence of this situation is that potentially superior but uncertified navigation algorithms are not deployed in safety-critical applications. The Simplex architecture[170] provides a gateway to using these unverified algorithms with the concept of advanced controllers (ACs). ACs are used alongside a verified baseline controller (BC) and a decision module, which chooses what controller is active. The decision module monitors the system, and if it determines that the AC will violate one

of the safety properties, then the BC takes control. Other work employs reachability analysis to generate a maneuver automaton for collision avoidance of road vehicles[71]. Here, differential equations model the continuous behaviour of the system. Runtime fault monitoring is useful for catching irregular behaviors in running systems, but it does not ensure safety in all situations. Mitsch et al.[118] use differential dynamic logic, designed for hybrid systems, to describe the discrete and continuous navigation behavior of a ground robot. Their approach uses hybrid programs for modeling a robot that follows the dynamic window algorithm and for modeling the behavior of moving objects in the environment. Using the hybrid theorem prover KeYmaera, Mitsch et al.[118] verify that the robot will not collide with, and maintains a sufficient distance from, stationary and moving obstacles. In proving these safety properties, 85 of the proof steps were carried out automatically. In further work, Mitsch et al.[119] verify a safety property that makes less conservative driving assumptions, allowing for imperfect sensors, and add liveness proofs to guarantee progress. They extend their approach to add runtime monitors that can be automatically synthesized from their hybrid models. They note that all models deviate from reality, so their runtime monitors complement the offline proofs by checking for mismatches between the verified model and the real world. Formal and nonformal models of the real world are prone to the problem of the reality gap, where models produced are never close enough to the real world to ensure successful transfer of their results[52, 204]. This is especially problematic when real-world interactions can impact safety. Bridging the gap between discrete models of behavior and the continuous nature of the real world in a way that allows strong verification is often intractable[34]. Moreover, in a multirobot setting one robot can be part of another's environment[86]. There is also a tradeoff between ensuring that the system is safe and ensuring that it is not so restrictive that it is unusable in practice [204].

A model that combines probabilistic and hybrid approaches is needed. This model should reconstruct the environment safely and it should update information deriving from the environment and capable of handling unexpected events. Localization and navigation are the key technologies of autonomous mobile service robots, and simultaneous localization and mapping (SLAM) is considered an essential basis for this. The main principle of SLAM is to detect the surrounding environment through sensors on the robot, and to construct the map of the environment while estimating the pose of the robot.

In the last decades, service robotics has reached a level of performance and credibility that makes it appealing to the general public. Some progress has been made in introducing autonomy, however only small individual actions have been demonstrated. Because of this, commercial robots still rely on hard automation, as industrial robots, or are teleoperated as surgical and underwater robots. In critical scenarios like surgical procedures, knowledge based approaches are the preferred way since they provide a clearer description of the workflow. For instance, in [59] an ontology-based framework for the automation of the peg&ring task has been proposed. The main drawback was the lack of real-time reconfiguration of the system. In fact, ontologies are much

more used in the field of situation understanding by humans [37]. A solution to the limitation of the ontologies can be found in non-monotonic programming, where the planning is carried out in a more flexible way, thus the knowledge can be updated in real-time from the sensing information. In [60] Answer Set Programming (ASP) has been used to define the reasoning module and has been successfully applied to an automated peg-and-ring task. The drawback of non-monotonic programming resides in the computational complexity required to solve a planning problem, which makes this approach often unsuitable for real-time applications.

1.2 Autonomy in surgery

Looking into surgical applications, all robotic platforms within an operating room primarily rely on surgeons to provide all guarantees through their experience and direct instrumental control via teleoperation. For instance, the most advanced robotic system available today in the operating room is the da Vinci® Surgical System, a remote teleoperation platform for minimally-invasive surgery (MIS) that does not present any automation capability and provides only video as feedback to the surgeon to ensure control stability under all circumstances

A significant part of current research in Robotic-assisted Minimally Invasive Surgery (R-MIS) is focussing on the development of autonomous systems for the execution of repetitive surgical steps, such as suturing, ablation and microscopic image scanning [40]. This would potentially help surgeons, who could focus on the more cognitive demanding parts of the procedure, leaving repetitive actions to the robot.

Every R-MIS system has to comply with tight requirements to be allowed within an operating room and has to provide safe interaction for the tools with both soft tissues and hard surfaces, such as needles, clips, and the tools themselves. Soft-tissue surgery in non-rigid anatomical environments should take into account hardly predictable scene changes, complicated tasks requiring collision-free motion planning and physical interaction with the environment

A proposal for the classification of autonomy grade in a surgical system [213], identifies five progressive levels:

- Level 0: no autonomy. The robot is fully teleoperated.
- Level 1: robot assistance. The robot provides support during teleoperation, such as virtual fixture or assisted guidance.
- Level 2: task autonomy. The robot can perform autonomously specific task initiated by the user, i.e. the user determines which task has to be performed and where.
- Level 3: conditional autonomy. The robot can generate autonomously different strategies to perform a task an the user decides which one should the robot apply.
- Level 4: high autonomy. The robot can take decision on the task to be performed in the surgery but under the supervision of the user.

Fig. 1.1: Different levels of autonomy as mapped to robotic surgery.

- Level 5: full autonomy. The robot can perform autonomously the entire surgery.

Within this scale (reported in Figure1.1), this thesis is located at a level 2: the system is bounded to operate reactively to the surgeon's actions and follow her/his lead during the operation while providing assistance to complete the tasks.

An important step in the development of cognitive surgical architectures is represented by the EU funded I-SUR project. It addressed the automation of needle insertion and suturing tasks[128] by means of a dual-arm robot with hybrid parallel/serial kinematics. The cognitive control architecture proposed by I-SUR[151] was able to operate in either teleoperated[48] or autonomous mode[152], guaranteeing a stable switch between the two and an adaptive interaction with the environment in both modes[49].

The introduction of autonomy requires systems with advanced capabilities in perception, reasoning and motion planning, together with specific methods to handle the interaction with the physical environment. In the surgical domain such tasks are very challenging due to the complexity of the anatomical environment, which is patient-specific and composed of soft tissues with dynamic behaviours.

Specifically, better medical imaging and vision techniques have significantly improved the performance of robotic surgical systems in a range of clinical scenarios, such as orthopaedics and neurosurgery [1]. Recently, several projects are actively working on introducing some level of autonomy in these robotic systems, as demonstrated by ARS (Autonomous Robotic Surgery) and SARAS (Smart Autonomous Robotic Assistant Surgeon) projects.

1.2.1 Autonomous Robotic Surgery (ARS)

The Autonomous Robotic Surgery (ARS) project aims at making the scientific advances that will enable the autonomous execution of complete procedures in uncertain and partially unknown environments.

The first scientific objective of the ARS project is to fully analyze and formally represent real surgical interventions with abstract models, integrating a priori knowledge from textbooks with the structures identified by big data analysis. This objective allows identifying not only the intervention details, but also the reasoning during surgery and action motivations, by comparing interventions done by different surgeons. Another key aspect of this objective is environment modeling, since data will support the identification of structures and properties of the anatomy and the creation of realistic phantoms. From this analysis the project develops the intervention specification to be verified during the demonstration phase.

The second scientific objective of the project isto develop methods to plan an intervention for a specific anatomy. Task planning overcomes the combinatorial explosion by instantiating the intervention model to the patient specific anatomy, thus limiting the number of possible choices. However, since not all steps can be planned in advance because of the changes occurring during the intervention, a key aspect of this objective is on-line and reactive planning.

The third scientific objective of the project is to develop methods for the real time control of the surgical instruments during the execution of the intervention. Hybrid controller are designed to account for the discrete evolution of the intervention and the continuous tool motions. Furthermore, since instruments must be localized with respect to the patient anatomy, an important aspect of this objective is the identification of the organ positions in the surgical area and their biomechanical properties.

The fourth and most ambitious scientific objective of the ARS project is the development of situation awareness and reasoning methods capable to handle a real surgical intervention. The objective is to reach autonomy at the level of sub-tasks of a surgical procedure, i.e. repetitive, yet tedious operations (e.g., dexterous manipulation of small objects in a constrained environment, as needle and wire for suturing). This will help reducing time of execution, hospital costs and fatigue of surgeons during the whole procedure, while further improving the recovery time for the patients. By using answer set programming (ASP), a logic programming paradigm, for task planning (i.e., coordination of elementary actions and motions). Logic programming allows to directly encode surgical task knowledge, representing plan reasoning methodology rather than a set of pre-defined plans. This solution introduces several key advantages, as reliable human-like interpretable plan generation, real-time monitoring of the environment and the workflow for ready adaptation and failure recovery.

The fifth and final objective of this project is to demonstrate the autonomous execution of a representative surgical intervention using the *daVinci Research Kit* (DVRK) setup and a patient specific physical phantom. This objective will first aim at improving the hardware setup and addressing the robot safety and security. Finally, the quality of the autonomous execution is

measured, by developing specific benchmarks. The integration of these five scientific objectives into a unified framework will permit the design, testing, and validation of autonomous surgical robots that is the core of the ARS project.

1.2.2 Smart Autonomous Robotic Assistant Surgeon (SARAS)

The goal of the project is to define the required technologies and to pursue the development of an effective robotic substitute to the assistant surgeon that currently works next to the patient within the operating room during R-MIS operations. All the instruments involved are general-purpose products for minimally invasive surgery, like scissors, graspers, clip appliers. However, to effectively validate the SARAS concept, the project focuses on radical prostatectomies, i.e. the resection of the whole prostate gland in male patients with prostate cancer while preserving urinary continence and erectile function, and partial or radical nefrectomies.

The project aims at developing three increasingly complex autonomous platforms to assemble a data-driven cognitive control architecture in which the surgeon and the robots operate seamlessly together. In the first, called Multirobots-Surgery platform, the main surgeon controls the da Vinci® tools from the console, whereas the assistant surgeon teleoperates standard laparoscopic tools mounted at the end effectors of the assistant robotic arms from a remote control station equipped with virtual reality and haptic devices. The assistant surgeon will perform the same actions as in standard robotic surgery, but this time by teleoperating the tools instead of moving them manually. The Multirobots-Surgery platform is an example of multi-master/multi-slave (MMMS) bilateral teleoperation system, where two users cooperate on a shared environment by means of a telerobotic setup. This setup already improves over standard robot-assisted radical prostatectomies as the assistant surgeon controls a sophisticated system that emulates standard laparoscopy tools and provides force feedbakc and virtual fixtures to the user. Moreover, the platform allows to acquire the relevant video and kinematic data from expert operators. In the second architecture, called the Solo-Surgery platform and the intended case study for the work of this thesis, the assistant surgeon will be replaced by the cognitive control architecture controlling the SARAS arms and adapting to the operator's actions to provide assistance. This platform is a very sophisticated example of a shared-control system: a surgeon operates remotely a pair of robotic laparoscopic tools (e.g. the da Vinci® Surgical Platform) and cooperates with the two novel SARAS autonomous robotic arms inside a shared environment to perform complex surgical procedures.

Finally, in the Laparo-2.0 platform, the only robot operating next to the patient is the SARAS assistant robot as the surgeon handles standard laparoscopy instruments instead of robotic tools. The use of a single robot in the operating room increases the challenges of controlling the collaborative robot as it introduces the requirement of visual tracking for all the instruments that, otherwise, can be achieved by exploiting the robots' kinematics.

1.3 Thesis contribution and outline

This thesis contributes to the ambitious goal of achieving surgical autonomy. Starting from the study of the sensors available in robotics, to the calibration of these with the robot and finally how to reconstruct an unknown environment. This Section has specified the evolution of autonomy in robotics, what are the major difficulties, some proposed solutions, and projects I contributed to during my doctorate.

This thesis focuses on three aspects that are deemed fundamental for an autonomous surgical robot:

- sensors available in robotics, how they integrate into robotic surgery, their use and the development of a new endoscope prototype;
- hand-eye calibration between the sensor and the robot itself, developing a new approach with the aim of improving the accuracy of interaction with the anatomical environment;
- 3D reconstruction of the anatomical environment, developing a new *simultaneous localization and mapping* (SLAM) algorithm.

The first contribution is the analysis of sensors that can be used in robotic surgery. Their advantages and disadvantages motivate the need for new technology to improve the accuracy of the reconstruction of the environment. The proposed solution is a new type of endoscope, which combine a standard stereo camera with a *Time of Flight*(ToF) sensor, in order to obtain an accurate 3D depth map at a greater distance than standard endoscopes.

The second contribution is the development of a calibration procedure which is adaptable to all types of sensors and allows better accuracy in the workspace than state-of-the-art procedures. The experiments for validating the first two contributions of this thesis are performed on a benchmark training task for surgeons and to the development of new algorithms to recognize structures that are present in anatomical environments.

One challenge of autonomous robotic surgery is the unpredictability of the anatomical environment and its behavior intra-operatively, depending on the specific patient. Hence, the third contribution brings together the others works and aims to develop a new SLAM algorithm that adapts to the anatomical environment. Starting with the registration of the pre-operative images to the 3D reconstruction, then to distinguish dynamic features from static ones to adapt the algorithm to a dynamic environment.

The experimental setup provided by the ARS and SARAS projects allowed the entire system to be tested. However, the next chapters will clarify that the generality and the applicability of the proposed methodologies in this thesis are preserved.

The remainder of the thesis is organized as follows. First, Chapter 2 describes all the sensors we used and the endoscope we developed. This chapter is useful to clarify the challenges of each type of sensor, their advantages and disadvantages and and the need to build new technologies to increase accuracy in the medical field. In Chapter 3 a new hand-eye calibration procedure is presented, while chapter 4 describes our benchmark for validating calibration

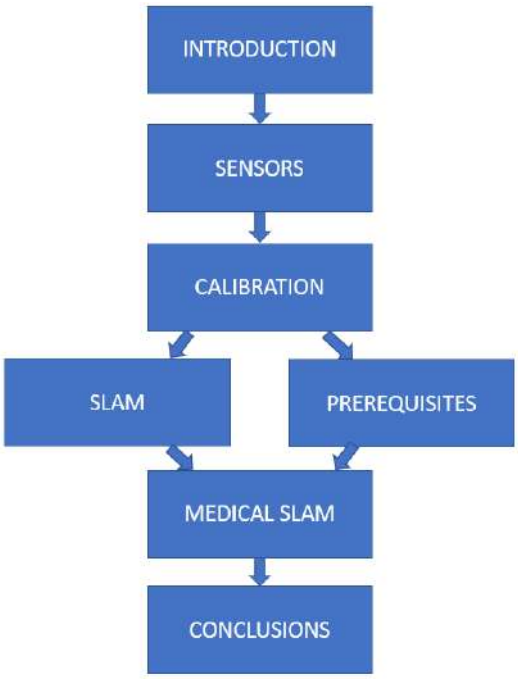


Fig. 1.2: Organigram of the Chapters.

and algorithms useful for the development of autonomy in surgery. Chapter 5 describes the methods to reconstruct an unknown environment, reviewing the state of the art of SLAM algorithms. Chapter 6 and 7 describe what modifications were necessary to use the SLAM in an anatomical environment. Finally, Chapter 8 summarizes the results of this thesis and proposes future extensions. Figure 1.2 presents a reading organigram of the Chapters.

2

Sensors

This chapter describes all the sensors available in a robotic system. explaining first of all how the cameras work and then the differences between the various sensors. Highlighting what are the strengths and weaknesses of each one up to the new endoscope model that we developed for Robotic-assisted Minimally Invasive Surgery (R-MIS).

2.1 Image acquisition

An imaging apparatus works by collecting the light reflected from the objects in the scene and creating a two-dimensional image. If we want to use the image to obtain information on the scene, we must study the nature of this process. The most common types of cameras are shown in 2.1

(a) Monocular (b) Stereo (c) RGB-D

Fig. 2.1: Sensors available.

The simplest geometric model of image formation is the pinhole camera, represented in figure 2.2. This is the same principle as the Renaissance dark-room. Let M be a point of the scene, with coordinates (X, Y, Z) and let M' be his projection on the image plane, with coordinates (X', Y', Z'). If f is the distance of the projection center from the image plane (focal distance), then from the similarity of the triangles we obtain:

$$\frac{-X'}{f} = \frac{X}{Z} and \frac{-Y'}{f} = \frac{Y}{Z} \tag{2.1}$$

then,

Table 2.1: A comparison of the sensors

Sensor	Advantages	Drawbacks
Monocular	Smallest Lowest power consumption Cheapest Minimal calibration	Scale is unobservable Scale drift 3D only from multi-view No mapping under pure rotations
Stereo	3D from one stereo frame	More processing per frame Extrinsic calibration
RGB-D	Directly provide dense depth map Dense maps 3D metric system	Active sensor Only indoors Comples calibration Power consumption
IMU	Inter-frame motion estimation Pitch and roll are observable	Varying sensor biases Gravity must be compensated Observability issues Visual-inertial calibration Synchronization

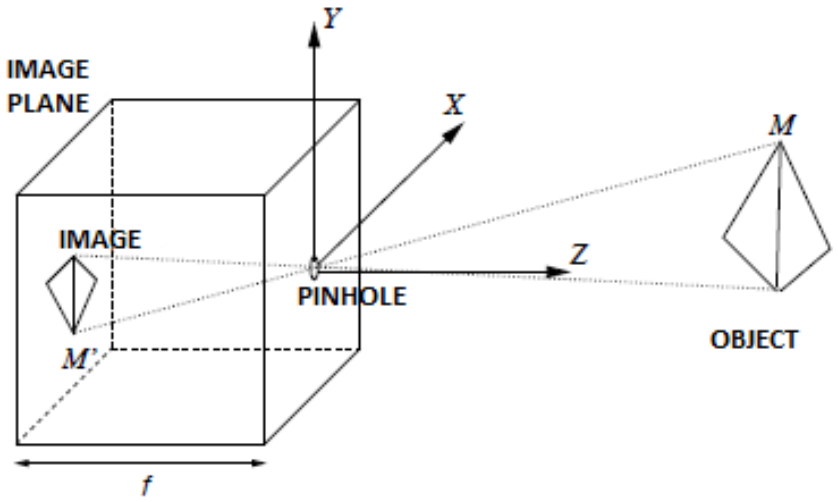


Fig. 2.2: Formation of the image in the pinhole camera.

$$X' = \frac{-fX}{Z} \qquad Y' = \frac{-fY}{Z} \qquad Z' = -f \tag{2.2}$$

Note that the image is inverted with respect to the scene, both right-left than above-below, as indicated by the minus sign. These equations define the process of image formation that takes the name of perspective projection. We can model the projection perspective by placing the image plane in front of the projection center, thus eliminating the negative sign. The division by Z is responsible for the foreshortening effect, so the size of the image of an object varies according to its distance from the observer.

2.1.1 Lens

Vertebrate eyes, cameras and video cameras use lenses. A lens is able to collect more light. The downside is that not the whole scene can be in focus at the same time. The approximation we make for the system's perspective acquisition,

which in general is very complex, being constituted from more lenses, is that of the *thin lens*. Thin lenses have the following properties:

- the rays parallel to the optical axis incident on the lens are refracted in order to pass through a point of the optical axis called focus F.
- the rays passing through the center C of the lens are unaffected.

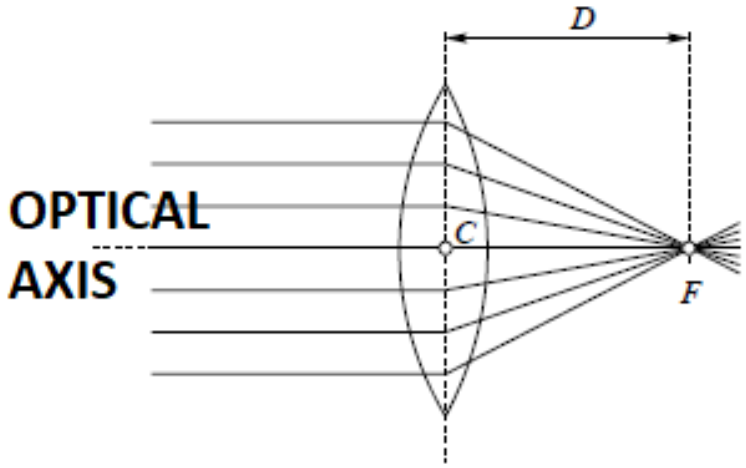


Fig. 2.3: Thin lens

The focus distance F from the center of the lens C is called distance focal length D (figure 2.3). It depends on the radii of curvature of the two surfaces of the lens and the refractive index of the constituent material. Given a point of the scene M it is possible to graphically construct the image M' (or conjugate point) using two particular rays which start from M: the beam parallel to the optical axis, which after refraction passes through F and the radius that passes unaltered through C (figure 2.4).

Thanks to this construction and the similarity of the triangles, we obtain the conjugate point formula (or thin lens equation):

$$\frac{1}{Z} + \frac{1}{Z'} = \frac{1}{D} \tag{2.3}$$

The image of a point of the scene distant Z from the lens is produced (in focus) at a distance from the lens Z', which depends on the depth to Z of the point and the focal distance D of the lens. To focus an objects at different distances, the eye lenses change focal length deforming as the camera lenses move along Z. Basically, the image of point M, when it is in focus, is formed on the image plane in the same point foreseen by the pinhole model with the hole coinciding with the center of the lens C, in fact the ray passing through C is common to the two geometric constructions. The other light rays that leave M and are collected by the lens serve to increase the light reaching M'. If 2.3 is not verified you get a blurry image of the point, that is, a circle that takes the name of a circle of confusion. Plan image is covered by photosensitive elements which

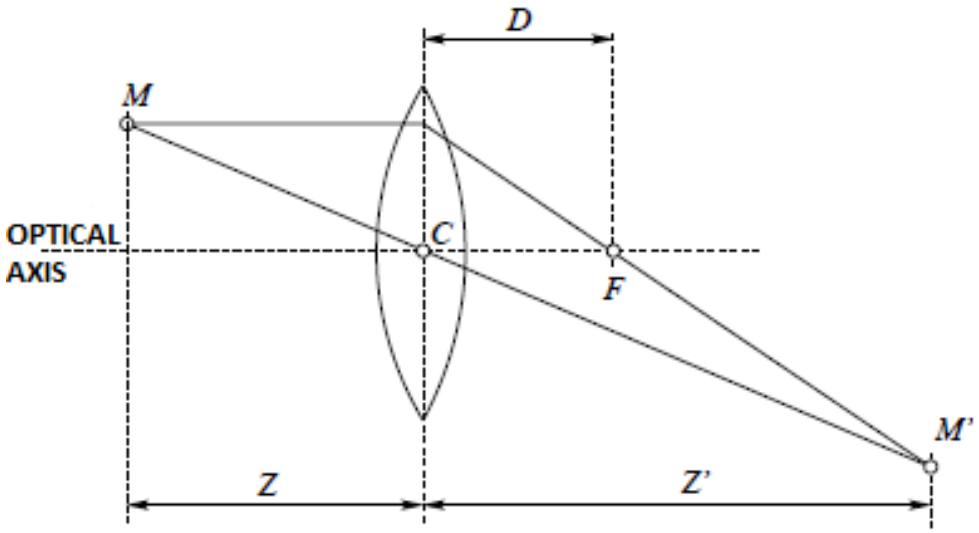


Fig. 2.4: Construction of the image of a point.

have one dimension small but finished. Until the circle of confusion exceeds the dimensions of the photosensitive element the image is in focus. So, there is a depth range for which the points are in focus. This interval is called *depth of field*. The depth field is inversely proportional to the diameter of the lens. Indeed the pinhole camera has an infinite depth of field. The light which is collected, instead, is directly proportional to the diameter of the lens.

2.2 Monocular camera

Monocular cameras are the most common, we use them every day with our smartphone and there are many types and resolutions.These cameras are mainly composed of the image sensor and lens. We assume that the camera can be accurately modeled as a pinhole camera, once lens distortion has been removed, so that a $3D$ point $X_c \in \mathbb{R}^3$ in the camera coordinate reference system C is projected into $2D$ pixel coordinates x with the projection function $\pi_m : \mathbb{R}^3 \to \mathbb{R}^2$:

$$x = \pi_m(X_c) = \begin{bmatrix} f_x \frac{X}{Z} + c_x \\ f_y \frac{Y}{Z} + c_y \end{bmatrix} \tag{2.4}$$

$$X_c = [X, Y, Z]^T, \qquad x = [u, v]^T$$

where f_x and f_y are the horizontal and vertical focal lenghts, and c_x and c_y the horizontal and vertical coordinates of the principal point. These are intrinsic calibration parameters that can be computed from several images of a known calibration pattern. The camera coordinate system C has its origin at the optical center and follows the standard directives for its transform: the Z axis is looking forward, the X axis points to the right and the Y axis points downwards. The projection function π assumes no distortion introduced by

the lens. Well known software and libraries like Matlab or OpenCV include toolboxes for camera calibration, including distortion.

2.3 Stereo camera

Stereo cameras are composed of two rigidly connected cameras. Ideally both cameras are hardware synchronized so that image capturing is triggered at the same time. Depth can be estimated from just one stereo frame through the stereopsis, which is the process that allows to obtain information on the three-dimensional structure from a pair of images, coming from two cameras that frame a scene from different locations. The distance between both cameras, known as the baseline b, along with focal length and image resolution will determine the depth range at which depth estimation is accurate. We can identify two subproblems: computation of correspondences and triangulation. The first consists in the coupling between points in the two images that they are projection of the same point of the scene. These points are called *conjugate points*. Notice the couplings between the points of the two images and note the position reciprocal of the cameras and the intrinsic parameters of the sensor is It is possible to reconstruct the position in the scene of the points that are projected on the two images. This triangulation process requires the calibration of the stereo system, or the calculation of the parameters intrinsic and the reciprocal position (extrinsic parameters) of the cameras. In order to facilitate stereo matching, images are typically rectified, removing distortion and rotating them so that the epipolar lines are horizontal, i.e. the correspondence of a pixel in the left image lies on the same row in the right image. The projection function for a rectified stereo camera $\pi_s : \mathbb{R}^3 \to \mathbb{R}^3$:

$$x = \pi_s(X_c) = \begin{bmatrix} f_x \frac{X}{Z} + c_x \\ f_y \frac{Y}{Z} + c_y \\ f_x \frac{X-b}{Z} + c_x \end{bmatrix} \tag{2.5}$$

$$X_c = [X, Y, Z]^T, \qquad x = [u_L, v_L, u_R]^T$$

where (u_L, v_L) are the coordinates in the left image and u_R is the horizontal coordinate in the right image. The vertical coordinates in both images are the same and we assume that both cameras have the same intrinsic parameters after rectification.

2.4 RGB-D Camera

RGB-D cameras are the combination of a monocular RGB camera and a depth sensor, based on structured light or time of flight. The measured depth can be registered into a depth map with $1:1$ pixel correspondences to the RGB image, after the extrinsic calibration between the camera and the depth sensor The main advantage of this camera is that for every pixel in the image we

know its depth value without needing to perform a stereo matching as in the case of the stereo camera. However, their use is restricted to indoors and the depth range is limited.

2.5 Inertial Measurement Unit (IMU)

Inertial Measurement Units (IMU) are composed of a gyroscope that measures the angular velocity, and an accelerometer that measures the linear acceleration of the sensor. IMU provides information of self-motion while the vision camera observes the scene, it can be used to estimate motion between camera frames or to estimate the metric scale and the gravity. The IMU, whose reference we denote with B, measures the acceleration a_B and angular velocity ω_B of the sensor at regular intervals Δt. The measurements are affected by sensor noise and by slowly varyng biases b_a of the accelerometer and b_g of the gyroscope. Moreover the accelerometer is also subject to gravity g_W which must be subtracted to compute the motion. The discrete evolution in the world reference W of the IMU orientation $R_{WB} \in SO(3)$, position ${}_W p_B$ and velocity ${}_W v_B$, ca be computed as follows:

$$R_{WB}^{k+1} = R_{WB}^{k} Exp((\omega_B^k - b_g^k)\Delta t) \tag{2.6}$$

$$\begin{aligned} {}_W v_B^{k+1} &= {}_W\, v_B^k + g_W \Delta t + R_{WB}^k (a_B^k - b_a^k)\Delta t \\ {}_W p_B^{k+1} &= {}_W\, p_B^k + {}_W\, v_B^k \Delta t + \frac{1}{2} g_W \Delta t^2 + \frac{1}{2} R_{WB}^k (a_B^k - b_a^k)\Delta t^2 \end{aligned}$$

where Exp is the exponential map for $3D$ rotation group $SO(3)$. The calibration between the IMU sensor and the vision is essential in order to synchronize with the same clock and without drift. Also the extrinsic calibration is needed to obtain the transformation $T_{CB} = [R_{CB}|_C p_B]$ between the reference of the camera and the IMU sensor.

2.6 Endoscope

An endoscope is a long, thin, flexible or rigid tube that has a light and camera at one end. Images of the inside of a body are shown on a television screen. Such a device could provide minimally invasive access to sections which are accessible only through invasive methods. Conventional endoscopes are highly flexibile devices for minimally invasive inspection in interior lumens and cavities, such as the stomach, colon, urinary tract, respiratory tract, etc. These medical imaging devices are available in many sizes to suit different purposes. Different types of endoscope cameras are used for different areas of the body. Upper endoscopes, bronchoscopes, colonoscopes, sigmoidoscopes, and many other endoscopes are named for the areas where they are used. Endoscope cameras are beneficial pieces of medical equipment. They are versatile enough to be used as both diagnostic and treatment instruments. These cameras provide visual imaging as a means of replacing unnecessary or exploratory surgery.

With the availability of several options, doctors can choose the best endoscope for each area of the body. The introduction of the stereo endoscope has played an important role in video $3D$ laparoscopy system, that provides stereo view to the surgeon. The simple 3D endoscope consists of two complementary metal-oxide-semiconductor (CMOS) camera modules that provide real-time stereo view to the surgeon via the stereo viewer. This helps the Robotic Minimally Invasive Surgery (R-MIS), which is nowadays a standard for many surgical procedures, where it demonstrated to provide benefits over traditional laparoscopic or open surgery, namely patient's safety, minimized collateral surgical trauma and quicker recovery. Currently, all surgical robotic systems on the market are teleoperated by a main surgeon from a remote console [30, 141, 213], while an assistant surgeon directly operates next to the patient with laparoscopic tools. This setup helps the main surgeon in smoothing his/her movements and increase their precision through motion scaling, while also reducing both the required physical and cognitive effort to perform the operation. On the other hand, having no direct contact with the patient visual perception of the anatomical environment becomes a crucial point in such systems.

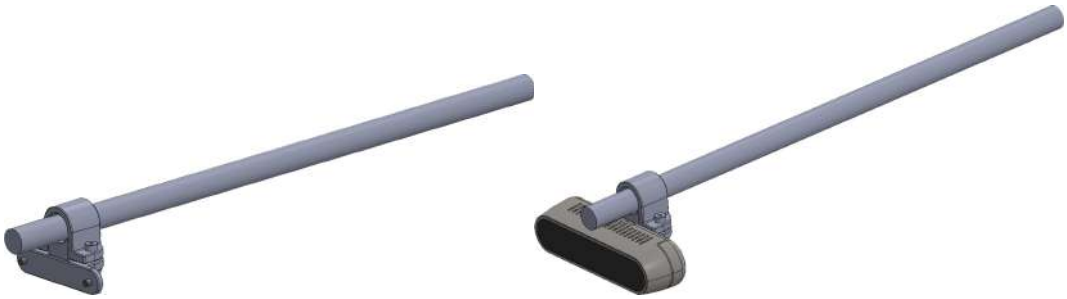
The next step forward in R-MIS is expected to come with an autonomous robot that can either perform some routine operations by its own or assist the main surgeon during an intervention [31, 50, 129, 137]. The introduction of autonomy degrees requires systems with advanced capabilities in perception as well as in reasoning, motion planning and physical interaction.

As of today, all the systems on the market use endoscopes equipped with a RGB stereo pair that is able to provide a nice visual representation of the environment by directly streaming the two channels on a 3D visualization device. This is a robust solution for teleoperated systems, with extremely limited delay (no processing of the video streaming is required) and high visual quality. Nevertheless, building a metric 3D reconstruction of the anatomical structures with such an instrument is very critical for two main reasons: first, disparity estimation in anatomical structures is very challenging due to the poor textures and the highly deformable nature of the bodies; second, the error in depth estimation is inversely proportional to the baseline, which must be a small value for mechanical constraints (i.e. limited by the diameter of the endoscope itself). Moreover, the depth estimation error grows quadratically with the distance from the sensor, which translates in a highly variable accuracy between the near field and the far field views [57]. Therefore, for autonomous surgical robots the 3D reconstruction provided by traditional stereoscopic endoscopes is not satisfactory, thus exploring new technologies is deemed of primary importance.

To overcome the limitations of standard endoscopes we used two different approaches. First approach was to integrate a RealSense sensor to the endoscope to provide the 3D reconstruction of the environment.

Therefore, we developed a stop-gap solution to account to this issue: we designed an adapter, shown in Figure 2.6, that attaches rigidly to the da Vinci® endoscope an "Intel RealSense®" camera. Its technical specifications are reported in Table 3.1.

Fig. 2.5: The first prototype of the depth stereo endoscope



(a) (b)

Fig. 2.6: The adapter to attach the RealSense to the da Vinci® endoscope.

The positioning of this camera via the adapter, although clearly not viable to be used internally in its current form, allows the hand-eye calibration procedure to complete successfully, thus it permits the alignment of the pre-operative and the reconstructed 3D pointclouds. This adapter and camera configuration allows to maintain separate the point-of-view of the point cloud reconstruction from the stereoscopic image pair directed to the da Vinci® console. This fact reduces the overall error since keeping the 3D camera further away from the target reduced the overall noise and improves distance measurement. The addition of the separate camera does not hamper the scene seen by the surgeon nor prevent the calibration procedure of the endoscope arm.

This temporary solution was adopted to test the calibration and reconstruction algorithms while we were waiting for the manufacturing of the new endoscope.

	Camera specifications
Resolution	1280×720
Field of view (FOV)	$91° \times 65° \times 100°$
Frame rate	90 fps
Baseline	50 mm
Z-accuracy	$\leq 2\%$ of the working distance

Table 2.2: RealSense d435 specifications

The second approach was to develop the *SARAScope*, a new model of endoscope that combines a Time-of-Flight (ToF) sensor with a RGB stereo pair towards a high resolution accurate depth estimation combined with color information. This multimodal acquisition allows us to build an accurate model of the anatomical structures needed to drive the autonomous robots, while also streaming high resolution images to the main surgeon visualization device for teleoperation.

In the following sections we provide a brief excursus on the evolution of 3D endoscopes for MIS and the shortcomings that lead to the novel design. Then, we present the details of the endoscope design through the mechanical, optical, and control software specifications and the calibration procedure that allows to combine the depth and the color information into a high-precision 3D reconstructed image. Finally, we discussed the experimental to draw some conclusions about the potential of the prototype endoscope.

2.7 Endoscopy Evolution and Applications

The continuous development in endoscopy carries profound implications to the betterment of care: more advanced endoscopes directly translate into better vision for the surgeons and thus to improved patient care and reduced costs [190]. In less than 20 years we moved from endoscopes based on optical fibers to convey light to external analogic cameras to the most advanced market-available endoscopes today, that are *chip-on-tip* digital stereoscopic endoscopes which provide exceptional image quality in a very compact packaging. However, this improvement brings little to no benefit to autonomous R-MIS applications, as the reduction in endoscope's size is actually detrimental to stereoscopic 3D reconstruction.

Many solutions have been devised to overcome this issue, most of them trying to eliminate the dependency to the stereoscopic camera altogether. Mahmoud *et al.* [106] applies the ORB-SLAM algorithm [125] to generate a 3D reconstruction from a high resolution monocular camera stream. The primary issue of this technology arises in the restriction imposed to the camera movement by the remote center of motion (RCM) requirement due to the patient's safety: this kinematic constraint prevents pure translation in the camera's point-of-view that induce additional errors in an algorithm designed for mobile robotic platforms. The few solutions that overcome this constraint propose mechanical [89] or soft-robotics [35] flexible endoscope tips that intend to

improve exploratory capabilities for the surgeon; unfortunately, these devices also introduce position uncertainty in their kinematic chains.

Additional solutions to the small baseline issue in stereoscopic endoscopes have been investigated by researchers: these include the use of trinocular vision [24], controlled aberration [212], and monocular shading [23]. These are all passive technologies that operate at a software level, i.e. image processing, to improve depth estimation but experimental results showed only a marginal improvement over traditional triangulation techniques.

Within active technologies, the most studied approach in endoscopic imaging is based on structured light [55, 94, 110, 169]. It consists in the projection of a pattern by infrared laser light at low intensity that is captured by a specialized camera to obtain a 3D map from the deformation of the observed pattern. Unfortunately, these deformations could interfere with reflections over the illuminated surfaces which greatly reduce reliability of depth estimation. At consumer-level products, the comparison between Structured Light and Time-of-Flight technologies presented in [15] for the Kinect® camera demonstrated how the latter produces more precise and stable results.

The integration of visible and depth images has been explored in [84]. In this paper the authors present a hybrid 3D endoscope that exploits the high-resolution of a single RGB camera to estimate subpixel motion used as a cue for super-resolution imaging. The proposed prototype is compact, but it does not integrate stereoscopic vision and does not provide in output a 3D pointcloud. Moreover, as acknowledged by the authors, the high computational cost for the super-resolution makes it unfeasible in real time applications. In this work, we exploit the depth channel as a prior and a soft-constraint for the estimation of a 3D pointcloud directly from the stereo camera. As such, we aim at providing the both the surgeon and autonomous reasoning algorithms with a real-time 3D reconstruction of the anatomical environment.

2.8 Time-of-Flight 3D Stereo Endoscope

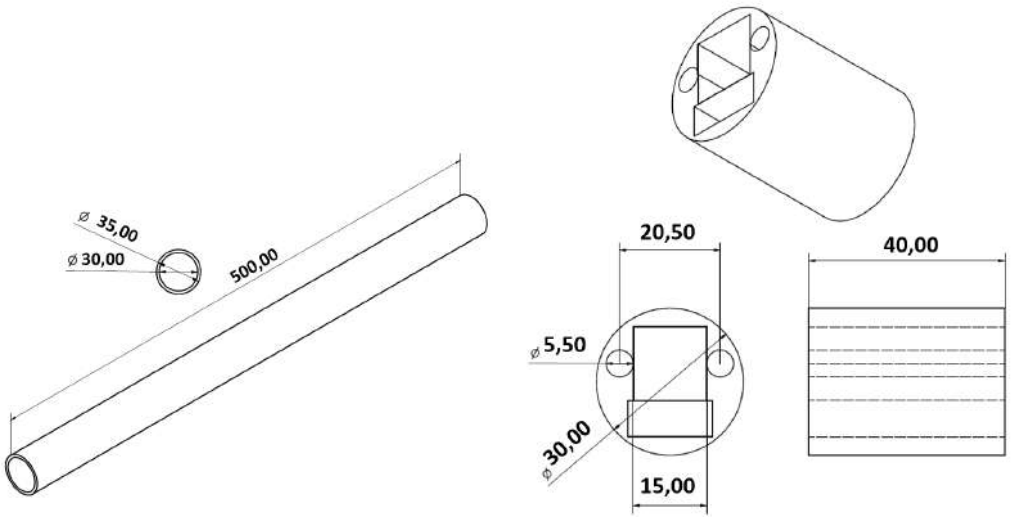
In this section we present in details our novel 3D endoscope designed to improve the view over the operating field by merging a stereoscopic camera with a Time-of-Flight (ToF) sensor. Both these technologies have been available on the consumer market for a long period of time, but to the best of our knowledge this is the first time they are presented in a single package operating as a unit in endoscopy.

We named the prototype *SARAScope*, and we will refer to it as such throughout the thesis, as it has been designed within the EU-funded SARAS project[1]. The *SARAScope* is composed of two RGB cameras rigidly coupled with a short range ToF depth sensor. The purpose of this endoscope is to provide a dense 3D point-cloud reconstruction of the anatomical environment while mounted on a robotic platform that provides accurate positioning, as well as to present an accurate 3D representation to the main surgeon operating at the da Vinci® console or wearing a virtual reality device. Standard

[1] `https://saras-project.eu/`

ToF cameras on the market operate within a range starting from 15 cm, which make them not directly suited for this application. Indeed, the endoscope is usually placed by the surgeon at a working distance of about 5 to 10 cm from the main anatomical structures of interest. Fulfilling this requirement has only recently become possible as more compact ToF cameras started appearing on the market with increased close-range capabilities compatible with the required workspace constraints. In the *SARAScope* the stereo cameras provides the 3D vision to the main surgeon, with the color information provided by the same being reprojected back on the 3D reconstructed environment to achieve a setup similar to larger, high-end ToF cameras such as the Intel® RealSense™. A first prototype of the *SARAScope* is shown in Figure 2.5.

The endoscope prototype is made of an aluminium cylinder with a length of 500 mm and outer diameter of 35 mm. The inner size of the endoscope is 30 mm. The cylinder is waterproof and it contains the two RGB cameras, the depth sensor and a high brightness white led. Figure 2.7 shows the CAD model of the endoscope.



(a) (b)

Fig. 2.7: The CAD model of outer part of the endoscope and the stereo vision system.

The whole vision system is attached to the one of the end of the body, the RGB cameras are placed co-planar with a baseline of 21.5 mm as shown in Figure 2.8. The technical specifications of the *SARAScope* are reported in Table 2.3, while the technical specifications of the vision systems forming the endoscope are in Table 2.4.

We are aware that the outer diameter (35 mm) is too large to be used in R-MIS (even though single port surgery relies on trocar close to such diameter). However, this is only a first prototype and we argue that it is valuable as a proof of concept. As a future work we plan to improve the optomechanical

design, by miniaturizing the optical components, to further reduce the outer diameter of the endoscope to 20 mm, thus making it suitable for real surgeries.

Characteristic	**Value**
Operating temperature (°C)	0 to 45
Nominal depth resolution (mm)	≤ 1 % (500-4000), ≤ 2 % (50-1000)
Maximum diameter (mm)	35
Length (mm)	500
Weight (g)	≤ 450
Visible bandwidth light source	High brightness white LED
Depth sensor light source	VCSEL Classe 1 Laser (850 nm)

Table 2.3: Technical specifications of the *SARAScope*

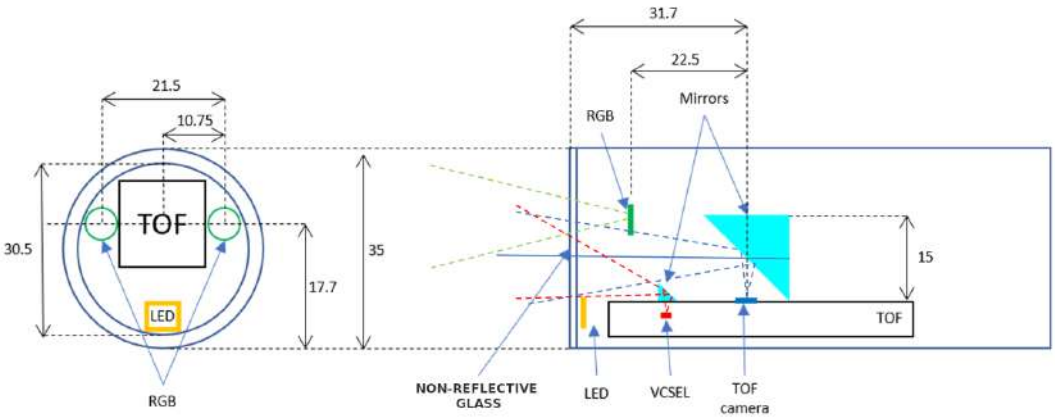


Fig. 2.8: A detailed view of the relative position of the RGB cameras and the depth sensor.

2.8.1 Endoscope characterization

The endoscope has been characterized using a specific target, shown in Figure 2.9, composed of a set of concentric circles with radius 1.25 cm, 2.5 cm, 5 cm, 8 cm and 11 cm. These radius refers to a FoV of 28 deg^2

at the distances of 5 cm, 10 cm, 20 cm, 30 cm and 40 cm between the endoscope and the target. Fixing the endoscope over a rail we verified the expected requirements (FoV and working distance) at each of the fixed distances previously defined. This analysis permits the identification of issues in the optical design of all cameras whenever the resulting view presents distortions of the target.

These experiments, a visible representation of which is shown in Figure 2.10, verify that the optical properties of each single camera unit are generally main-

Characteristic	**RGB cameras**	**Depth sensor**	**Endoscope**
Horizontal FoV	55	62×45	> 30, > 40
Working distance (mm)	70-Inf	50-4000	70-Inf
Optimal working distance (mm)	70-400	100-4000	70-400
Resolution (pixel)	1280×720	224×171	853×640,145×108
Communication interface	USB 2.0	USB 3.0	-
Frame rate (full resolution)	15 FPS	45 FPS	15 FPS
Driver	UVC	Royale	-

Table 2.4: Technical specifications of the vision systems used in the *SARAS-cope*.

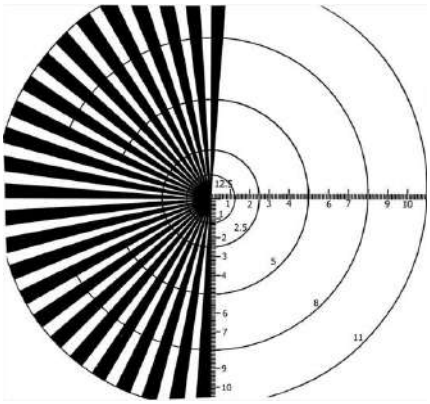


Fig. 2.9: Example of target used during the experimental characterization of the endoscope.

tained with little to no observable distortion of the characterization target through either the RGB or the ToF sensors. Nevertheless, the optical design does present minimal issues in the distance characterization that require some adjusting during the calibration phase (which is presented in Section 2.9). Specifically, both the distances and the relative orientation of the acquired target plane present measurements errors that can be attributed to the sensitivity of the infrared sensors to the mounted lenses and mirrors that amplify the noise in the infrared spectrum.

2.8.2 Software design

The endoscope is integrated and controlled in the SARAS platform using the Robot Operating System[2] (ROS) for which we developed a specific ROS package that provides a driver for all camera units within the `sarascope` that is in charge of performing the RGB-depth alignment and publishing the coloured point cloud. Moreover, it captures the images from the RGB cameras and makes them ready to be provided to the main surgeon at the dVRK console [3].

[2] `https://www.ros.org`
[3] `https://github.com/jhu-dvrk/sawIntuitiveResearchKit`

The software has been developed as two nodes that act as a foundation for the communication with the sensors, and a node built on top the previous two that deals with merging the information coming from the lower driver level.

The RGB driver has the task of instantiating a stream of data coming from the camera to the PC. The stream takes place via USB serial communication and contains the frames captured by the camera sensor. These frames are, then, made available for reading via the standard ROS image transport interface

The node needs the following inputs:

- **left id** is the identification number of the serial port for the left camera;
- **right id** is the identification number of the serial port for the right camera;
- **cam res** is a string of character which indicates if we want the RGB cameras at high or standard definition. Its value can be "HD" or "SD";
- **calib** is a boolean value which indicates if the cameras are calibrated or not;
- **left config file** when the calib value is true, this file contains the path to the configuration file for the left camera. The configuration file is formatted as yaml and collects the calibration parameters;
- **right config file** same as left config file, but for the right camera.

The chip-on-tip cameras embedded in the SARAS endoscope provide an image quality comparable with the current da Vinci® albeit with a reduced field-of-view (they operate with a pixel resolution of (1280×1024)). The Depth Driver is the node in charge of managing the data stream from the depth sensor to the PC. The depth sensor is also connected via standard USB.

The node takes as input the following arguments:

- **exposure time** which sets the depth sensor exposure time;
- **exposure mode** which sets the depth sensor exposure mode;
- **min filter** which sets the minimum distance where to filter out the points;
- **max filter** which sets the maximum distance where to filter out the points;

The depth driver is open-source software developed by Tom Panzarella, written in the C++ language and can be downloaded at the prodived link [4].

2.8.3 SARAScope 3D

The SARAScope 3D node takes as inputs the frames of the RGB cameras from the RGB Driver and the depth image from the Depth driver, and fuses those information to create a colored point cloud.

The node uses the following parameters estimated using the calibration procedure described in the Section 2.9:

- **left displacement** is the displacement between the left camera and the depth sensor.
- **left Euler** is the orientation expressed as Euler angles between the left camera and the depth sensor.

[4] `https://github.com/ifm/royale-ros`

- **right displacement** is the displacement between the right camera and the depth sensor.
- **right Euler** is the orientation expressed as Euler angles between the right camera and the depth sensor.
- **cam info subscribers** contains the path to the files with the extrinsic parameters of each camera.

By using the camera matrix and the distortion model, the node projects the pixels of the camera on the corresponding points of the point cloud. In order to achieve good results, an accurate and robust calibration is needed. The projection could be performed by using only one between the left or the right camera, but, in this case, we could encounter occlusion problems. Thus, when an object is occluded in the chosen camera, it could still be visible by the depth camera, resulting in coloration inaccuracies. To address this issue, we implemented a solution that merges the results from the left and the right cameras to obtain the correct pixel coloring for the point cloud also in the event of occlusions of a single perspective. Clearly, the problem could still occur if the object were to occlude both the left and right camera perspectives, but this eventuality is deemed unlikely since it implies that also the surgeon's view of the scene is occluded.

2.9 Calibration For Surgical Robots

The application to 3D reconstruction of the novel endoscope to surgical robotics use cases requires to calibrate the pose of the endoscope and its reconstructed image in a common reference frame. This calibration process is performed in three steps: the first step involves finding the reconstructed image location with respect to a known position of the endoscope (in our case, this position is the tooltip of the endoscope); the second step operates the extrinsics calibration between the RGB cameras and the ToF relative to the tooltip, which allows to match the visible and infrared images; finally, the last step is to perform a *hand-eye* calibration to compute the position of all robots in the scene relative to the endoscope. The hand-eye calibration also allows to correctly position every 3D-reconstructed anatomy within a common Cartesian frame [161, 196].

2.9.1 Tooltip Estimation

As the current endoscope represents a proof-of-concept for future developments, it requires particular care for every measurement. The ToF camera location estimate relative to the tooltip has been performed by placing a flat surface perpendicular to the main axis of the endoscope (in a similar manner to the characterization in Section 2.8) at various known distances within the estimated workspace range (i.e. between 40 and 120 mm). We evaluated both the distance and the orientation, the latter being the angles (in rad) of rotation along the longitudinal and latitudinal axes, perpendicular to the main axis. We process the data acquired by the ToF camera through RANSAC [51] to find

the best fitting plane given the noisy raw measurements. Table 2.5 reports the estimates of such fit with estimates on the mean and standard deviation from the ground truth error. We, then, evaluate the distance between the ToF camera to the tooltip by calculating the average of the difference of each of the known distances to the three reference measurements. Primarily due to construction inaccuracies of the prototype, the pose identification required minor adjustments to both the estimated distance plane orientation. Therefore, we applied a simple first-order polynomial fit to the position to minimize the error, which produced the coefficients $0.9587D + 3.3$, and we used the average orientation estimation on the latitude $\bar{\rho} = -0.15$ rad and the longitude $\bar{\theta} = -0.016$ rad as the relative orientation of the endoscope to the view.

Table 2.5: Overall system accuracy evaluation (in mm and rad)

Measures @	40 mm	80 mm	120 mm
$\bar{D}$	64.99	105.78	141.69
$\max \epsilon$	2.57	3.83	3.42
Σ	0.59	0.89	1.04
σ^2	0.46	0.65	0.76
ρ (rad)	-0.17	-0.15	-0.13
θ (rad)	0.01	-0.04	-0.02

To apply the colour information to the 3D reconstruction we require the extrinsics calibration of the endoscope, thus we adopted the procedure presented by Zhang [220] We acquired from the depth sensor and from the RGB cameras a set of images of a checkerboard with a square size of 2 mm. Then we perform the stereo calibration twice: the first time to find the transformation between the ToF reference frame and the left camera reference frame, the second time to find the transformation between the ToF and the right camera reference frame. Figure 2.11 shows an example of the images used during the estimation of the intrinsic and extrinsic camera parameters.

2.9.2 Hand-eye calibration

To find the transformation between the 3D endoscope and the common reference frame for all the robots involved in the setup, we slightly modify the procedure described in [161]. We use the point-cloud generated by the 3D endoscope, which is a dense point-cloud but limited by the size of the cropped depth image. The first step of the calibration procedure consists, as before, in the computation of the rigid transformations $T_{\star}^{w}$ between the common reference frame (*world*) and the base frame of the arms. Instead the estimation of the transformation $T_{endoscope}^{arm}$ between the camera reference frame and the robot's arm reference frame is based on the points recognition on a new calibration board.

The recognition algorithm is based on the *SimpleBlobDetector* class of the OpenCV library. This class implements a simple four-step algorithm to extract blobs from an image:

1. Convert the source image to binary images by applying thresholding with several thresholds, from a minimum (inclusive) to a maximum (exclusive), with distance between neighboring thresholds.
2. Extract the connected components from every binary image by finding contours and calculate their centers.
3. Group up the centers extracted from multiple binary images by their coordinates. For each image, the centers that are close enough per group are fused into a single blob as controlled by the *minimum distance between blob* parameter.
4. For each resulting blob groups, estimate the final center points of the blobs and their radii to return them as locations and sizes of the desired keypoints.

The *SimpleBlobDetector* also implements several filtering techniques over the recognized blobs:

- **Color**: it compares the intensity of a binary image at the center of a blob to *blobColor*.
- **Area**: the extracted blobs have an area between *minArea* (inclusive) and *maxArea* (exclusive).
- **Circularity**: the extracted blobs have circularity between *minCircularity* (inclusive) and *maxCircularity* (exclusive).
- **Ratio of the minimum inertia to maximum inertia**: the extracted blobs have this ratio between *minInertiaRatio* (inclusive) and *maxInertiaRatio* (exclusive).
- **Convexity**: the extracted blobs have convexity (area / area of blob convex hull) between *minConvexity* (inclusive) and *maxConvexity* (exclusive).

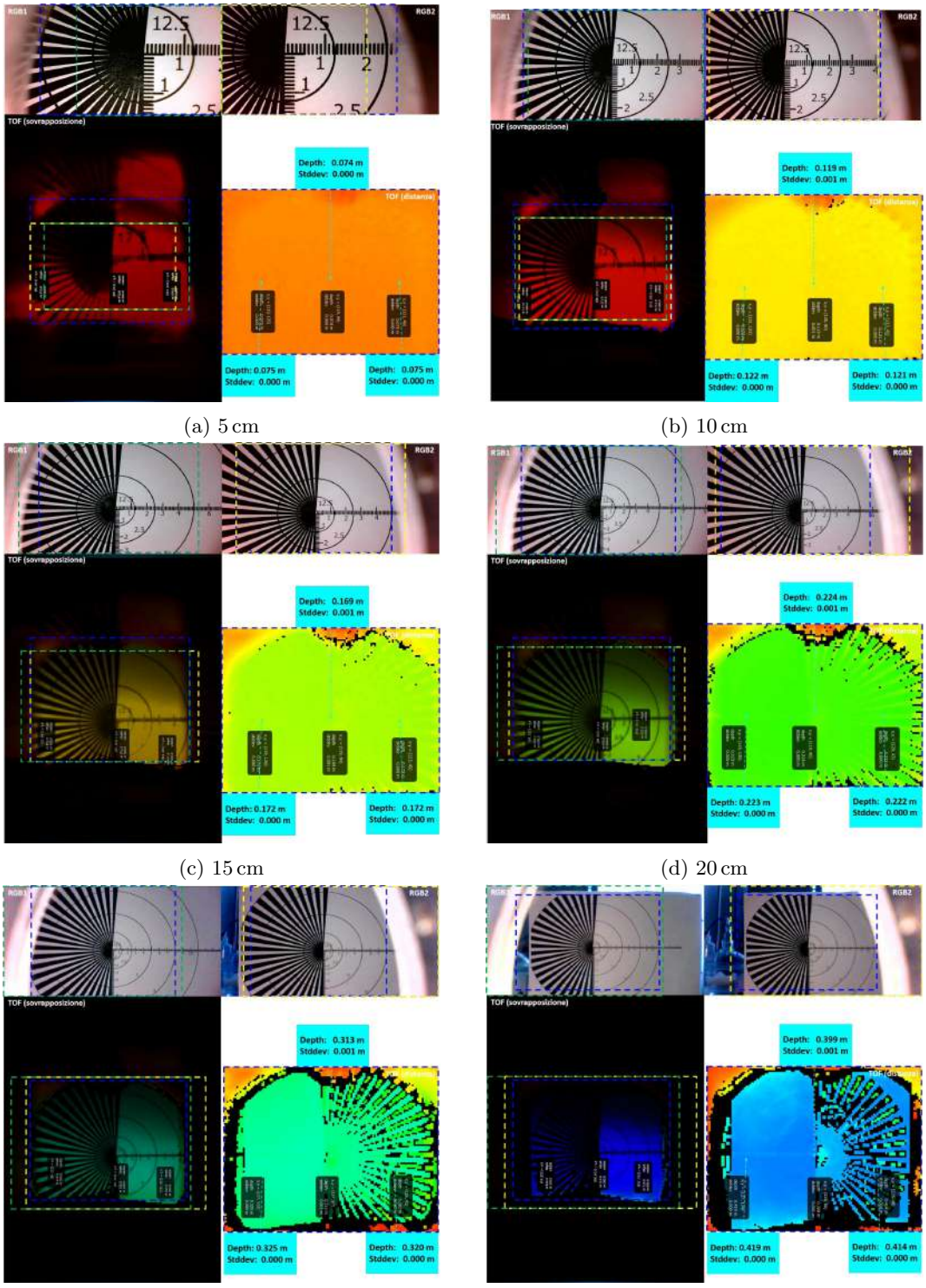
Since the calibration board is simplified, we only need two filters: area and circularity. Once we obtain the pixel coordinates of the three blobs in the image plane, we can map them to the point cloud and consequently to their poses P in the 3D space.

Once the pose set P is obtained we can proceed with the calibration procedure presented in [161] to find out the best fitting plane and the final Hand-eye calibration step. Figure 2.12 shows the qualitative results of the calibration by touching with the instrument one of the reference points used to calibrate the system.

2.10 Discussion and Conclusions

The calibration of the endoscope allowed to perform a required mapping procedure to correctly position the camera-wielding robot in the 3D space surrounding the operational scenario. The endoscope did suffer from noisy measurements that induced errors in the camera orientation and, thus in subsequent errors in positioning itself in the space, but the adopted procedure with RANSAC proved robust enough for maintaining the errors manageable. This is primarily a sign of the prototyping status of the current platform which can clearly be refined in its overall optical geometry and materials to operate more

precisely in the infrared spectrum adopted by the ToF sensor and emitter. Nevertheless, if we perform a direct comparison to the technologies available on the market, it is possible to find solutions that provide a higher visible image quality, like the current da Vinci® endoscope technology offered to surgeons, or joint ToF and stereoscopic infrared reconstruction provided by Intel RealSense®. This camera represents an innovative device to combine current and future surgical robotic applications. Being just a prototype, additional work is required to reduce the bulky optical solution hardware and the high operation temperatures that preclude its adoption to minimally-invasive applications.

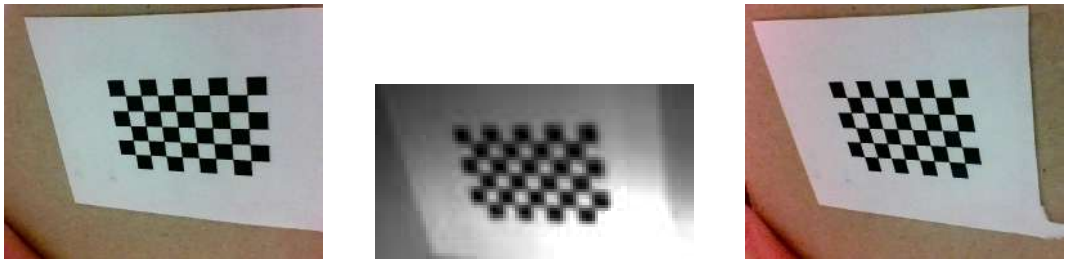


(a) 5 cm (b) 10 cm

(c) 15 cm (d) 20 cm

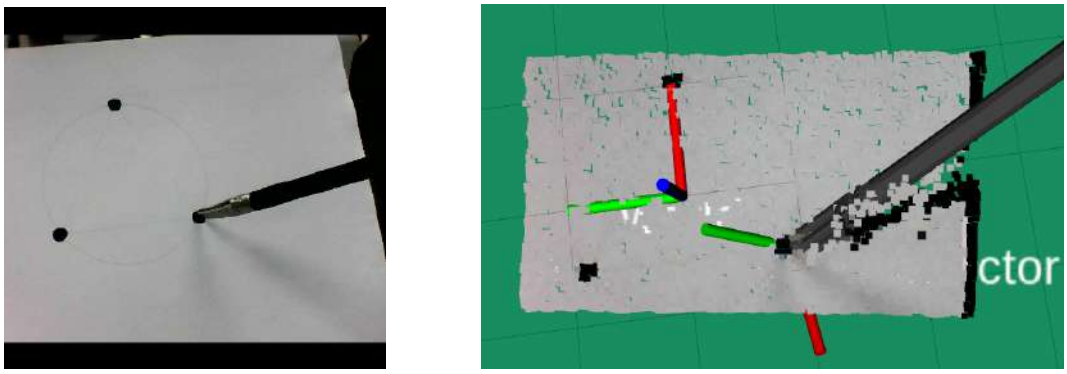
(e) 30 cm (f) 40 cm

Fig. 2.10: Endoscope characterization at various working distances. The blue, yellow and green dotted rectangles are the FoVs for the depth, left and right camera respectively.



(a) Left camera (b) Tof sensor (c) Right camera

Fig. 2.11: An example of the images used during the calibration procedure



(a) 2D (b) 3D

Fig. 2.12: Qualitative result of the hand-eye calibration: with the instrument we touch one of the reference points used for the calibration (a) and its respective 3D point cloud with the world reference frame obtained with the calibration

3

Camera calibration for robotic surgery

Autonomy requires systems with advanced perception, reasoning and motion planning, as highlighted in [31, 50]. Specifically, better medical imaging and vision techniques have significantly improved the performance of robotic surgical systems in a range of clinical scenarios, such as orthopaedics and neurosurgery [1]. Vision systems can retrieve pre and intra operative information from tomography (CT) [148], magnetic resonance (MR) and ultrasound to plan toll trajectories and support surgeons' decision making. However, image-guided interventions require an accurate calibration to map poses of robots, instruments and anatomy to a common reference frame.

In order for a robot to use a video camera to estimate the 3D position and orientation of a part or object relative to its own base within the work volume, it is necessary to know the relative position and orientation between the hand and the robot base, between the camera and the hand, and between the object and the camera. These three tasks require the calibration of robot, robot eye-to-hand, and camera. These three tasks normally require large-scale nonlinear optimization, special setup, and expert skills.

3D robotics hand/eye calibration is the task of computing the relative 3D position and orientation between the camera and the robot gripper in an eye-on-hand configuration, meaning that the camera is rigidly connected to the robot gripper. The camera is either grasped by the gripper, or just fastened to it. More specifically, this is the task of computing the relative rotation and translation (homogeneous transformation) between two coordinate frames, one centered at the camera lens center, and the other at the robot gripper. The gripper coordinate frame is centered on the last link of the robot manipulator and it must possess enough degrees of freedom so as to be able to rotate the camera around two different axes while at the same time keeping the camera focused on a stationary calibration object in order to resolve uniquely the full 3D geometric relationship between the camera and the gripper.

3.1 Theory of calibration

Hand-eye calibration has been widely studied within the robotics literature [172]. The most common way to describe the hand-eye calibration problem is using the homogeneous transformation matrix as:

$$AX = XB \tag{3.1}$$

where A and B are known homogeneous matrices, and X is the unknown transformation between the robot coordinate frame and camera coordinate frame. For each homogeneous matrix, it is in the form of

$$\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{3.2}$$

where R is a 3×3 rotational matrix, and t is a 3×1 translational vector. Thus, we can expand (3.1) as

$$\begin{bmatrix} R_A & t_A \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_x & t_x \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_x & t_x \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_B & t_B \\ 0 & 1 \end{bmatrix} \tag{3.3}$$

where R_A, R_X and R_B are the rotational matrix parts of A,X and B, and t_A, t_X and t_B are the translational parts, respectively. Equation (3.3) can be further simplified as:

$$\begin{bmatrix} R_A R_X & R_A t_X + t_A \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_X R_B & R_X t_B + t_X \\ 0 & 1 \end{bmatrix} \tag{3.4}$$

The purpose of hand-eye calibration is to find R_X and t_X given j pair of A_i and the corresponding B_i, where $i = 1, 2 \ldots J$. The majority of the approaches regards the rotation estimation decoupled from the translation estimation. At least two rotations containing motions with nonparallel rotation axes are required to solve the problem (Tsai and Lenz 1989). Several approaches have been proposed for the estimation of R_X from (3.4): using the rotation axis and angle [174, 196], quaternions [22] and canonical representation [97].

3.1.1 State of the art

However, the first simultaneous consideration of rotation and translation in a geometric way was presented by Chen (1991) [18], who first introduced the screw theory in the hand-eye calibration. Daniilidis introduced the algebraic entity for a screw: the unit dual quaternion [28].

The dual quaternions approach proves that:

- the hand-eye transformation is independent of the angle and the pitch of the camera and hand motions, and depends only on the line parameters of their screw axes.
- the unknown screw parameters, including both rotation and translation, can be simultaneously recovered using the singular value decomposition (SVD).

In R-MIS systems, where the patient-side arms are constrained by a Remote Center of Motion (RCM), it is challenging to obtain the camera motion range needed to guarantee an accurate calibration. Wang [202] takes advantage of this constraint by finding a unique relationship between the endoscope and the surgical tool using camera perspective projection geometry. A different approach

is followed in [140, 222] where the instruments themselves are used as calibration tools. Thus far, several closed-form solutions for 2d images have been proposed for hand-eye calibration that use linear methods that separate rotations and translations. In [174], the orientation component was derived by utilizing the angle-axis formulation of rotation, then the translational component was estimated using standard linear systems techniques. Chou and Kamel [22] introduced quaternions to represent orientation and solved the quaternion coefficients as a homogeneous linear least squares problem. A closed form solution was then derived using the generalized inverse method with singular value decomposition analysis. Other works [98, 130, 142] used the Kronecker product to get a homogeneous linear equation for the rotation matrix. However, separating the rotational and translational components neglects the intrinsic correlation between them. Working directly in 3D space is then a better solution. In [80] the authors studied the comparison between hand-eye calibration based on 2D and 3D images, introducing quantitative 2D and 3D error metrics to assess the calibration accuracy. They proved that the 3D calibration approach provides more accurate results on average but requires burdensome manual preparation and much more computation time than 2D approaches. Kim used 3D measurements at the center of markers for the hand-eye calibration [82]. Fuchs [54] proposed a solution based on depth measurements instead of 2D images, using a calibration plane with known position and orientation. The hand-eye calibration was then obtained by estimating the best fitting calibration plane of the measured depth values.

We propose a novel calibration method for the surgical robotic scenario using the da Vinci® Research Kit (dVRK) and an RGB-D camera. Differently from [54], the accuracy and computational time of our method do not depend on the placement of the calibration board within the workspace. We perform exhaustive experimental validation on relevant use cases for surgery. We separate the calibration of the robotic arms (two Patient-Side Manipulators, PSM1 and PSM2, and an Endoscope Camera Manipulator, ECM) from the hand-eye calibration of the camera. For both calibrations we propose a three-step method with closed-form solution:

1. touching reference points on a custom calibration board with the end-effectors of the surgical robot.
2. recognizing the same reference points with the RGB-D camera.
3. mapping the poses reached by the robotic arms in the first step to the 3D points computed in the second step.

The main advantage of the proposed method is the improved accuracy in a 3D metric space, which is increased by a factor of four with respect to the state-of-the-art results with comparable sensors [80]. Moreover, with our method the camera can be mounted on the moving endoscopic arm of the dVRK, overcoming the limitations of a fixed camera.

In the following sections we describe our calibration technique and the setup used to test our method. In Section 3.4 we describe the validation of the proposed method by evaluating the workspace through simple kinematic tasks. We also compare our calibration method with Tsai's [196], which is the gold

standard for hand-eye calibration, in two different tasks: grasping and camera projection to 3D space. Finally we present our conclusions and plans for future works.

3.2 Proposed method

The aim of the calibration procedure is twofold. First, we perform computation of the rigid transformations $T_{\star}^{w}$ between the common reference frame (*world*) and the base frame of the arms, $\star \in \{ecm_b, psm_{1b}, psm_{2b}\}$. Second, we estimate the transformation T_{ecm}^{cam} between the camera reference frame and the ECM reference frame. The resulting transformation tree is shown in Figure 3.1.

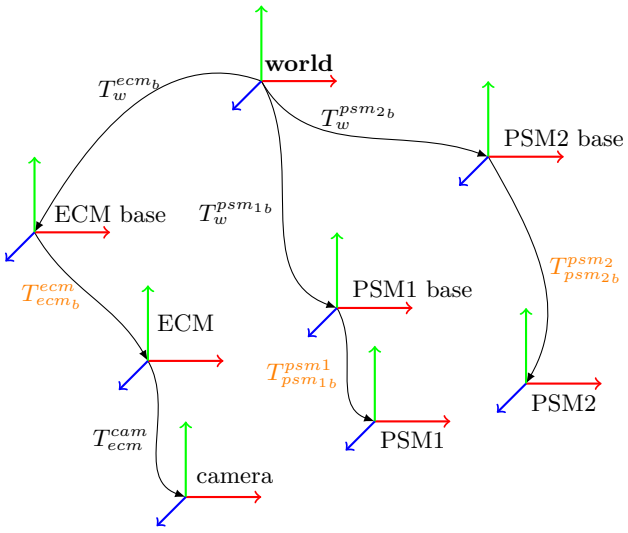
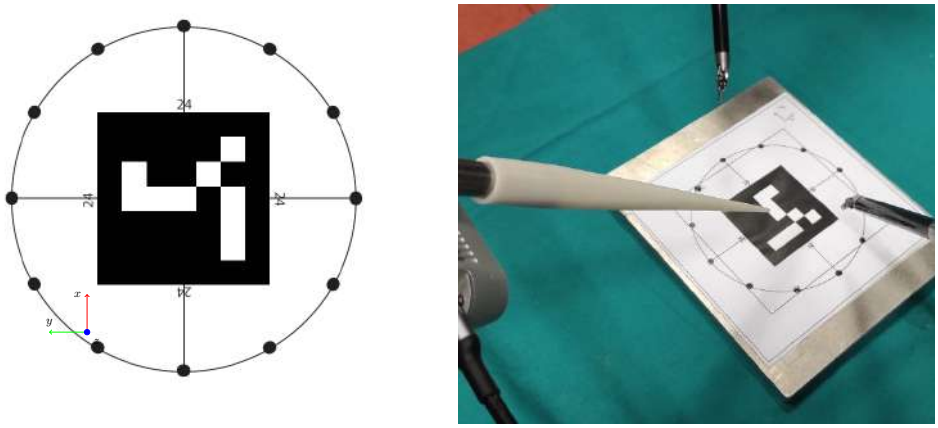


Fig. 3.1: The reference frames produced by our proposed method (the axes direction of the reference frames are only for visualisation purpose). The orange transformations are known, whereas the black transformations are to be estimated.

We use a custom calibration board, shown in Figure 3.2a, with an ArUco marker in the center of a circumference of 50 mm radius, with several reference dots. We equipped the ECM with a 3D-printed adapter, shown in Figure 3.2b. The adapter has a smaller tip than the ECM to guarantee precise positioning on the dots on the board.

The procedure starts by positioning the calibration board in the robot workspace. We choose a set of reference points P such that each point $p \in P$ is reachable by the three arms and visible from the camera. The points in P must be symmetric with respect to the center of the board to compute the origin of the common reference frame; at least three points are needed to estimate the plane coefficients. The best fitting plane is characterized by the centroid of the point set P, $\mathbf{c}$, and the normal vector $\mathbf{n}$. Their optimal estimations are the solution of the optimisation problem



(a) Calibration pattern (b) ECM adapter

Fig. 3.2: The calibration components. a) the calibration board with the marker, the coloured axes represents the common reference frame directions b) the adapter for the ECM positioning.

$$\{\hat{\mathbf{c}}, \hat{\mathbf{n}}\} = \underset{\mathbf{c},|\mathbf{n}|_2=1}{\arg\min} \sum_{i=1}^{n} ((\mathbf{p}_i - \mathbf{c})^T \mathbf{n})^2 \tag{3.5}$$

As in [58] the centroid is estimated by

$$\hat{\mathbf{c}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{p}_i. \tag{3.6}$$

The normal vector $\mathbf{n}$ is obtained by factorizing the distance matrix A with Singular Value Decomposition (SVD)

$$A = USV^T = \left[\mathbf{p}_1 - \hat{\mathbf{c}}, \dots \mathbf{p}_n - \hat{\mathbf{c}}\right] \in \mathbb{R}^{3\times n} \tag{3.7}$$

and taking the third column of the matrix $U = \left[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\right]$, $\hat{\mathbf{n}} = \mathbf{u}_3$. To generate a common reference frame for all the tools we implement the following three main steps:

(1) Arm calibration
(2) Camera calibration
(3) Hand-eye calibration

3.2.1 Arm calibration

To find the transformation of the arms base frame with respect to the common reference frame we record the end effector pose of the arms (PSMs and ECM with adapter) on each point in the set P. In order to obtain the ECM effective pose, we remove the known rigid transformation between the adapter and the ECM. On this set we estimate the best fitting plane using (3.5). The set P is then augmented by adding a point above the calibration board acquired by moving the arm's end effector. This last point is used to define the desired plane normal direction

$$\mathbf{n}_d = \frac{\mathbf{p}_{n+1} - \mathbf{c}}{|\mathbf{p}_{n+1} - \mathbf{c}|_2}$$

where $\mathbf{p_{n+1}}$ is the last point in the ordered set P, $\mathbf{c}$ is the centroid of P and $|*|_2$ is the vector norm. For each arm, the homogeneous transformation $T^w_\star$ of the common reference with respect to the arm base frame is defined using the direction versors

$$\begin{aligned}\mathbf{u} &= \text{sign}(\mathbf{n} \cdot \mathbf{n_d})\mathbf{n} \\ \mathbf{l} &= \mathbf{u} \times \frac{\mathbf{p}_1 - \mathbf{c}}{|\mathbf{p}_1 - \mathbf{c}|_2} \\ \mathbf{f} &= \mathbf{l} \times \mathbf{u}\end{aligned}$$

and the centroid $\mathbf{c}$,

$$T^w_\star = \begin{bmatrix} \mathbf{f}_x & \mathbf{l}_x & \mathbf{u}_x & \mathbf{c}_x \\ \mathbf{f}_y & \mathbf{l}_y & \mathbf{u}_y & \mathbf{c}_y \\ \mathbf{f}_z & \mathbf{l}_z & \mathbf{u}_z & \mathbf{c}_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.2.2 Camera calibration

To find the transformation T^w_{cam} for the RGB-D camera we first detect the center of the ArUco marker on the board with respect to the camera frame. Once we find a camera position that ensures good visibility and a stable pose of the ArUco marker, we align the pose on the point cloud generated from the depth map acquired by the RGB-D camera. We use the marker pose and its known radius to generate the pose of every dot in the set P in the marker reference frame, as well as the point above the calibration board that is needed to define the desired plane normal direction.

Once the pose set P is obtained we find the best fitting plane using (3.5) and then we build the homogeneous transformation T^w_{cam} between the common reference frame to the camera base frame by adapting the previous approach used for the arms.

3.2.3 Hand-eye calibration

The hand-eye calibration problem is formulated using the homogeneous transformation matrices:

$$AX = XB$$

where A and B are known homogeneous matrices representing the frames of the base of the robot and the camera, respectively. The unknown transformation X is between the robot coordinate frame and the camera coordinate frame. Given T^w_{cam}, we can compute X as the relative homogeneous transformation between the end effector of the ECM and the RGB-D base frame:

$$T^{cam}_{ecm} = T^{cam}_w (T^{ecm}_w)^{-1}.$$

3.3 Experimental setup

The validation of the proposed method has been carried out with the dVRK robot shown in Figure 3.3.

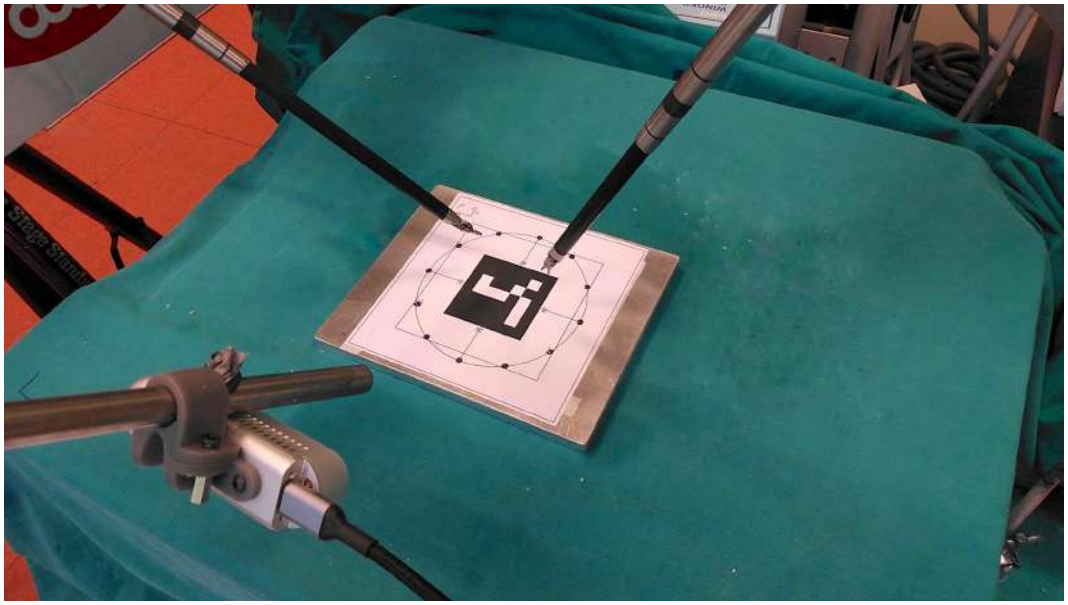


Fig. 3.3: The proposed setup for calibration, with the RealSense d435, the PSMs and the calibration pattern.

The stereo endoscope has been augmented with an Intel RealSense d435 RGB-D camera rigidly attached to the endoscope through a 3D printed adapter as explained in the previous chapter. The camera specifications are reported in Table 3.1. The whole calibration method has been implemented in Robot Operating System (ROS) using the Point Cloud Library (PCL) and OpenCV. The present setup is not compatible with a surgical scenario. However it is well possible that in the near future small RGBD cameras could be integrated within the endoscope.

Table 3.1: RealSense d435 specifications

Camera specifications	
Resolution	1280×720
Field of view (FOV)	$91^\circ \times 65^\circ \times 100^\circ$
Frame rate	90 fps
Baseline	50 mm
Z-accuracy	$\leq 2\%$ of the working distance

3.4 Experimental results

To experimentally validate our methodology we compared our calibration with the Tsai's method [196] in two benchmark tests for surgical robotics:

- Localization and grasping of small targets,
- Dual-arm manipulation

Finally we evaluated the accuracy of the projection from 2D camera image plane to the 3D workspace.

3.4.1 Localization and grasping

In the first scenario (Figure 3.4) the two PSMs must autonomously grasp a ring placed on the calibration board, in this case on location 2. The RGB-D

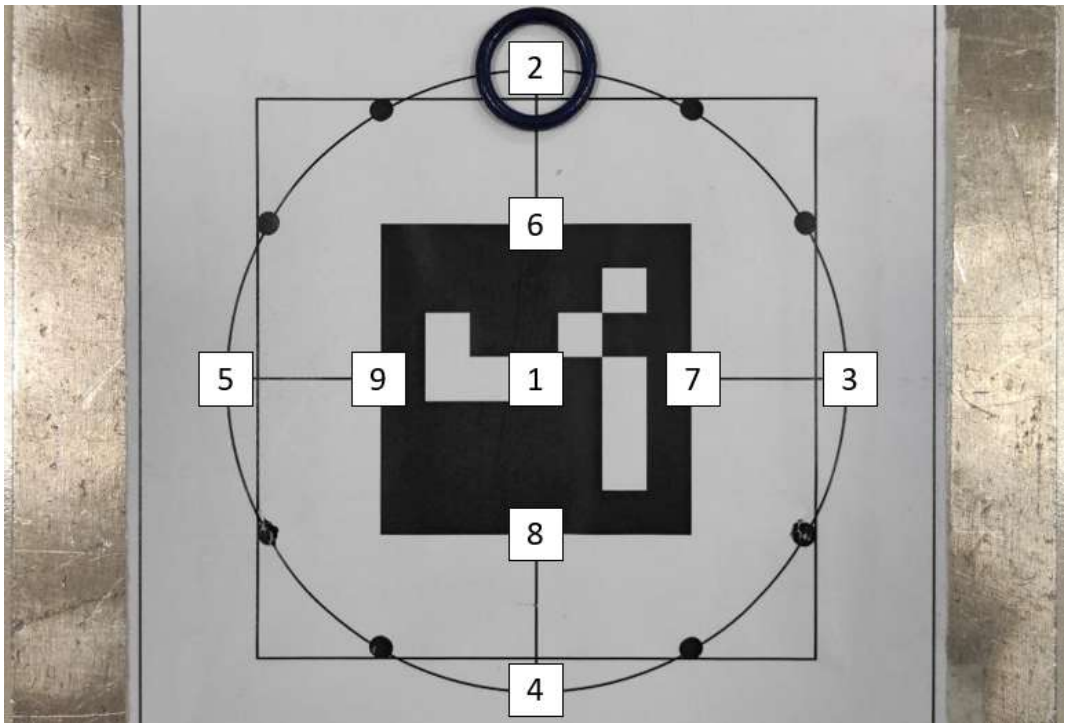


Fig. 3.4: Setup for the localization and grasping experiment. The numbers on calibration board represents the nine locations used during the experiment. The ring is identified by the camera and then reached by the PSMs.

camera identifies the point cloud corresponding to the ring after color and shape segmentation, and points are transformed from the camera to the common reference frame. The ring has a diameter of 15 mm, and the target point for both PSMs is chosen as the center of the ring. The ring is placed in the 9 different locations on the board to cover the full $x-y$ plane, as shown in Figure 3.4. The arms reach the target points ten times, and for each iteration we compute the Euclidean distance between the target and the final positions of the PSMs. In this way, we estimate the mean accuracy of our calibration procedure on the $x-y$ plane. The results are reported in Figure 3.5 and compared with state-of-the-art Tsai's calibration method [196]. It is worth mentioning that errors are comprehensive of the estimated kinematic accuracy of the da Vinci®: 1.02 mm on average when localizing and reaching fiducial markers [65], with a maximum error of 2.72 mm [91].

Table 3.2: A comparison of the error in the localization and grasping test

	Max error (mm)	Mean error (mm)	Std dev (mm)
Our method	1.07	0.53	0.15
Tsai [196]	3.17	1.83	0.33

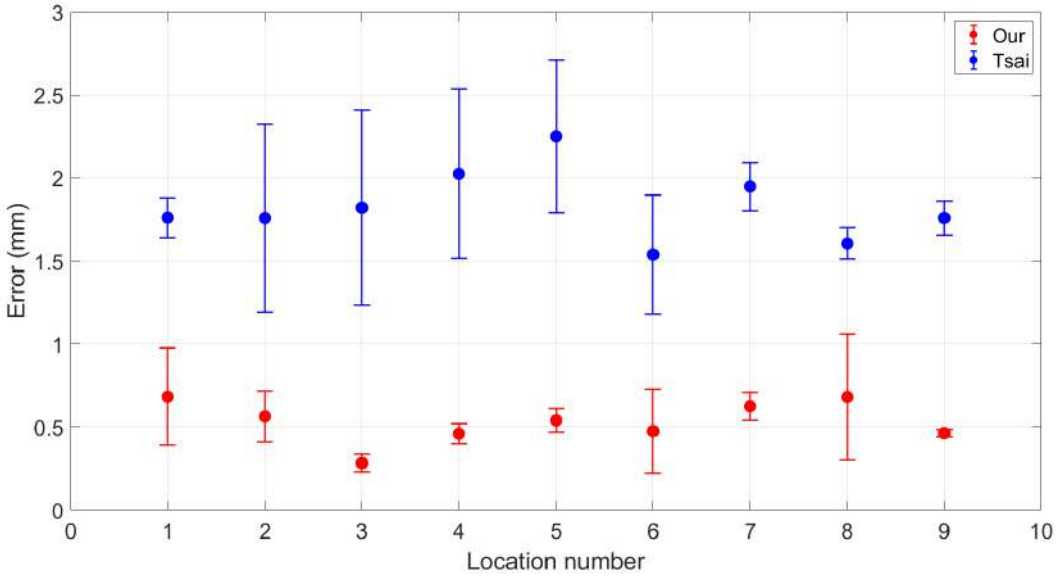


Fig. 3.5: The measured 3D positioning errors between the robot end effector and the grasping point

Table 3.2 shows that our method achieves significantly better accuracy (0.53 mm average error against 1.83 mm with Tsai's calibration). The error does not depend on the location of the ring on the $x - y$ plane.

3.4.2 Dual arm manipulation

In the second scenario (Figure 3.6) the PSMs start holding the same ring, and they must execute simultaneous pre-computed circular trajectories with center on the z axis of the common reference frame (45 mm above the calibration board) and radius r ranging from 10 mm to 40 mm. Circumferences are first defined in the $x - z$ plane of the common reference frame (normal to the calibration board), and then replicated in planes rotated around the z axis with a step of 10 deg. In this way we define a spherical workspace by interpolation between the recorded trajectories. PSMs are commanded with the transformed waypoints in their relative frames. This task validates the accuracy of the transformations between the arms computed with the proposed method. We measure the difference between the trajectories of the two PSMs, and we consider the standard and the maximum deviations from the mean for each radius. In absence of calibration and kinematic errors, the difference between the trajectories would have null standard deviation. Figure 3.7 shows the absolute error through the workspace for spheres with radii 20 mm, 30 mm

Fig. 3.6: Dual arm manipulation experiment. The two arms carries a ring while performing circular trajectories through the workspace.

and 40 mm, by using the Lambert equal-area cylindrical projection [203]. In Table 3.3 we report the errors for all the spheres. We notice that the mean error increases with the radius of the sphere, as the PSMs move away from the calibration plane. The standard deviation of the error increases with the radius but remains below 0.11 mm, hence the overall error does not change significantly on the surface of the spheres. This ensures good repeatability of motions in the whole workspace. The accuracy of our calibration method in 3D is compatible with the requirements of surgery (the mean error between the arms is below 1 mm, comparable with the known kinematic accuracy of the da Vinci®).

Table 3.3: The positioning error between the PSM1 and PSM2 during the dual-arm manipulation experiment

Radius (mm)	Max error (mm)	Mean error (mm)	Std dev (mm)
10	0.61	0.11	0.06
20	0.37	0.13	0.08
30	0.51	0.14	0.10
40	0.62	0.16	0.11

3.4.3 2D/3D projection

In the last scenario the PSM1, with a colored marker on its tip, executes a spiral-shaped trajectory along the entire workspace. The RGB-D camera

identifies the marker in the image plane, and the corresponding 3D point can be computed using the depth value. The trajectory starts near the origin of the common reference frame and then increases in radius and altitude according to the following parametric equations

$$x(t) = \kappa t \cos(\omega t)$$
$$y(t) = \kappa t \sin(\omega t)$$
$$z(t) = \kappa t$$

where ω is the constant angular speed and $\kappa \in \mathbb{R}$ is a time-scaling factor. The orientation of the end effector is kept fixed towards the camera along the trajectory. We measure the Euclidean error between the points in the trajectory executed by the arm and the re-projected points from the camera image plane. Figure 3.8 and Table 3.4 show that the re-projection accuracy with our method significantly outperforms the one reached with Tsai's. In fact, the mean error (4.71 mm) and the maximum error (11.76 mm) are four and two times smaller than the one achieved by Tsai's method. It is important to remark that the measured error also includes the marker detection accuracy.

Table 3.4: A comparison of the error between the marker tip trajectory and the measured tip trajectory for the projection test

	Max error (mm)	Mean error (mm)	Std dev (mm)
Our method	11.76	4.71	0.89
Tsai [196]	20.85	16.41	1.21

Finally Figure 3.9 shows the re-projection of PSMs end effector position onto the camera image plane with both calibration methods. Our method achieves a better re-projection better of the 3D instruments.

3.5 Discussion and Conclusions

In this work we proposed a novel 3D calibration procedure for the patient-side manipulators and the ECM of the da Vinci® surgical robot. Our procedure exploits an RGB-D Realsense camera. We have validated our calibration procedure by evaluating the 2D/3D projection errors on two relevant use cases for surgery localization and grasping of a small object and dual-arm manipulation. Both tasks require an accurate estimation of the transformation tree connecting the arms and the camera, to guarantee precise positioning and coordination of the PSMs. In our experiments the proposed method outperforms the state-of-the-art solution proposed by Tsai. Our method reaches an accuracy below 1 mm on the $x-y$ plane and in the dual arm manipulation scenario, which is comparable with the intrinsic kinematic precision of the da Vinci®.

The main drawback of our solution is the use of a RGB-D camera, which limits its actual application in surgery. We think that our methodology can be

extended to a setup with a standard surgical endoscope. The main issue with an endoscope is that the small baseline between the stereo cameras introduces additional complexities in computing depth maps and reduces the depth range of view. We will address this problem in our future research. Moreover, we will develop an autonomous procedure for our calibration method, which can significantly reduce manual errors and simplify its implementation in a surgical setup.

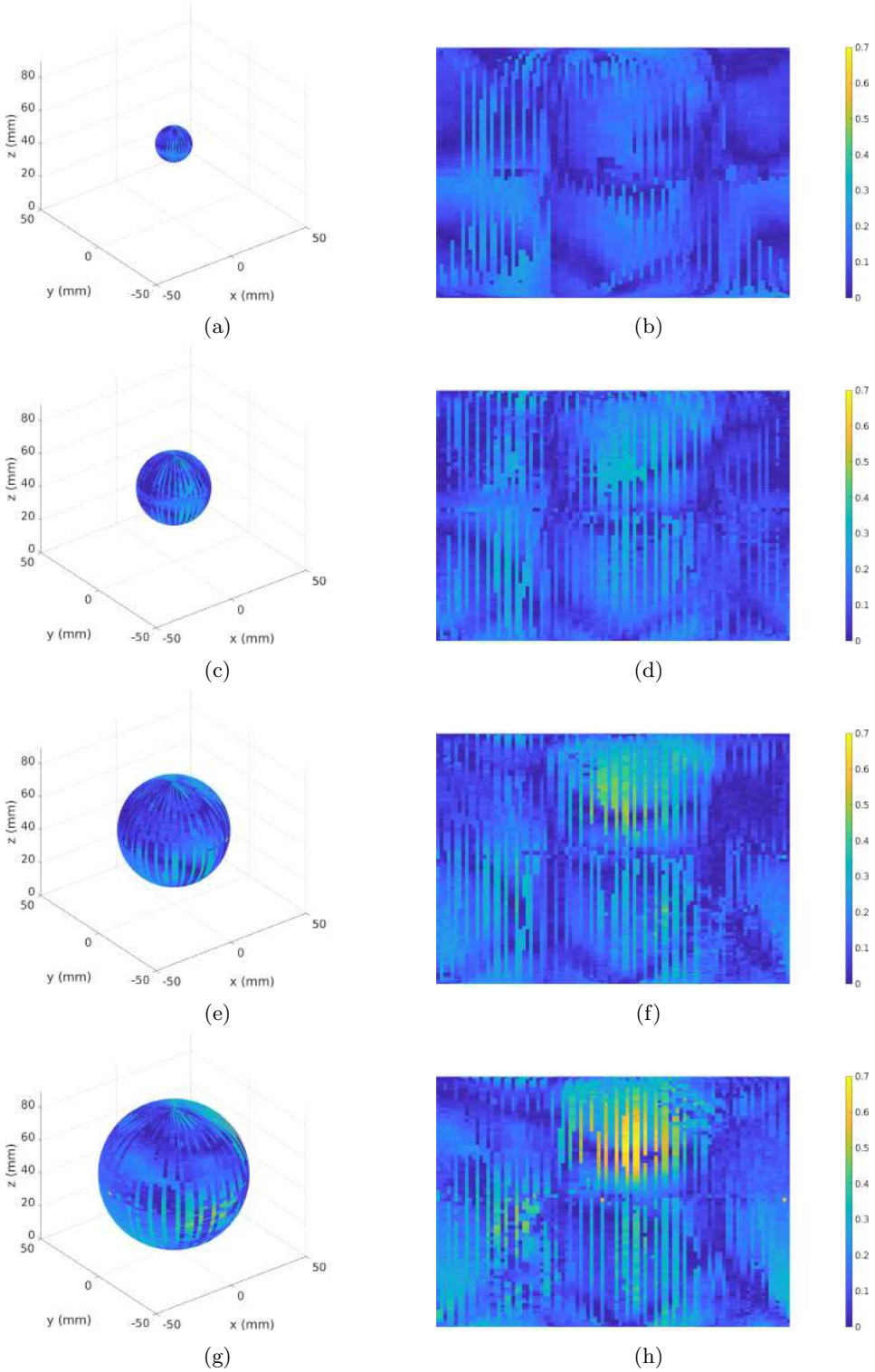
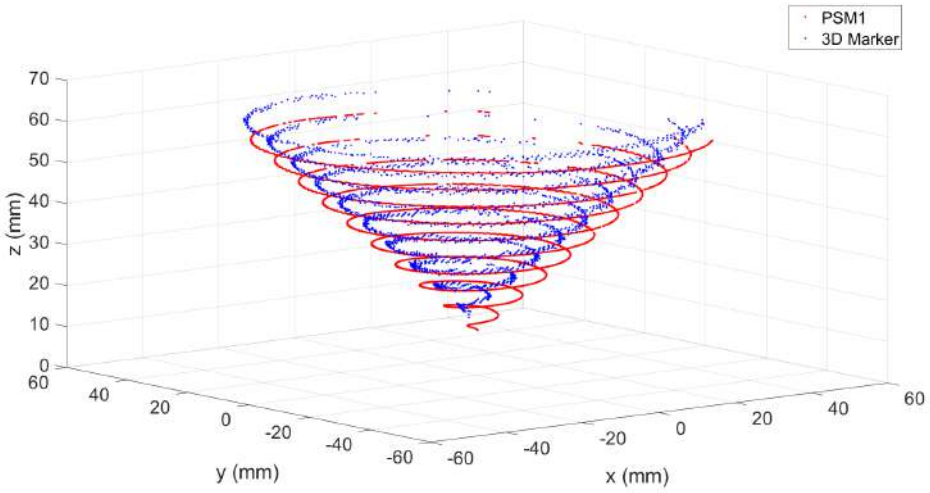
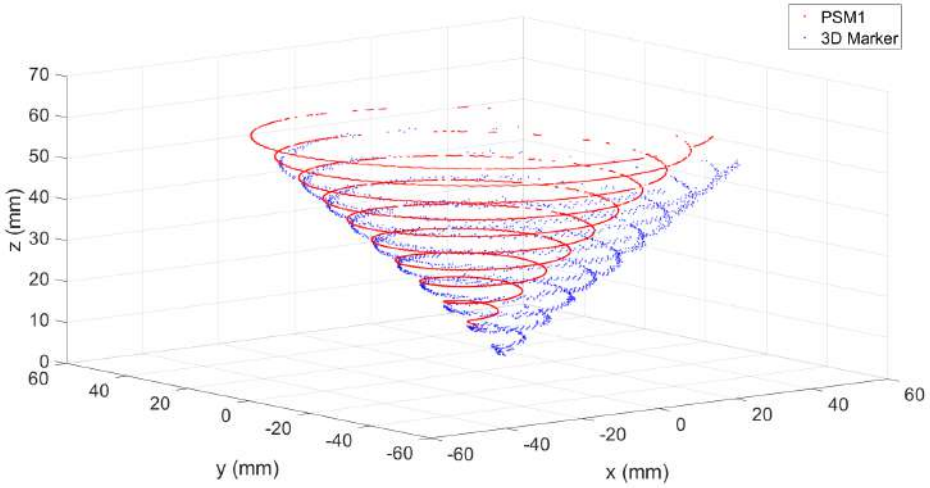


Fig. 3.7: Absolute error of the dual arm manipulation through the workspace. The workspace has been projected using the Lambert equal-area cylindrical projection, the error is reported in mm. a) the workspace surface of the sphere with radius 10 mm, b) the projected surface of the sphere with radius 10 mm, c) the workspace surface of the sphere with radius 20 mm, d) the projected surface of the sphere with radius 20 mm, e) the workspace surface of the sphere with radius 30 mm, f) the projected surface of the sphere with radius 30 mm, g) the workspace surface of the sphere with radius 40 mm, h) the projected surface of the sphere with radius 40 mm.

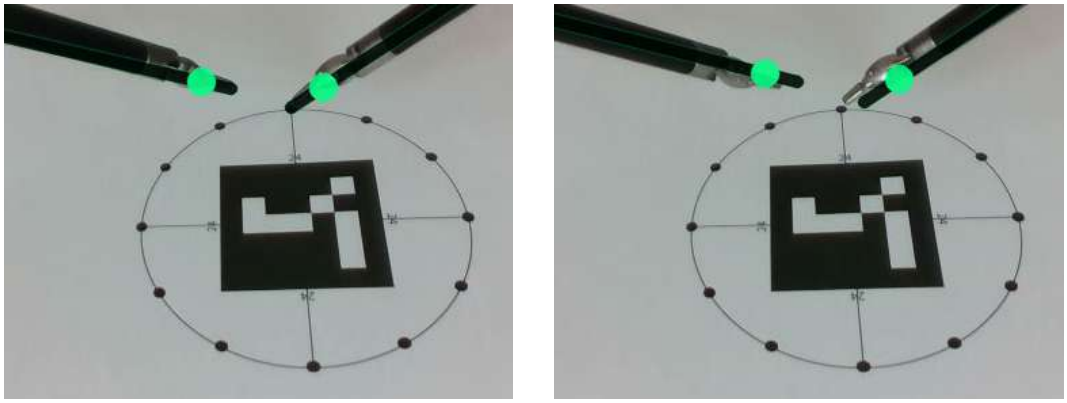


(a)



(b)

Fig. 3.8: Spiral-shaped trajectory executed by the PSM1 with our method in (a) and Tsai's method in (b). The red trajectory represents the kinematics of the PSM1, while the blue trajectory represents the marker identified in 3D space.



(a) Our method (b) Tsai's method

Fig. 3.9: An example of re-projection of da Vinci® surgical instruments by using kinematic re-projection of the model directly onto camera color image.

4

Prerequisites for autonomy in surgery

In the first chapters of this thesis we highlighted the importance of sensors and theirs calibration with a robot. The computer vision is the process of extracting information from a scene by analyzing the image. The vision sensors allow you to measure the environment without contact and in a global way the ever-increasing speed of computers and the improvement of image analysis techniques allow high performance even in the case of low-cost cameras.

In this chapter we present the experiments carried out to validate the accuracy of the calibration and of the algorithms designed for the development of autonomous applications in surgery performed on different scenarios, thought of as necessary prerequisites for autonomy. The first work is a about a benchmark training task for surgeons, the ring transfer from Fundamentals of Laparoscopic Surgery (FLS), consisting in placing rings on same colored pegs and requiring coordination of multiple actions depending on dynamic environment conditions. Other works instead address the recognition of features of objects of interest within the intervention, such as instruments or catheters during radical prostatectomy procedure, to trigger the state machine that represents the intervention to move to the next state.

4.1 Peg and ring

This task replicates several challenges of real surgery and is executed with the research version of the established surgical da Vinci robot, the da Vinci Research Kit (dVRK). As the original da Vinci system, dVRK consists of a patient side and a remote control side. The remote control side is equipped with a console for an expert surgeon. Through the console, the surgeon can tele-operate the instruments on the patient side, which consists of two cable-driven robotic arms with 6 degrees of freedom (DOFs), each called Patient-Side Manipulators (PSMs), and one Endoscopic Camera Manipulator (ECM) with 6 DOFs. The setup for the bimanual ring transfer task with dVRK is shown in Figure 4.1. The task consists of placing colored rings (4 rings with colors red, green, blue and yellow) on the same-color pegs. In standard task description from FLS, each ring is initially placed on a grey peg, and it must be transferred between arms of dVRK before placing on the corresponding peg. The standard task definition hence only requires the execution of a pre-defined sequence of

actions (moving to a ring, grasping it, transferring to the other arm and placing on the peg) without significant variations in the workflow, except for the failure condition when a ring may fall during the motion of a PSM.

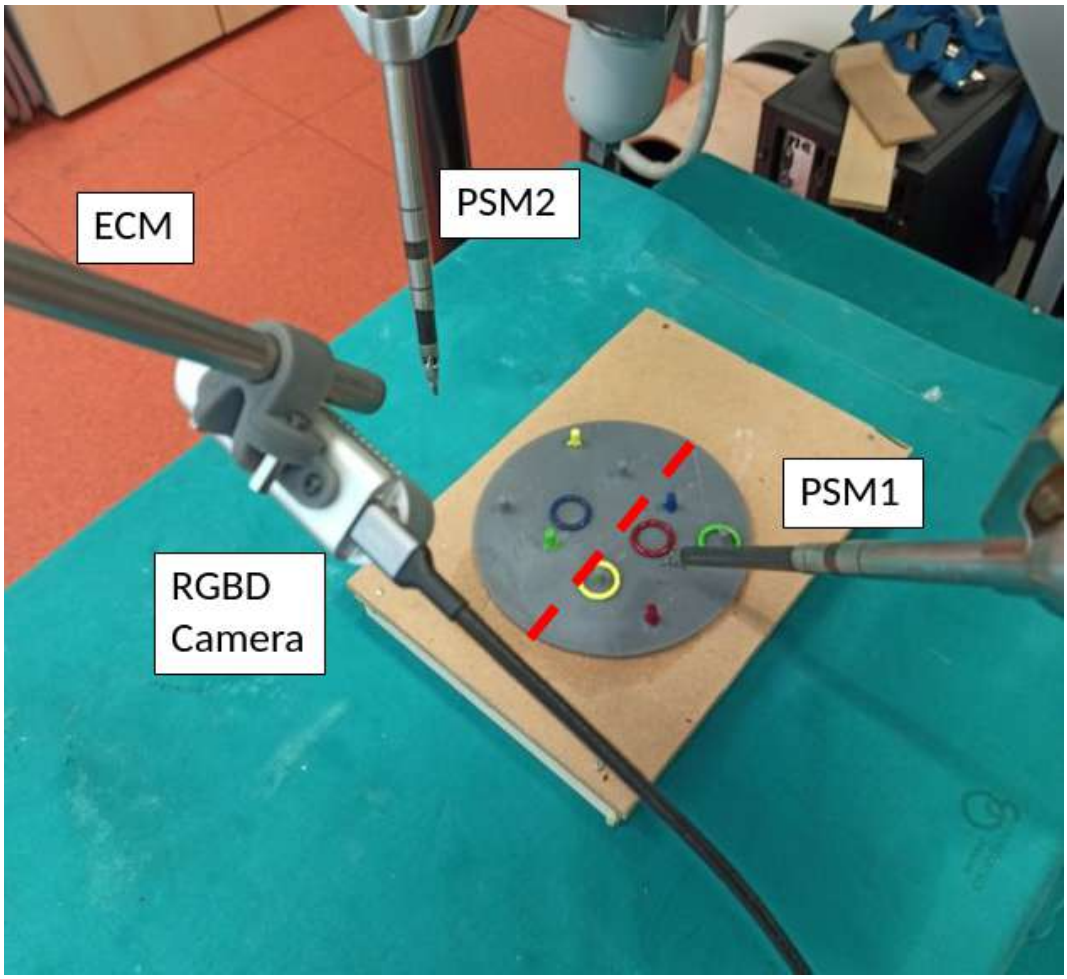


Fig. 4.1: The setup for the ring transfer task. The red dashed line defines reachability regions for the two arms.

In order to execute the task autonomously, the robotic system must be equipped with sensors to deal with the dynamic scenario. Though the endoscopic camera could be directly used, the localization accuracy typically suffers from the small baseline between the stereo cameras. For this reason, in the experiments in this thesis a Realsense D435 RGBD camera is mounted with a proper adapter on the fixed ECM, as explained in chapter 2 and shown in Figure 2.6. The chosen camera has 105mm minimal depth range of view. Before the task execution begins, the camera and the PSMs must be calibrated, in order to define a common reference frame (from now on, referred to as world) for localization and motion control. Then, an algorithm for real-time object recognition is executed, to identify and localize rings and pegs.

Algorithm 1 Detection Algorithm

Input: Point cloud $PC_{in}(t)$ in real time
Output: Point clouds of rings, PC_r, and pegs, PC_p
Initialize: $PC_r = []$, $PC_p = []$
for $t = 1 : \infty$ **do**
    Subsample $PC_{in}(t)$ to $PC_{sub}(t)$
    for each color **do**
        Color Segmentation of $PC_{sub}(t)$
        Euclidean clustering
        Update $PC_r \leftarrow$ RANSAC
        if $t == 1$ **then**
            Update PC_p
        end if
    end for
end for

Fig. 4.2: Vision Algorithm.

Thanks to the calibration procedure, the poses of the robotic arms as read from the built-in encoders of the dVRK can be easily transformed in the world frame, and related to the poses of the point cloud returned by the camera. From the point cloud, relevant objects in the ring transfer scenario must be identified using an appropriate 4.2. The algorithm is based on standard methods from the well-established Point Cloud Library. The point cloud is subsampled in order to guarantee real-time performance. The base and the pegs are assumed to be static during the whole execution, and they are identified only at the beginning of the task. The poses of all rings are retrieved at each time step. The identification of pegs and rings is performed in two steps. First, color segmentation allows to identify same-colored points. Then, Euclidean clustering allows to separate the clouds of ring and peg for each color. Finally, Random Sample Consensus (RANSAC) [51] is used to fit a torus shape on both clusters, and the best fitting cluster is identified as the ring, while the other as the peg. The output of 4.2 is the point cloud of rings and pegs (Figure 4.3).

4.2 Perception module

One of the most difficult challenges is to define an architecture to be able to perform autonomous or semi-autonomous operations. Within the SARAS project we designed a cognitive architecture (Figure 4.4) to control a surgical robot, which integrates pre-operative and intra-operative data, manages the multimodal interaction with the principal surgeon and the environment.

The *perception module* takes care of understanding the complexity of the surgical area, by reconstructing, labelling and tracking all its elements as observed by the available sensors (i.e. videos, kinematics and forces). this was a

great benchmark for testing calibration work and designing new technologies and making new observations.

The core of the cognitive architecture is the supervisory controller for a surgical semi-autonomous robotic platform, which uses a three-level Hierarchical Finite State Machine (HFSM) to define all the possible behaviours of the autonomous system. The transitions of the HFSM are triggered by the *Observers*, a set of functions fed with the state of the system (robot kinematics, anatomical structures, etc.) that output a logical description of the surgery state. We tested the supervisory controller performing the "bladder neck incision" phase of a Radical Prostatectomy (RARP) procedure.

The HFSM has three layers defined as: (1) **Procedure**, which encompasses a complete surgical procedure and is composed of a set of surgical phases and transitions between them; (2) **Phase**, which defines a complex of surgical actions with a defined intention or objective having a clear beginning and end; (3) **Action**, which defines a simple tool task with a specific objective and defined as a set of surgemes and their interactions, i.e. the atomic actions within a surgery that cannot be decomposed further. The medical knowledge used to define the HFSM is provided by surgeon interviews and literature review. The control and supervision of the procedure, phases and actions is conducted by means of finite state machines (FSM). This methodology enables a strict control of the status of each phase, action and surgeme executed by each robotic tool during a procedure. The procedure is modelled as a FSM, where each state represents a phase. Following with the decomposition, each phase FSM contains an action FSM whose states are surgemes representing atomic actions performed by a robot or a surgical tool.

The supervisory controller is in charge of providing commands to the robot controller and to trigger the transitions of the HFSM. Three sub-modules compose the controller: *Supervisor* uses the knowledge of the surgical procedure to choose the atomic movement (i.e. surgeme); *Observer* converts the information generated by the perception of the environment to the trigger events encoded into the surgical procedure; and *Dispatcher* is in charge of dispatching the surgeme execution to the lower level robot controller (i.e. the trajectory reconfiguration and obstacle avoidance modules).

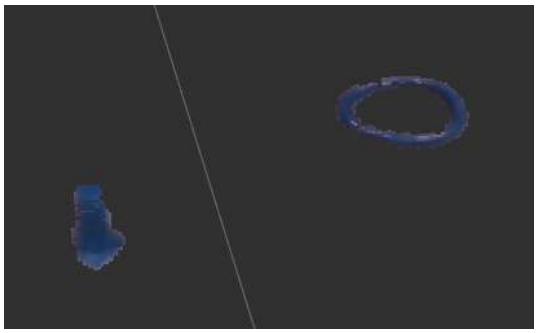


Fig. 4.3: Example of the segmentation of the blue peg and ring.

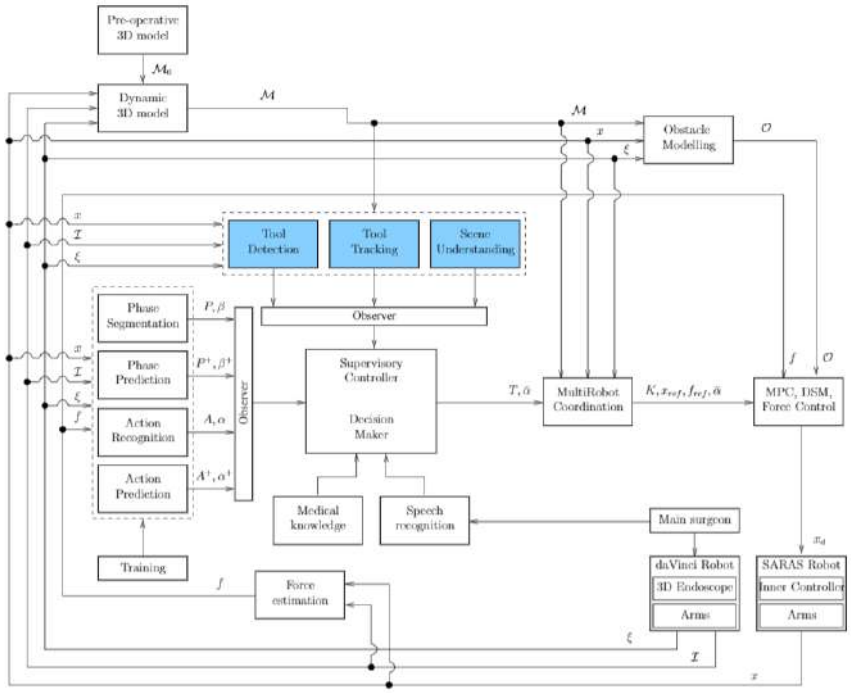


Fig. 4.4: Cognitive architecture: Perception blocks highlighted in blue.

Label	State description
s_1	Move to safe position
s_2	Wait until catheter is recognised
s_3	Follow the catheter position
s_4	Approach the grasping position
s_5	Idle, wait in the grasping position
s_6	Open the grasper
s_7	Reach the grasping position
s_8	Close the grasper
s_9	Pull up the grasper
s_{10}	Open the grasper

Label	Trigger description
t_1	Robot tool is in safe position
t_2	Catheter has been recognised
t_3	Catheter has been pulled-up
t_4	Robot tool is ready to start grasping
t_5	Robot tool is open
t_6	Catheter is on the grasping position
t_7	Robot tool is close
r_1	Catheter tracking is lost
r_2	Reset command by surgeon
r_3	Target position not reachable

Table 4.1: Description of the surgemes and triggers generated by the Observer.

The experimental setup consists of a da Vinci® surgical robot controlled through the da Vinci® Research Kit (dVRK), a SARAS robotic arm [138] acting as the assistant surgeon.

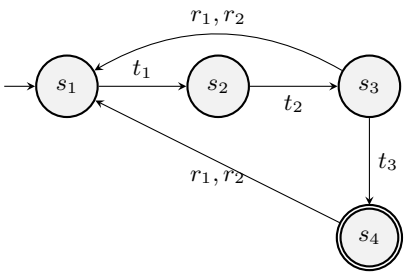


Fig. 4.5: The finite state machine of the first action (approach catheter) of the bladder neck incision. Labels s_i, r_i and t_i are defined in Table 4.1.

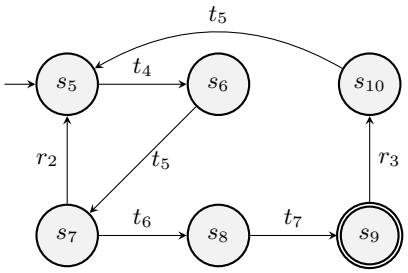


Fig. 4.6: The finite state machine of the second action (grasp catheter) of the bladder neck incision. Labels s_i, r_i and t_i are defined in Table 4.1.

The HFSM used in the experiment is composed of one phase FSM and two action FSMs shown in Fig. 4.5 (*wait until the catheter is detected before moving SARAS arm towards it*) and Fig. 4.6 (*grasping and pulling movements*). A detailed description of the surgemes and the transition triggers are presented in Table 4.1.

4.2.1 Catheter recognition and tracking

Given the registration between the pre-operative data of the patient and the reconstructed map in a common reference frame we can have an idea where the urethra is and consequently the static areas of the anatomy. With this hypothesis we can filter an area of interest in world space and re-project the points p to create a "bounding box" BB in the image acquired by the endoscope and work with a smaller image. The next step is to recognize and to track the catheter using the algorithm described in Algorithm 1.

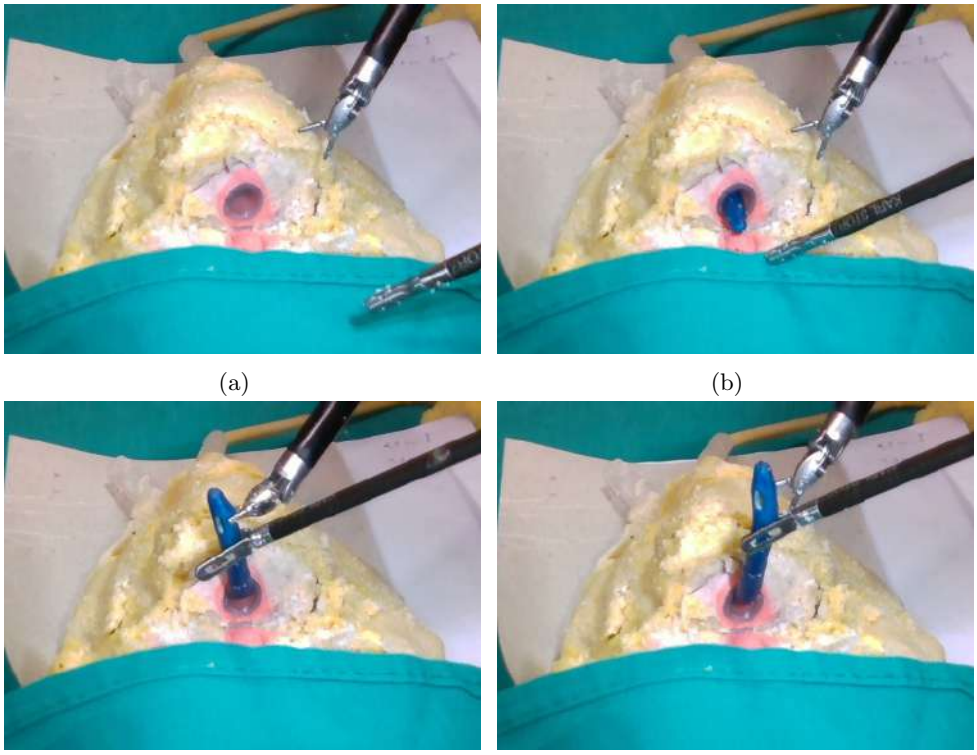
Algorithm 1: Catheter Tracking

```
1  Data: p(n) ∈ BB of the Bounding Box, Tools
2  p'(n) ∈ BB' ← re-projection 3D to 2D
3  for t = 1 to Inf do
4      if Initialisation then
5          Catheter recognition ← Matching
6          Feature extraction ← GoodFeatureToTrack()
7      else
8          Tracking with optical flow ← calcOpticalFlowPyrLK()
9          p'(n,t) Bounding Box update ← re-projection of the tool
10         Initialisation = false
```

To recognize the catheter as soon as it is visible in the image we used a template matching technique. It is a technique for finding areas of an image that match (i.e. similar) to a template image (patch).

Let I and T be

- **Source image (I)**: endoscope image
- **Template image (T)**: catheter.



(a) (b)

(c) (d)

Fig. 4.7: The automatic catheter grasping experimental validation. a) the initial position of the autonomous system, b) the arm starts moving to the catheter, c) the arm approaches the grasping point, d) once the catheter is grasped the main surgeon releases it.

We compare the template image against the source image by sliding it, moving the patch one pixel at a time. At each location, a metric is calculated so it represents how similar the patch is to that particular area of the source image. For each location of T over I, the metric is stored within a matrix R. Each location (x, y) in R contains the match metric. In our experiments we used the following metric

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x', y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x', y+y')^2}} \tag{4.1}$$

Once we have a match higher than a pre-defined threshold, the catheter is found, then good features can be identified and tracked from frame to frame.

To find features we use the method proposed by Shi-Tomasi [173]. The idea is to find the difference in intensity for a displacement of (u, v) in all directions.

$$E(u,v) = \sum_{x,y} \underbrace{w(x,y)}_{\text{window function}} [\underbrace{I(x+u, y+v)}_{\text{shifted intensity}} - \underbrace{I(x,y)}_{\text{intensity}}]^2 \tag{4.2}$$

For corner detection we have to maximize the function $E(u, v)$, that means we have to maximize the second term. Applying Taylor Expansion we get the final equation as:

$$E(u,v) \approx \begin{bmatrix} u\ v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{4.3}$$

where

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x I_x\ I_x I_y \\ I_x I_y\ I_y I_y \end{bmatrix} \tag{4.4}$$

Then we need to create a score to determine if a window contains a corner or not. A scoring function is the Harris Corner Detector given by:

$$R = \det(M) - k(trace(M))^2 \tag{4.5}$$

where $\det(M) = \lambda_1 \lambda_2$, $trace(M) = \lambda_1 + \lambda_2$, with $\lambda_{1,2}$ the two eigenvalues of M. Shi-Tomasi proposed

$$R = \min(\lambda_1, \lambda_2) \tag{4.6}$$

When its value is greater than a threshold value, it is considered a corner.

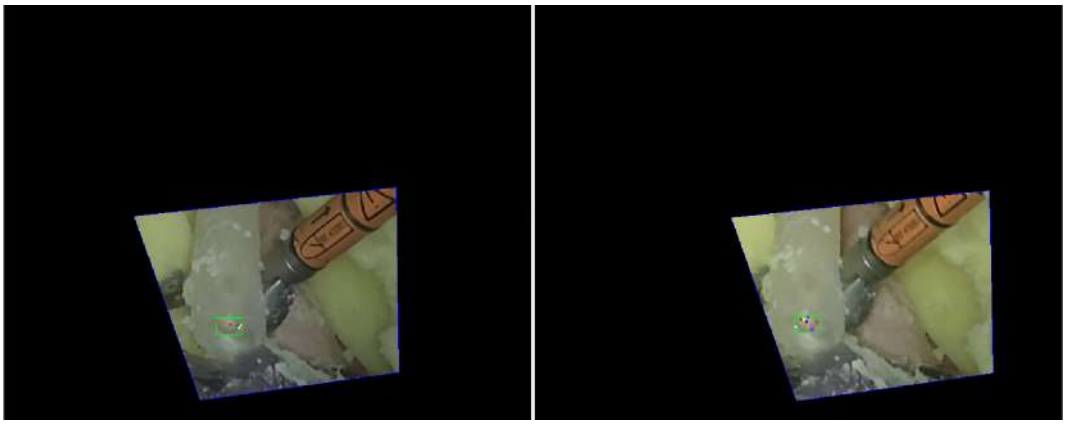


Fig. 4.8: Catheter recognition and tracking in both endoscope images

The extracted features are given to the optical flow function frame by frame to ensure that the same points are being tracked. Optical flow assumes that the pixel intensities of an object do not change between consecutive frames and the neighbouring pixels have similar motion. There are many implementations of sparse optical flow, including the Lucas–Kanade [105] method, Horn–Schunck [73] method, and Buxton–Buxton [14] method. We used the Lucas-Kanade method. Starting from the optical flow equation

$$I(x,y,t) = I(x+dx, y+dy, t+dt) \tag{4.7}$$

by computing the Taylor series approximation of the right-hand side, removing common terms, and dividing by dt we get the following equation

$$f_x u + f_y v + f_t = 0 \tag{4.8}$$

Lucas-Kanade method takes a 3×3 patch around the point. So all the 9 points have the same motion. We can find (f_x, f_y, f_t) for these 9 points. So now our

problem becomes solving 9 equations with two unknown variables which is over-determined. A better solution is obtained with least square fit method. The final solution is

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i {f_{x_i}}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i {f_{y_i}}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix}. \tag{4.9}$$

The tracking procedure starts when the tool grasps the catheter: then the bounding box is updated continuously following the instrument within the video stream. The position and the velocity of the catheter are projected in 3D (using the depth map of the RGBD camera) providing to the supervisory controller the odometry of the grasping point.

The proposed method, as shown in Fig. 4.7, is able to accomplish the catheter detection and grasping autonomously. Fig. 4.7a shows the robot in the initial position, corresponding to the state s_1 of the FSM of Fig. 4.5. When the vision module recognises the catheter, the SARAS tool reaches the approach grasping position (s_4) provided by the features in the 3D system space, as shown in Fig. 4.7b. The surgeon extracts the catheter with the da Vinci® arm (PSM) and pulls it up to the desired position, triggering the transition from state s_5 and s_6. Therefore, the SARAS tool starts moving towards the grasping point and can proceed with the grasping action shown in Fig. 4.7c. After that, the SARAS arm pulls the catheter autonomously as shown in Fig. 4.7d.

4.3 Instrument Tracking

The real-time knowledge of the pose of surgical tools with respect to the endoscope and to the underlying anatomy is of paramount importance in computer-assisted systems. Different approaches for instrumental localisation have been investigated including electro-magnetic (EM) [53, 92] and optical tracking [43], robot kinematics [159] and image-based tracking in endoscopic images, ultrasound (US) [76] and fluoroscopy [207]. Image-based approaches are highly attractive because they do not require modification to the instrument design or the operating theatre and they can provide positional and motion information directly within the coordinate frame of the images used by the surgeon to operate. A major challenge for image-based techniques is robustness and in particular to the diverse range of surgical specialisations and conditions that may affect image quality and visibility.

In computer vision, the detection of any object can be described quite generally as a parameter estimation problem over a set of image features. In general there are three strategies that have been used to solve the problem. The first two fit within a more holistic modelling paradigm and are separated into discriminative methods using discrete classification and generative methods which aim to regress the desired parameters in a continuous space. The third strategy encompasses ad-hoc methods that rely on empirical combinations of simple models for detection. The first step to perform an object detection consists on compressing the image data into a manageable, low dimensional

representation in the form of features. The second step is to add prior knowledge and details to increase the detection results. The strategies we evaluated for the detection of surgical instruments are based on five discriminative selections, namely *Feature Representation*, *Color*, *Gradient*, *Texture*, and *Shape*. In addition to these, the system relies also on the real-time semantic segmentation of the anatomical structures in the scene, which provides the location of the deformable organs along with a less refined segmentation of the robots' tools. The combined semantic and discriminative tool detection systems generate very accurate localization and characterization of them, which, in turns, unlocks the adaptation of advanced trajectory planning algorithms.

4.3.1 Feature representation

Features computed over the input images and aggregated into specific representations serve as a basis for object-specific model learning and classification. The selection of sufficiently distinguishable natural features is a challenging aspect for any detection system. One of the most common approaches is to combine different features to provide a potentially more discriminative feature space; unfortunately, this requires more computational power and increases the size of the required training set. In Pezzementi et al. [145], the authors relied on one of the most popular existing strategies: the Linear Discriminant Analysis (LDA) [112].

4.3.2 Color

The most common and widespread of the natural features is color. Nearly all of the existing methods for detecting surgical instruments in images use color information as the primary or sole visual aid due to its ease of computation and simplicity. Nevertheless, it is challenging when we need to cope with visual ambiguities created by shadows and by different lighting conditions. The RGB colorspace was initially investigated for tool detection as part of the framework developed by Lee at al [96] in a MIS surgical context. It has been directly used in [111, 157, 188, 226]. RGB is an additive color model. It means that different proportions of red, blue and green light can be used to produce any color. The RGB color model was created specifically for display purposes and often has been supplanted by the HSV or HSL colorspace, which are more representative [39, 182]. These colorspaces offer a separation between the chromaticity and the luminance component. By decoupling luminosity from other components, more robust results can be obtained against lighting changes. The CIELab color space, which is closely modelled on human visual perception, allows a wider range of possible colors than RGB, but each channel is described by more than 8 bits, which means longer process time [3]. While being relatively simple to compute, color features have significant shortcomings. Despite the obvious dissimilarity between the red hues of tissue and the monochromaticity of instruments, the lighting used in medical environments combined with the smooth tissue surface causes large specular reflections disrupting the white and grey appearance of metallic instruments. This leads to particular challenges when classifying the instruments using color alone.

4.3.3 Gradient

Gradients is the second most popular feature. Typically, gradients are generated from color image for example from intensity values or specific colorspace component (e.g. Saturation). Gradient features can be extracted either through the computation of image derivatives along $x-$ and $y-$axis [209] or by performing Sobel filtering [64]. However, such information are not used alone but they are often the inputs to other more sophisticated functions (e.g. Hough transform). A more robust representation of gradients is the Histograms of Oriented Gradients (HOG) [27]. Typically, not all the orientations, or oriented gradients, are represented but rather a discrete number corresponding to the amount of bins of the histogram [11]. Variants of the HOG framework have been preferred in other studies through the use of edges and dominant orientations [159, 187]. In general, those feature representations are useful for describing the oriented edges and corners but suffer heavily from noise which is common in medical images.

4.3.4 Texture

More robust representations of gradient features can be achieved by extracting texture information which can be defined as periodically repeated local patterns in an image. Originally, texture features have been extracted using filter responses for example through Gabor filtering, via textons [108] or Local Binary Patterns (LBP) [136]. A popular strategy for object detection lies in *interest points* detection since the emergence of the highly successful SIFT features [104], which has spawned numerous other attemps [4, 7, 27]. All of them are based on the principle that creating histograms of gradient orientation around a particular keypoint allows it to be correctly matched when viewed from a different viewpoint. Despite the popularity of these methods in other areas of computer vision, they have not been used extensively in the task of surgical instrument detection. One particularly successful attempt has been made by Reiter et al. [158], making use of SIFT features learned around the tip of da Vinci robotic instruments.

4.3.5 Shape

Amongst the least represented categories, surgical tool detectors can utilize shape features, generally represented as a set of numbers produced to describe a given shape. Different approach types can be followed such as region-based, space-domain, and transform-domain shape features [215]. Region moments, for instance Hu invariant moments, are very popular amongst region-based shape features. In Voros et al. [198] authors relied on the Otsu's thresholding technique to identify the tool-tip location by finding the optimal separation between instrument and background pixels by computing zeroth-, first-, and second-order cumulative moments. Region moments are mathematically formulated to offer invariance under translation, scale, and rotation for an average computational complexity. However, they provide a very limited robustness towards noise, occlusion or non-rigid deformation, for a highly redundant

information extracted. Within transform-domain shape features, Fourier descriptors have been previously used in Doignon et al. [38] to enable better classification of regions as instrument or background. As their color-based segmentation methods produce several outliers, they are forced to incorporate the shape of the region as part of their evaluation. Fourier descriptors describe the boundary of a region by computing a Fourier component for each pixel in the boundary. Exploiting the properties of the Fourier transform, this descriptor can be shown to be invariant to rotation, translation, scaling and origin. By extracting the outer contour of a region detected by a color classifier and taking the Euclidean distance between the region's Fourier descriptors, the most similar shape in the image is taken as the one with the minimal distance. The authors also combine the Fourier descriptors with affine invariant region moments to improve the robustness of their region detection, again using the Euclidean distance between the moments.

4.3.6 Proposed method

The proposed method is based on a discriminant color feature with robustness capabilities with respect to intensity variations and specularities. It uses color and feature approaches to fit the instrument model. The RGB space is the most well-known color representation since it is useful for data storage. However, some color image processings such as enhancement and restoration require that only the luminance component (as the amount of visible light) to be processed whereas some other applications require color (hue and saturation) components to be preserved or modified. It is known that the human eye can detect only in the neighborhood of one or two dozen intensity levels at any point in a complex image due to brightness adaptation, but it can differentiate thousands of color shades and intensities. The color saturation seems to be a discriminant attribute for gray regions segmentation since it is a measure of the amount of white within the color despite that it maybe affected by surface reflectance . A low saturation value indicates a low colored pixel and a high value corresponds to a purely colored pixel. Coordinate systems related to the psychological perceptual attributes (Hue, Saturation and Intensity—HSI for short) are more relevant for analyzing colors distribution in the image than RGB since the chromaticity plane (H and S) is perpendicular to the intensity axis and furthermore, RGB space brings a non-uniform chromaticity scale. The standard saturation S related with RGB is obtained from the following equation:

$$S = 1 - 3\frac{\min(R, G, B)}{R + G + B}. \tag{4.10}$$

This definition clearly shows that pixels may have the same saturation whatever their brightness or color hue values are (excepted for the singularity located at $(R = G = B = 0)$. With the objective of improving the detection of "dark" regions in the image, we slightly modify the above definition of saturation S' as follows:

$$S' = 1 - \frac{\min(R, G, B)}{\max(R + G + B)}. \tag{4.11}$$

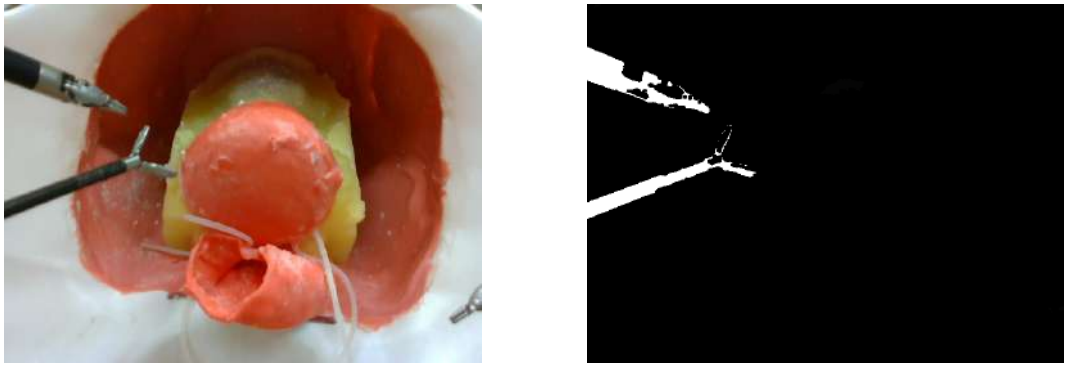
This can be thought as a color purity stretching with a simple non-linear point operation. Compared to the saturation attribute defined in (4.10), this new color purity attribute is a little bit more sensitive to brightness changes but mainly for chromatic pixels. Since S' rather affects more high values than low values, it tends to separate more chromatic pixels from achromatic ones. With this new definition of saturation it is easier to apply a simple color segmentation on the image to achieve good results. Once we get the color mask, we can filter out the noise. Noise filtering is commonly used as one of the first operations applied to digitized images. Non-linear filtering allows to detect lack of spatial coherence and either replace an inconsistent pixel value by using some or all pixels in a neighborhood. Such low-level processing is crucial to avoid over-segmentation results which are very awkward for pixels classification. Some of non-linear filters have the capabilities to smooth intensity values of pixels in a given region and to equally preserve the topological properties of edges. The homogeneity plays a significant role in separating the objects from each other, usually in separating the region of interest from the background and a very attractive color segmentation based on homogeneity histogram. Once we have a first instrument segmentation mask, we refine the output through morphology. Morphological transformations are some simple operations based on the image shape. It is normally performed on binary images. It needs two inputs, the first one is the original image, the second one is called structuring element or kernel which decides the nature of operation. Two basic morphological operators are erosion and dilation. The operations we used are called *Opening* and *Closing*. The opening is the erosion followed by dilation: it is useful for removing noise. The closing operation is the dilation followed by erosion which is useful in closing small holes inside the foreground objects, or small black points on the object.

To increase the detection and perform the tracking of the instruments during the procedure, we run a feature extractor in the image and select only the features that belong to the segmented mask in the previous step. The extracted features are given to the optical flow function frame by frame to ensure that the same points are being tracked. Optical flow assumes that the pixel intensities of an object do not change between consecutive frames and the neighbouring pixels have similar motion.

Once we have the instrument mask on the whole procedure (Figure 4.9), we can compute the corresponding 3D point-cloud using the depth value.

4.4 Discussion and Conclusions

This chapter has reviewed and discussed the experiments for validating the accuracy of the calibration and algorithms designed for the development of autonomous applications in surgery. Different scenarios have been evaluated as prerequisites for autonomy and various techniques to manage the anatomical environment. In the next chapters we will combine these acquired skills with a method to be able to reconstruct an anatomical environment and interact with it.



(a) (b)

Fig. 4.9: A frame acquired during one of the phases of the radical prostatectomy procedure: (a) shows the RGB image; (b) shows the segmented mask of the frame

5

Simultaneous localization and mapping

This chapter reviews the state of the art in *Simultaneous Localization and Mapping* algorithms to obtain the 3D structure of an unknown environment. In order to clarify the differences between the various approaches, how they can be adapted in laparoscopy and to motivate the final choice we made.

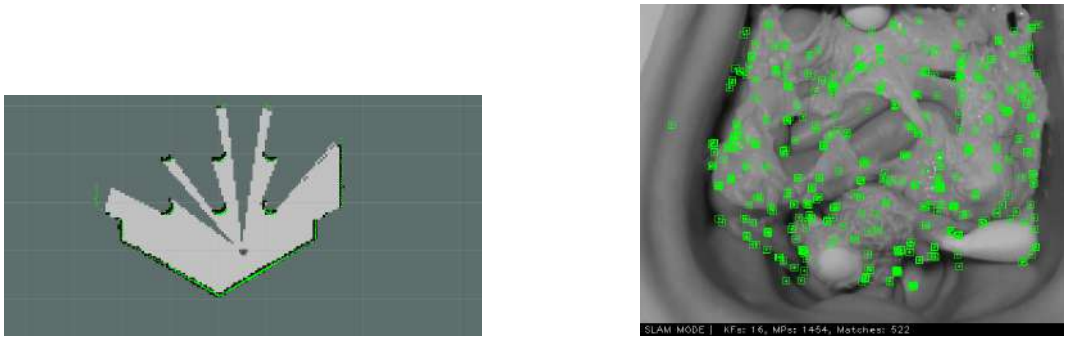
Fig. 5.1: Example of laparoscopic intervention

5.1 3D reconstruction in surgery

Today, numerous diseases are diagnosed or treated using interventional techniques to access the internal anatomy of the patient. While open surgery involves cutting the skin and dividing the underlying tissues to gain direct access to the surgical target, minimally invasive surgery (MIS) is performed through small incisions in order to reduce surgical trauma and morbidity. The term laparoscopic surgery refers to MIS performed in the abdominal or pelvic cavities. The abdomen is usually insufflated with gas to create a working volume (pneumoperitoneum) into which surgical instruments can be inserted via ports, as shown in 5.1. As direct viewing of the surgical target is not possible, an endoscopic camera (laparoscope) generates views of the anatomical structures and

of the surgical instruments. In contrast to open surgical procedures, MIS provides the surgeon with a restricted, smaller view of the surgical field, which can be difficult to navigate for surgeons only trained in open surgery techniques. To compound the visual complexity of MIS, laparoscopic instruments are operated under difficult hand-eye ergonomics and usually provide only four degrees of freedom (DoF) which severely inhibits the dexterity of tissue manipulation.

Minimally Invasive Surgery (MIS) is an indispensable tool in modern surgery for the ability of mitigating postoperative infections, but it also narrows the surgical field of view and makes surgeons receive less information. MIS has introduced significant challenges to surgeons as they are required to perform the procedures in narrow space with elongated tools without direct 3D vision. To solve this problem, 3D laparoscopy is applied to provide two images to create an 'imagined 3D model' for surgeons. Inspired by the fact that stereo vision can generate shapes for qualitative and quantitative purpose, a mosaic of all the 3D shape by taking account deformation will make better use of 3D information. Therefore, it is helpful if a dynamic 3D morphology could be incrementally generated and rendered for the surgeons intra-operatively and for future autonomous surgical robots for implementing surgical operation and navigation [216]. However, the small field of view of the scopes and the deformation of the soft-tissue limit the feasibility of using traditional structure-from-motion and image mosaicking methods. Even worse, rigid and non-rigid movement caused by motion of camera pose, breathing, heartbeat and instrument interaction increase difficulty in soft-tissue reconstruction and visualization.



(a) GMapping (b) ORB-SLAM

Fig. 5.2: Example of SLAM algorithms

Simultaneous Localization and Mapping (SLAM) [42] [191] is a technique used to obtain the 3D structure and for estimating sensor motion of an unknown environment (Figure 5.2). This technique was originally proposed to achieve autonomous control of robots. Then, SLAM-based applications have widely become broadened such as computer vision-based online 3D modeling, augmented reality (AR)-based visualization, and self-driving cars [74] [75]. In early SLAM algorithms, many different types of sensors were integrated such as laser range sensors, rotary encoders, inertial sensors, GPS, and cameras.

Usually, SLAM using cameras is referred to as visual SLAM (vSLAM) because it is based on visual information only. vSLAM can be used as a fundamental technology for various types of applications

SLAM systems have a state $x_t = (c_t, M1, ..., Mn)$ which describes the camera pose $c_t = (t, R)$, consisting of a translation vector t and a rotation matrix R and a set of n 3D landmarks $M_i = (x, y, z)$ which describe the 3D structure of the environment at time t. The live camera images are processed individually to update the state of the system. At each frame, a new camera pose is estimated, existing landmarks are re-observed and new 3D landmarks are added to the state. The computational complexity of SLAM is dictated by the size of the state (i.e. the number of landmarks) and not the number of images (as in Structure from Motion). This formulation of the problem makes it computationally feasible to sequentially estimate the camera pose and 3D structure in real time.

Modeling image noise and uncertainty is a fundamental component of SLAM. Sequentially updating the state with noisy observations of landmarks would lead to error propagation and an inconsistent state. Uncertainty in the state is modeled by a full covariance matrix. The state and covariance matrix are managed and updated using a probabilistic framework where the joint posterior density of the 3D landmarks and the camera pose is described by the probability distribution $P(x_t|Z_0 : t, U_0 : t, x_0)$ given the observations Z_i of visible landmarks and any control inputs U_i from position sensors on the camera (e.g. accelerometer) motion model. Motion models comprise a deterministic and a stochastic element. The deterministic part is a prediction based on a sensor measurement (e.g. Inertial Measurement Unit (IMU)) or on previous history of camera motion. The stochastic part is a probabilistic model of the uncertainty in the predicted motion, which may be derived experimentally. Given the predicted new pose of the camera it is possible to project the 3D landmarks into the image in preparation for the measurement and update steps. The measurement or observation step solves the association problem by establishing correspondence between 3D landmarks and features in the image space. In vision SLAM systems, the 3D landmarks may be associated to an image patch or template. Matching the template in the image provides new measurements of the location of the 3D landmarks relative to the camera. The measurement can be made in the image space for monocular cameras or in 3D for stereo cameras and RGBD sensors. A measurement model is defined which relates the measurement to the state. Finally, the state is updated using the predicted model, measurement model and the observed measurements of the 3D landmarks. A wide variety of solutions to the SLAM problem have been proposed [5].

Therefore, the existing vSLAM algorithms we studied are shown in the table 5.1 and are categorized according to feature-based, direct, and RGB-D camera-based approaches.

Algorithm	Method	Map density	Optimization	Loop Closure
MonoSLAM [29].	Feature	Sparse	no	no
PTAM [83]	Feature	Sparse	yes	no
ORB-SLAM [126]	Feature	Sparse	yes	yes
DTAM [132]	Direct	Dense	no	no
LSD-SLAM [45]	Direct	Semi-dense	yes	yes
RGB-D Slam [44]	RGB-D	Dense	yes	yes
Kinect Fusion [78]	RGB-D	Dense	yes	yes
SLAM++ [167]	RGB-D	Dense	yes	yes

Table 5.1: List of state-of-the-art algorithms for each type of SLAM

5.2 Elements of vSLAM

The framework is mainly composed of three modules as follows.

- Initialization
- Tracking
- Mapping

To start vSLAM, it is necessary to define a certain coordinate system for camera pose estimation and 3D reconstruction in an unknown environment. Therefore, in the initialization, the global coordinate system should first be defined, and a part of the environment is reconstructed as an initial map in the global coordinate system. After the initialization, tracking and mapping are performed to continuously estimate camera poses. In the tracking, the reconstructed map is tracked in the image to estimate the camera pose of the image with respect to the map.

In order to do this, 2D-3D correspondences between the image and the map are first obtained from feature matching or feature tracking in the image. Then, the camera pose is computed from the correspondences by solving the Perspective-n-Point (PnP) [156] [134] problem. It should be noted that most of vSLAM algorithms assumes that intrinsic camera parameters are calibrated beforehand so that they are known. Therefore, a camera pose is normally equivalent to extrinsic camera parameters with translation and rotation of the camera in the global coordinate system. In the mapping, the map is expanded by computing the 3D structure of an environment when the camera observes unknown regions where the mapping is not performed before.

The following two additional modules are also included in vSLAM algorithms according to the purposes of applications.

- Relocalization
- Global map optimization

The relocalization is required when tracking fails due to fast camera motion or disturbances. In this case, it is necessary to compute the camera pose with respect to the map again. Therefore, this process is called “relocalization.” If the relocalization is not incorporated into vSLAM systems, the systems do not work anymore after the tracking is lost and such systems are not practically useful. Therefore, a fast and efficient method for the relocalization has been discussed in the literature. Note that this is also referred to as kidnapped

robot problems in robotics. The other module is global map optimization [88]. The map generally includes accumulative estimation error according to the distance of camera movement. In order to suppress the error, the global map optimization is normally performed. In this process, the map is refined by considering the consistency of whole map information. When a map is revisited such that a starting region is captured again after some camera movement, reference information that represents the accumulative error from the beginning to the present can be computed. Then, a loop constraint from the reference information is used as a constraint to suppress the error in the global optimization. Loop closing is a technique to acquire the reference information. In the loop closing, a closed loop is first searched by matching a current image with previously acquired images. If the loop is detected, it means that the camera captures one of previously observed views. In this case, the accumulative error occurred during camera movement can be estimated. Note that the closed-loop detection procedure can be done by using the same techniques as re-localization. Basically, re-localization is done for recovering a camera pose and loop detection is done for obtaining geometrically consistent map. Pose-graph optimization has widely been used to suppress the accumulated error by optimizing camera poses. In this method, the relationship between camera poses is represented as a graph and the consistent graph is built to suppress the error in the optimization. Bundle adjustment (BA) [195] is also used to minimize the re-projection error of the map by optimizing both the map and the camera poses. In large environments, this optimization procedure is employed to minimize estimation errors efficiently. In small environments, BA may be performed without loop closing because the accumulated error is small.

5.2.1 Feature-based methods

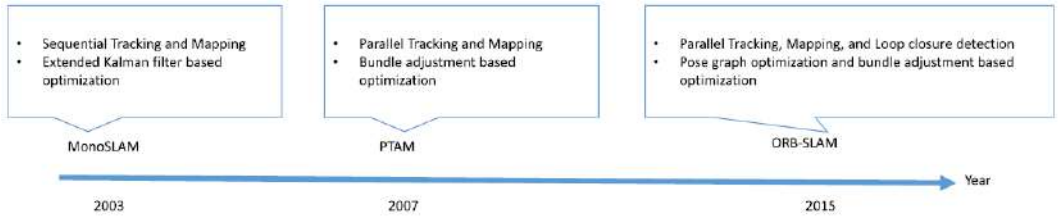


Fig. 5.3: Timeline of feature based methods

There exist two types of feature-based methods in the literature: filter-based and BA-based methods.
First monocular vSLAM was developed in 2003 by Davison et al. They named it MonoSLAM [29]. MonoSLAM is considered as a representative method in filter-based vSLAM algorithms. In MonoSLAM, camera motion and 3D structure of an unknown environment are simultaneously estimated using an extended Kalman filter (EKF) [160]. 6 Degree of freedom (DoF) camera motion and 3D positions of feature points are represented as a state vector in EKF. Uniform motion is assumed in a prediction model, and a result of feature point

tracking is used as observation. Depending on camera movement, new feature points are added to the state vector. Note that the initial map is created by observing a known object where a global coordinate system is defined. In summary, MonoSLAM is composed of the following components.

- Map initialization is done by using a known object.
- Camera motion and 3D positions of feature points are estimated using EKF.

The problem of this method is a computational cost that increases in proportion to the size of an environment. In large environments, the size of a state vector becomes large because the number of feature points is large. In this case, it is difficult to achieve real-time computation.

To solve the problem of reducing the computational cost in MonoSLAM, PTAM [83] split the tracking and the mapping into different threads on CPU. These two threads are executed in parallel so that the computational cost of the mapping does not affect the tracking. As a result, BA [195] that needs a computational cost in the optimization can be used in the mapping. This means that the tracking estimates camera motion in real-time, and the mapping estimates accurate 3D positions of feature points with a lower computational cost. PTAM is the first method which incorporates BA into the real-time vSLAM algorithms. After publishing PTAM, most vSLAM algorithms follow this type of multi-threading approaches. In PTAM, the initial map is reconstructed using the five-point algorithm. In the tracking, mapped points are projected onto an image to make 2D–3D correspondences using texture matching. From the correspondences, camera poses can be computed. In the mapping, 3D positions of new feature points are computed using triangulation at certain frames called keyframes. One of the significant contributions of PTAM is to introduce this keyframe-based mapping in vSLAM. An input frame is selected as a keyframe when a large disparity between an input frame and one of the keyframes is measured. A large disparity is basically required for accurate triangulation. In contrast to MonoSLAM, 3D points of feature points are optimized using lobal BA with some keyframes and global BA with all keyframes with the map. Also, in the tracking process, the newer vision of PTAM employs a relocalization algorithm. It uses a randomized tree-based feature classifier for searching the nearest keyframe of an input frame. In summary, PTAM is composed of the following four components.

- Map initialization is done by the five-point algorithm [133].
- Camera poses are estimated from matched feature points between map points and the input image.
- 3D positions of feature points are estimated by triangulation, and estimated 3D positions are optimized by BA.
- The tracking process is recovered by a randomized tree-based searching.

Compared to MonoSLAM, in PTAM, the system can handle thousands of feature points by splitting the tracking and the mapping into different threads on CPU. There have been proposed many extended PTAM algorithms. Castle et al. developed a multiple map version of PTAM. Klein et al. developed a mobile phone version of PTAM. In order to run PTAM on mobile phones,

input image resolution, map points, and number of keyframes are reduced. In addition, they consider rolling shutter distortion in BA to get an accurate estimation result because a rolling shutter is normally installed in most mobile phone cameras due to its low cost. Since PTAM can reconstruct a sparse 3D structure of the environment only, the third thread can be used to reconstruct a dense 3D structure of the environment.

Geometric consistency of the whole map is maintained by using BA for the keyframes as explained above. However, in general, BA suffers from a local minimum problem due to the large number of parameters including camera poses of the keyframes and points in the map. Pose-graph optimization is a solution to avoid the problem in the loop closing, camera poses are first optimized using the loop constraint. After optimizing the camera poses, BA is performed to optimize both 3D positions of feature points and the camera poses. For the loop closing, a visual information-based approach is employed. They used a bag-of-words-based image retrieval technique to detect one of the keyframes which view is similar with the current view. In a vSLAM system, a stereo camera is selected as a vision sensor. In this case, the scale of the coordinate system is fixed and known. However, in monocular vSLAM cases, there is a scale ambiguity and a scale may change during camera movement if global BA is not performed. In this case, a scale drift problem occurs and the scale of the coordinate system at each frame may not be consistent. In order to correct the scale drift, camera poses should be optimized in 7 DoF. Strasdat et al. proposed a method for optimizing 7 DoF camera poses based on similarity transformation. As an extension of PTAM, ORB-SLAM [126] includes BA, vision-based closed-loop detection, and 7 DoF pose- graph optimization. As far as we know, ORB-SLAM is the most complete feature-based monocular vSLAM system. ORB-SLAM is also extended to stereo vSLAM and RGB-D vSLAM.

5.2.2 Direct methods

In contrast to feature-based methods described in the previous section, direct methods use an input image without any abstraction derived from handcrafted feature detectors and descriptors. They are also called feature-less approaches. In general, photometric consistency is used as an error measurement in direct methods whereas geometric consistency such as positions of feature points in an image is used in feature-based methods. LSD-SLAM [45] is a leading method in direct methods. The core idea of LSD-SLAM follows the idea from semi- dense Visual Odometry (VO). In this method, reconstruction targets are limited to areas which have intensity gradient compared to DTAM [132] which reconstructs full areas. This means that it ignores textureless areas because it is difficult to estimate accurate depth information from images. In the mapping, random values are first set as initial depth values for each pixels, and then, these values are optimized based on photometric consistency. Since this method does not consider the geometric consistency of the whole map, this method is called visual odometry. In 2014, semi-dense VO was extended to LSD- SLAM. In LSD-SLAM, loop-closure detection and 7 DoF pose-graph optimization are

added to the semi-dense visual odometry algorithm. In summary, LSD-SLAM is composed of the following four components.

- Random values are set as an initial depth value for each pixel.
- Camera motion is estimated by synthetic view generation from the reconstructed map.
- Reconstructed areas are limited to high-intensity gradient areas.
- 7 DoF pose-graph optimization is employed to obtain geometrically consistent map.

Basically, these semi-dense approaches can achieve real-time processing with CPU. In addition, they optimized the LSD-SLAM algorithm for mobile phones by considering the CPU architecture for them. In the literature, is also evaluated the accuracy of the LSD-SLAM algorithm for low-resolution input images and is also extended to stereo cameras and omni-directional cameras

5.2.3 RGB-D methods

Recently, structured light-based RGB-D cameras such as Microsoft Kinect or Intel RealSense become economic and small. Since such cameras provide 3D information in real-time, these cameras are also used in vSLAM algorithms. By using RGB-D cameras, 3D structure of the environment with its texture information can be obtained directly. In addition, in contrast to monocular vSLAM algorithms, the scale of the coordinate system is known because 3D structure can be acquired in the metric space. The basic framework of depth (D)-based vSLAM is as follows. An iterative closest point (ICP) algorithm have widely been used to estimate camera motion. Then, the 3D structure of the environment is reconstructed by combining multiple depth maps. In order to incorporate RGB into depth-based vSLAM, many approaches had been proposed as explained below. It should be noted that most of consumer depth cameras are developed for indoor usages. They project IR patterns into an environment to measure the depth information. It is difficult to detect emitted IR patterns in outdoor environments. In addition, there is a limitation of a range of depth measurement such that the RGB-D sensors can capture the environment.

Salas-Moreno et al. [167] proposed an object level RGB-D vSLAM algorithm. In this method, several 3D objects are registered into the database in advance, and these objects are recognized in an online process. By recognizing 3D objects, the estimated map is refined, and 3D points are replaced by 3D objects to reduce the amount of data. Similar algorithm, is proposed in Tateno et al. consisting of a real- time segmentation method for RGB-D SLAM [44]. Segmented objects are labeled, and then, these objects can be used as recognition targets.

5.3 Application to laparoscopy

In MIS, SLAM approaches can be used to localize the pose of the endoscopic camera and build a 3D model of the tissue surface in vivo while the endoscope is

navigated by the surgeon. The in vivo organ model can be used for registration to a pre-operative model. A fundamental component of Augmented Reality (AR) or image guidance is knowing the camera's pose relative to the object or organ of interest. Real-time SLAM provides two fundamental components of computer-assisted surgery (CAS): 3D in vivo tissue model and camera pose estimation while allowing camera movement.

Burschka et al.[13] proposed using an approach called V-GPS to create long-term SLAM-style maps/reconstructions for sinus surgery using a monocular endoscope. A method is proposed for estimating the scale of the 3D reconstruction which cannot be recovered from a monocular camera. The scaled 3D reconstruction of the rigid sinus is registered to a pre-operative CT to enable AR overlay of critical subsurface anatomy. The system was reported to run at 10 Hz with sub-millimeter registration accuracy on phantom data.

An EKF SLAM approach was proposed[123] to build sparse 3D reconstructions of the abdomen and recover the motion of a stereo laparoscope. With the addition of an image pre-processing step, the system was used with low resolution stereo fiber image guides (10,000 fibers)[135] and demonstrated reconstruction accuracy of less than 3 mm of error on phantom data. Monocular EKF SLAM has also been proposed for MIS[61], combining randomized list relocalization with RANSAC outlier removal for recovering from tracking failure. The system reports run times of around 12 Hz. It increases the number of actively tracked landmarks, creating a denser reconstruction which can be used for relocalization. In EFK SLAM the reconstructed surface of the tissue is represented by the set of 3D landmarks. These landmarks can be meshed and textured with images from the endoscope to create visually more realistic tissue models [122, 193]. Such models are an approximation of the organ's surface and may contain inaccuracies. Combining sparse SLAM with dense stereo techniques [194] creates more comprehensive 3D reconstructions without increasing the computational complexity of SLAM. The models discussed so far are based on the assumption that the physical world is static. In anatomical environments such as the nasal passage this assumption holds, however, in the abdomen, respiration causes tissue motion. In [124] dynamic mapping is proposed where the tissue model deforms with periodic motion caused by respiration. The error in the estimated camera position was less than 2 mm for ex vivo data and the system demonstrated accurate recovery of respiration models. Tougher evaluation of SLAM systems for MIS remains a challenge for the community. Optical tracking systems have been used to obtain ground truth for camera motion, however these are still subject to errors from tracking, camera calibration and hand-eye calibration [196]. Validation of the 3D reconstruction can use CT/MRI phantom or ex vivo data for rigid environments and synthetic data for non-rigid environments. No solutions have been proposed for validation of in vivo non-rigid tissue. The SLAM systems described above are sequential and capable of running in real time at up to 25 Hz, however the increased complexity of non-rigid modeling, dense surface reconstruction and recovery from failure introduce additional computational burdens.

5.4 Discussion and Conclusions

SLAM is a mature technology and its use in MIS is attractive due to its real-time capabilities and integration with existing laparoscopic imaging equipment. The feasibility of SLAM has been demonstrated for the MIS environment but there remains a number of theoretical and practical research challenges in transferring this technology to the operating room. A fundamental assumption in SLAM is the rigid environment. Although this holds for some anatomy, fully non-rigid tissue motion is regularly observed in cardiac and abdominal soft-tissue surgery. A theoretical framework must be established for dealing with deformation caused by respiration, cardiac motion, organ shift and tissue tool interaction. Periodic biological signals (respiration, cardiac motion) have been well modeled in the medical imaging community and such models can be incorporated into SLAM [123] However, complex tissue tool interaction and organ shift are likely to require complex biomechanical modeling. Tissue cutting and removal is an additional complication which remains an open research question. SLAM's real-time capabilities rely on establishing a set of 3D landmarks which can be repeatably matched in the image over long periods of time. Correct matching directly affects robustness and reconstruction accuracy. In well illuminated, well textured MIS environments SLAM has been shown to work well.

The MIS environment can be challenging and procedure-long tracking is challenging due to repetitive textures, large changes in lighting conditions, specular reflections and deformation. Partial occlusion due to tools, blood and smoke can generally be dealt with by using outlier removal. Tissue surfaces without texture or detectable features will require additional information from alternative approaches such as structured light or shape from shading (SfS) algorithms.

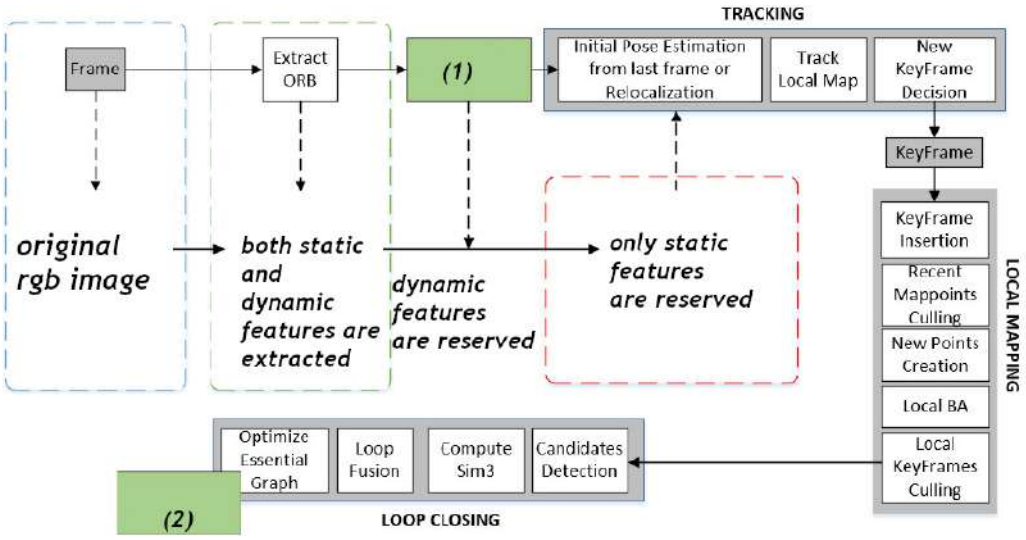


Fig. 5.4: ORB-SLAM system overview, showing all the steps performed by the tracking,local mapping and loop closing threads.

The main aim of this thesis, was to develop a SLAM algorithm capable of handling the deformable nature of the environment by modifying the standard algorithm of ORB-SLAM2 [127] which outperforms many SLAM systems such as Mono-SLAM, PTAM and LSD-SLAM, for the task of monocular endoscopic camera tracking and mapping. ORB-SLAM2 combines many state-of-the-art techniques into one SLAM system, such as using an ORB descriptor for tracking, local keyframe for mapping, graph-based optimization, the Bag of Words algorithm for re-localization, and an essential graph for loop closure. Real-time performance is crucial in time-demanding medical interventions. Since ORB is a binary feature point descriptor, it is an order of magnitude faster than SURF and more than two orders faster than SIFT with better accuracy. In addition, ORB features are invariant to rotation, illumination and scale, which means that it is capable of dealing with some of the main challenges in MIS scenes including rapid movements of endoscope cameras (rotation and zooming) and the change of brightness.

A common problem for monocular scene analysis using SLAM is the initialization, a step required for generating an initial map, because the depth cannot be recovered from a single image frame. An automatic approach,based on the estimation of the scaling factor using the inverse of the mean depth of the scene, is used in ORB–SLAM to calculate homography for planar scenes and a fundamental matrix for non-planar scenes dynamically. This approach can greatly increase the success rate of initialization and reduce the time required for the initialization step. It also facilitates the initialization on an organ surface or to compute a fundamental matrix when the endoscopic camera is pointing at complex structures.

Starting from the ORB-SLAM2 architecture, we modified it to make it suitable for the Robotic Minimally Invasive Surgery (R-MIS). In the following chapters we face the problems of classic SLAM trying to adapt this technology in an anatomical environment. First of all by increasing the performance of the initialization phase and then starting from the diagnostics and the study of the preoperative images with the registration to the map obtained from the SLAM. Finally to study how the dynamic features of the tissues evolve.

Figure 5.4 shows the ORB-SLAM system overview, showing all the steps performed by the tracking, local mapping and loop closing threads. The areas where we modify the SLAM algorithm are highlighted in the green box in the figure. After acquiring the image, the feature extraction section takes place, and the box with label 1 represents the methods where both dynamic and static features are extracted. Then, we need to find an efficient way to separate the two types of feature to be able to do an accurate reconstruction using only the static one. Once, we obtain the map we can work on how to optimize the reconstruction in the box labelled with 2, developing filters to remove those points that represent outliers caused for example by breathing or other involuntary movements of the human body.

6

Rigid 3D Registration of Pre-operative Information for Semi-Autonomous Surgery

In general, a surgical intervention is planned using pre-operative information about the patient and then the surgeons use this knowledge to decide which actions to take. The pre-operative assessment is an opportunity to identify co-morbidities that may lead to patient complications during the anaesthetic, surgical, or post-operative period. Patients scheduled for elective procedures will generally attend a pre-operative assessment $2-4$ weeks before the date of their surgery.

In autonomous systems, the mapping of this planning to the patient anatomy is not trivial since it has to be updated in real-time during the intervention, based on the actual patient condition. This scenario is further complicated by the fact that the environment is soft, which means that it could be subjected to deformation and some anatomical structures can be removed during surgery. Perception of the current surgical environment during robotic minimally invasive surgery (R-MIS) is usually performed with the stereo endoscope that provides visual feedback to the main surgeon. Although it is able to produce a 3D dense reconstruction of the environment, the main drawback of a stereo vision system is that the endoscope has to be sufficiently close to the surface of interest in order to provide a reliable reconstruction. Monocular vision systems, on the other hand, cannot reconstruct a dense point cloud in real-time, but exploiting the multi-view approach, can provide sparse 3D reconstruction. In order to apply the multi-view approach the camera must be moved over time and the scene is in general assumed to be static.

In the previous chapter we presented the Simultaneous localization and mapping (SLAM) algorithm and, as in recent years, there have been many efforts done to evaluate the feasibility of applying it in laparoscopy to reconstruct a sparse or even dense soft-tissue surface [61, 123, 193, 194]. A very popular approach for SLAM in minimally invasive surgery (MIS) relies on oriented FAST and rotated BRIEF (ORB) features [126]. Several works have proved that this approach can successfully support the process of endoscope localisation by providing the poses which are necessary to create a quasi-dense map of the environment [180], [107]. When a monocular vision system is used, there is an additional challenge based on the scale factor estimation. In [13], for example, they retrieve the scale factor manually through the computed tomography (CT) scan and apply it to the 3D reconstruction before the registration.

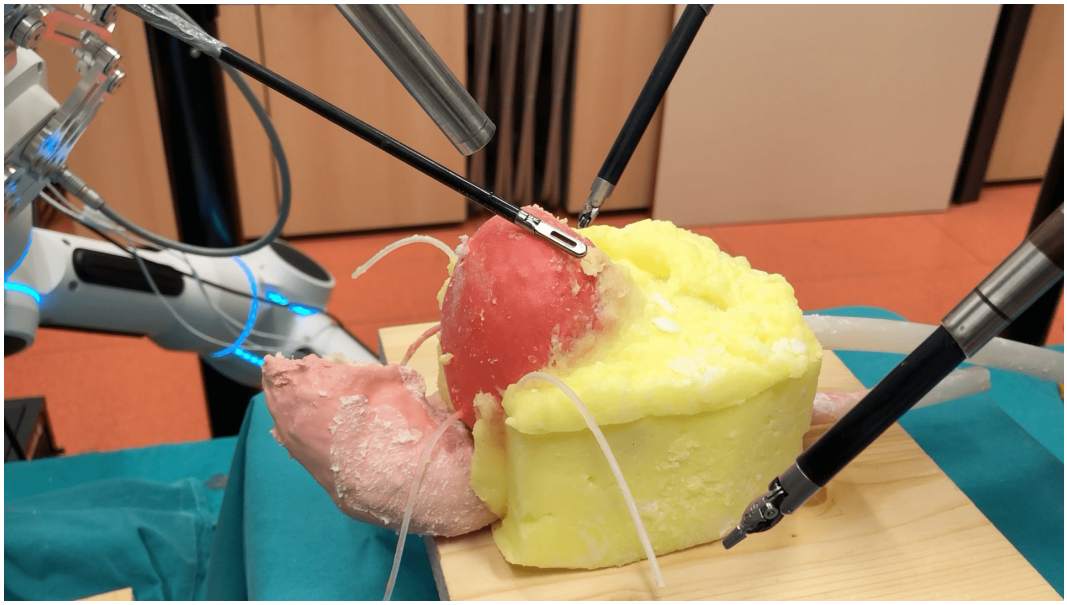


Fig. 6.1: The da Vinci robotic tools, the SARAS robotic tools and the phantom used during the experimental validation.

Once the partial intra-operatively reconstruction is available, its correct registration with the pre-operative information is a fundamental capability to enable the robot to perform tasks like object detection and recognition, navigation and 3D dense map reconstruction. In the context of autonomous robotic surgery, this represents an essential step to plan the robot motion. For example, in [10, 129] a CT scan is used to select the ablation points and the extracted model is then registered to the phantom using embedded spherical landmarks. However, from a practical point of view, the use of landmarks is rarely feasible within a realistic surgical environment. The most popular approach that allows to rigidly align models when no information about landmarks or correspondences is available is the Iterative Closest Point (ICP). The main drawback of this method is that the standard implementation does not always guarantee to find the globally optimal transformation because it can be easily trapped in local minima [9]. Several advanced implementations of ICP have been proposed to tackle this issue, which either rely on the extraction of robust features [214] or exploit more efficient ways to search the 3D space [95, 166].

In this work, we propose a SLAM-based rigid 3D registration method for a semi-autonomous surgical robotic system. The rigid registration allows to provide target points and the volumes of interest, assuming that the environment is not subject to deformations. These volumes represent the bounding regions where the robot has to look to detect the desired target points. In our workflow, we register the anatomical model extracted from Magnetic Resonance Imaging (MRI) to its sparse 3D reconstruction obtained by ORB-SLAM. Such registration is then refined using ICP, which allows to obtain a common reference frame. The main contributions of the work
are the following:

- a scale factor estimation using the da Vinci® kinematics for monocular SLAM,
- an accurate registration between the reconstructed sparse point cloud and a pre-operative model using point-to-plane ICP,
- the integration of the registered 3D anatomical model within the robotic path planner for executing semi-autonomous surgery.

6.1 Method

In a semi-autonomous intervention the robotic system assists the main surgeon when performing specific tasks, and it is triggered by an action recognition system and/or directly by the surgeon. The proposed method aims to register a pre-operative 3D model with the real anatomical environment in order to drive the robotic arm holding a laparoscopic tool towards the target points. The regions of interest are defined on a pre-operative model of the anatomical structures, extracted from pre-operative images (CT or MRI).

Since the registration is rigid and since the pre-operative model is not deformed during the procedure we extract only anatomical information which can be assumed to be fixed over time or subjected to small deformations. The surface of the 3D model is registered with the sparse 3D reconstruction of the environment obtained by SLAM exploiting 3D feature matching and refined using ICP. Once the registration is available the precomputed regions are transformed into the reference frame of the robot and can be used as a set of way-points during the surgery. The overall architecture is depicted in Figure 6.2.

6.1.1 Experimental setup

Figure 6.1 shows the considered semi-autonomous robotic system. It consists of a da Vinci® robot controlled through the da Vinci® Research Kit (dVRK) and of two more robotic arms which play the role of the assistant surgeon. In order to acquire the intra-operative point cloud of the environment, we use only one channel of the da Vinci® stereo endoscope, thus relying on a monocular vision system. This choice allows us to reconstruct larger areas with respect to those that can be obtained with the da Vinci® stereo camera, whose small baseline (around 5 mm) would require the scope to be placed too close to the surface (approximately 5 cm) in order to achieve a reliable reconstruction.

The most important preliminary step that needs to be performed in order to guarantee reliable results is the accurate camera calibration. In our setup, we can measure the pose of both tool manipulators and the endoscope camera arm. It is then possible to map their poses in a common reference space by reaching several points on a custom calibration board (Figure 3.2b), presented in the chapter 3. The overall software architecture is built on top of ROS Kinetic.

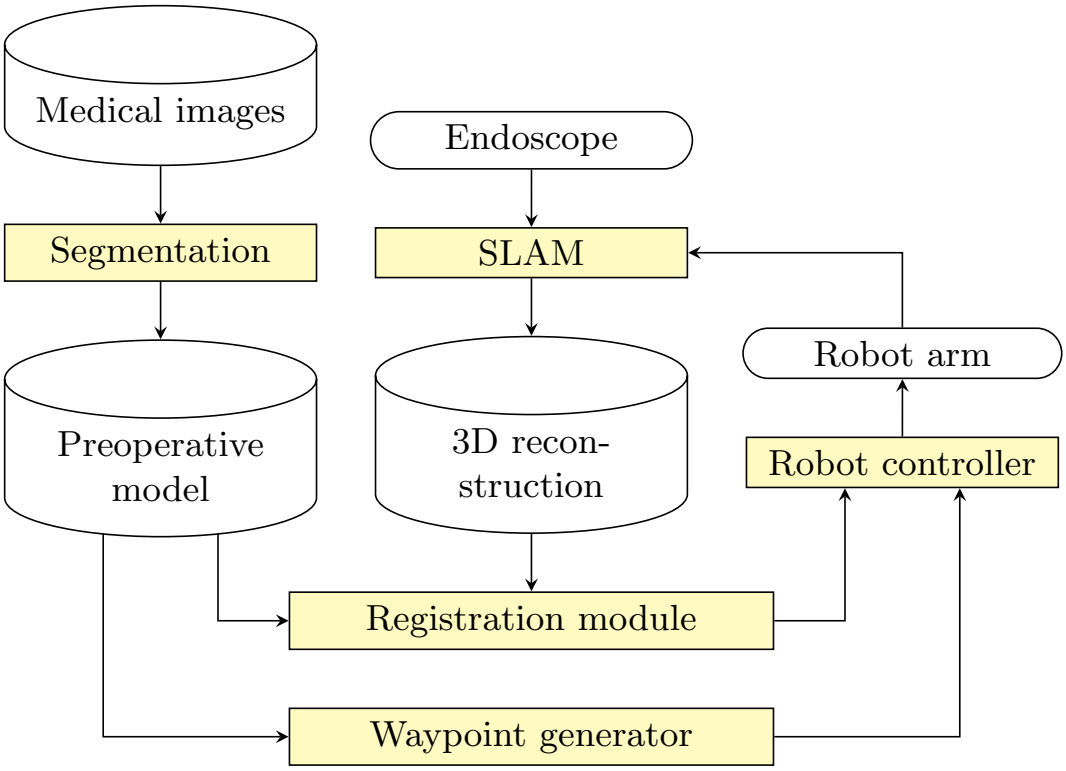


Fig. 6.2: The proposed software architecture.

6.1.2 Pre-operative model

The pre-operative model used in this work is obtained from an MRI scan of the anatomical phantom used in our experiments. In particular, the semi-automatic segmentation approach provided by ITK-SNAP framework [218] is exploited to extract the structures of interest.

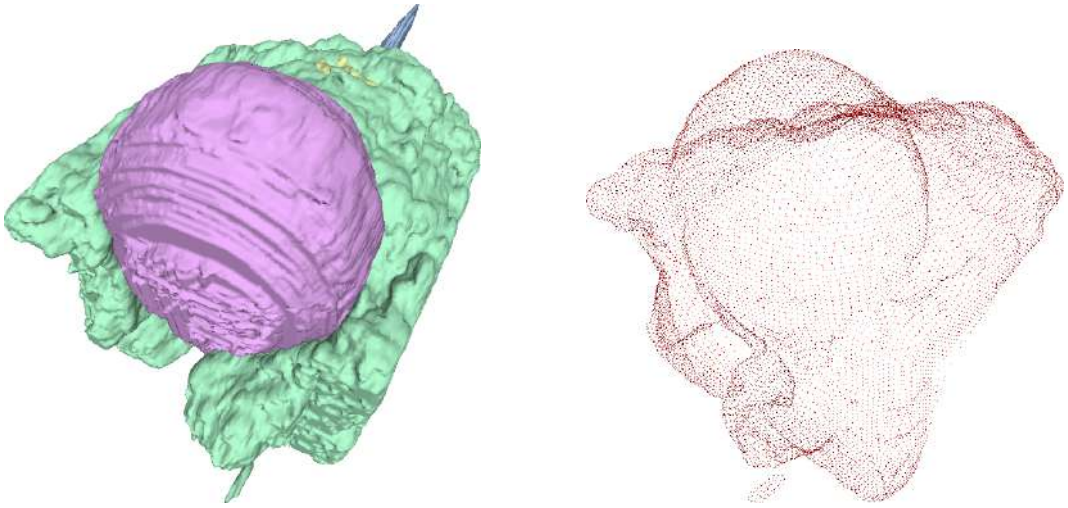
In order to register the pre-operative model (Figure 6.3) to the partial view provided by the endoscope, we perform an initial step to extract the visible surface from the complete model, which is the only portion that can be aligned with the camera view. Removal of occluded and unreachable parts is possible if an a priori estimation of the ECM pose with respect to the operational area is available. Since the endoscope movements are restricted by the remote centre of motion of the camera holder, it is possible to manually select and discard the parts of the model which cannot be seen.

Once the portion of the MRI used for the registration is defined, as shown in Figure 6.3b, we used the Poisson surface reconstruction approach [81] to smooth the artefacts introduced during the segmentation and we recompute the normal per vertex on the aforementioned surface.

The final 3D pre-operative model is converted in a point cloud representation discarding the faces from the mesh.

6.1.3 Scaled ORB-SLAM

The most important part of a SLAM system is the initialisation step. It is necessary to define a coordinate system for camera pose estimation and 3D



(a) 3D model (b) Point cloud

Fig. 6.3: The anatomical model extracted from the MRI. a) the segmented 3D model, b) the final point cloud used in the registration phase.

reconstruction in an unknown environment. Therefore, in the initialisation, the global coordinate system should be first defined, and a part of the environment is initially reconstructed in the global coordinate system. For monocular SLAM it is harder, because the depth cannot be recovered from a single image frame.

To initialise the map ORB-SLAM computes two geometrical models: a homography assuming a planar scene, and a fundamental matrix assuming a non-planar scene. The map initialisation is done when the two-view configuration is safe, detecting low parallax cases and well-known twofold planar ambiguity [102], otherwise the initialise map would be corrupted.

Once the map initialisation is started the first step of the algorithm consists in finding the initial correspondences (x_c, x_r) between the current frame F_c and the reference frame F_r extracting the ORB (Oriented FAST and Rotated BRIEF) features, which are rotation invariant and robust against noise. When there are enough matches, the initialisation procedure starts to compute a homography H_{cr} and a fundamental matrix F_{cr} using the following equations:

$$x_c = H_{cr} x_r \qquad x_c^T F_{cr} x_r = 0 \tag{6.1}$$

with the normalised DLT and 8-point algorithms respectively as explained in [66]. At each iteration it computes a score S_M for each model $M \in \{H, F\}$, H for the homography, F for the fundamental matrix. Afterwards a model is chosen according to this equation

$$R_H = \frac{S_H}{S_H + S_F} \tag{6.2}$$

where R_H is the robust heuristic. If $R_H > 0.45$ the selected model is the homography, otherwise the fundamental matrix. If the scene is planar it can be explained by a homography, otherwise it is selected the fundamental matrix.

Once the model is selected the pose and the motion of the camera has been estimated and the map reconstruction can be started.

Our method adds to the standard initialisation the kinematic measurements of the ECM to estimate the scaling factor. During the initialisation phase we keep track of the real position of the camera with respect to the remote centre of motion of the ECM. Let P_i and P_f be the homogeneous transformation of the camera pose, with respect to the ECM base frame, registered at the beginning and at the end of the correspondence matching of ORB-SLAM. The first virtual camera pose of the SLAM is no longer the one estimated during the correspondence matching but is substituted by the relative transformation of $P_f^i = P_i^{-1} P_f$. Then assuming zero rotation during the initialisation phase the scaling factor s can be computed as the ratio of the measured translation t_f^i and the translation $\hat{t}_f$ estimated by the SLAM

$$s = \frac{|t_f^i|}{|\hat{t}_f|}. \tag{6.3}$$

The translation vectors t_f^i and $\hat{t}_f$ are extracted from the homogeneous matrix P_f^i and $\hat{P}_f$, where $\hat{P}_f$ is the position of the virtual camera computed by the SLAM during the initialisation phase.

Finally, we apply the transformation P_f^i and the scaling factor s to the initial scene reconstruction in order to keep the same relative distance between the points and the new camera position. The assumption of zero rotation during the initialisation phase ensures that the relative motion of the 2D ORB features detected by the correspondence matching is due only to a translational movement and so Equation 6.3 is the desired scaling factor.

6.1.4 3D model registration

The registration is based on a feature-based initial alignment followed by a non linear least squares minimisation of the point-to-plane distance between the two point sets (the SLAM 3D reconstruction and the 3D pre-operative model). To prevent convergence issues related to the different spatial sampling we voxelise both the SLAM map and the pre-operative model with the same step size in order to have the same spatial point density.

Initial alignment

The initial alignment of the two point clouds is required because the ICP algorithm converges easier to a feasible solution if the input point clouds have been already partially aligned. The partial alignment is done using a featured based correspondence grouping which provides an initial transformation matrix M_g applied later to the source point cloud to increase the robustness of ICP. From the two point clouds we extract two sets of Intrinsic Shape Signatures (ISS) [224] key-points,K_s for the source point cloud and K_t for the target one [223].

The ISS key-points are local descriptor which are generally used in application like registration, object recognition, categorisation and are known to be stable, repeatable and discriminative. An ISS is obtained counting the weighted sum of points laying in a local 3D histogram built in a spherical angular space constructed around each feature point. For each point cloud the initial set of ISS key-points is encoded in a set of vectors using the Fast Point Feature Histogram (FPFH) [165] which allows robust multi-dimensional descriptor of the local geometry around a point.

To estimate M_g, a correspondence set C must be computed. Let $C = \bigcup_i C_i$ where C_i is the correspondences set for each vector f_i in the source feature set. The correspondences set is composed of the nearest neighbours of f_i in the FPFH feature space. The set C is then refined applying a cascade of correspondence rejection methods. First of all we enforce a normal direction matching and subsequently on the resulting subset we make the correspondence injective applying a duplication filtering which keeps only the closest neighbour. Finally C is further refined applying Random Sample Consensus (RANSAC) to estimate a transformation between the two correspondences set. The elimination of outlier correspondences is based on the Euclidean distance between the points once the computed transformation is applied to the source point cloud. The final transformation M_g is computed on the filtered set C using SVD.

Iterative closest point

The Iterative Closest Point (ICP) algorithm exploiting the point-to-plane method provides a more robust and much faster convergence [21] than the classical position based implementation. It differs from the standard ICP technique since it minimised the distance between the source point s_i with the plane defined by the target point t_i and its normal n_{t_i}

$$M_{opt} = \arg\min_M \sum_{i=1}^{N} ((M s_i - d_i) \cdot n_{t_i})^2 \tag{6.4}$$

where M_{opt} is the transformation matrices which aligns the source point cloud to the target one. The source point set s_i in our method is previously transformed according to the initial alignment procedure using the transformation matrix M_g.

Given a source, i.e. actually the SLAM output, and a target point cloud, i.e. the pre-operative model, each iteration of the ICP algorithm establishes a set of pair-correspondences between points. The output of an ICP iteration is the 3D rigid-body homogeneous transformation M that aligns the source points to the target point cloud such that the total error between the corresponding points is minimised. For a rule of thumb in case of registration of dense to sparse point cloud ICP performs better if the source point cloud is the sparse one. For the 3D pre-operative model the normal vectors are precomputed; for the SLAM point cloud the normal vectors are computed just before the registration routine. The estimation process is also taking into account the camera position in order to have all the normal vectors pointing towards it.

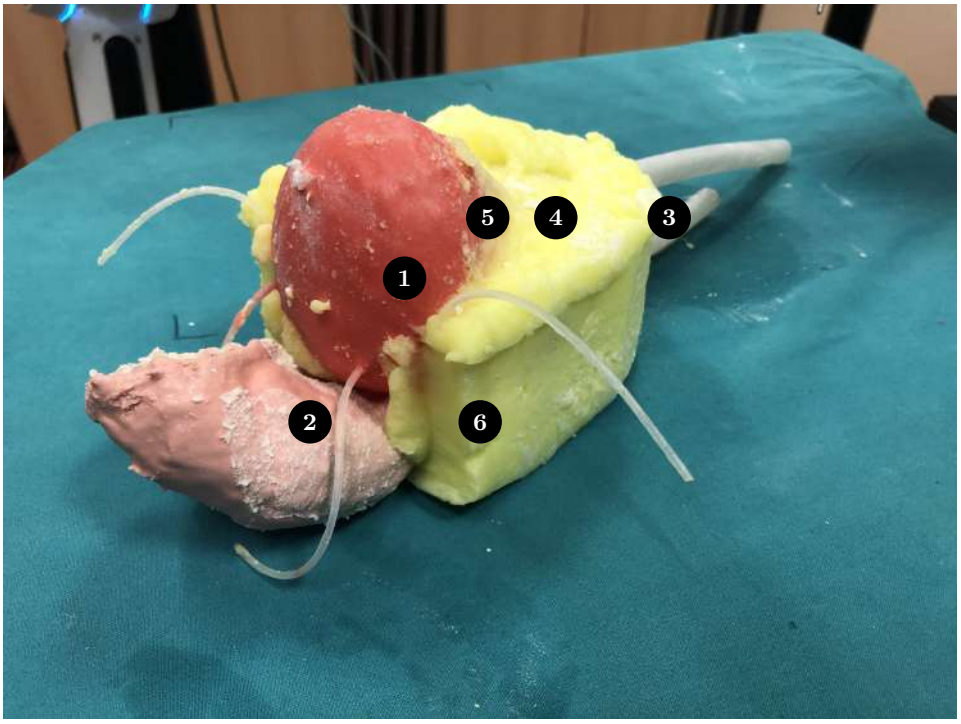


Fig. 6.4: The phantom (ACMIT *Gmbh*, Austria) used during the experiment. The phantom is composed of bladder (1), rectum (2), urethra (3), prostate (4), seminal vesicles (5), fat (6).

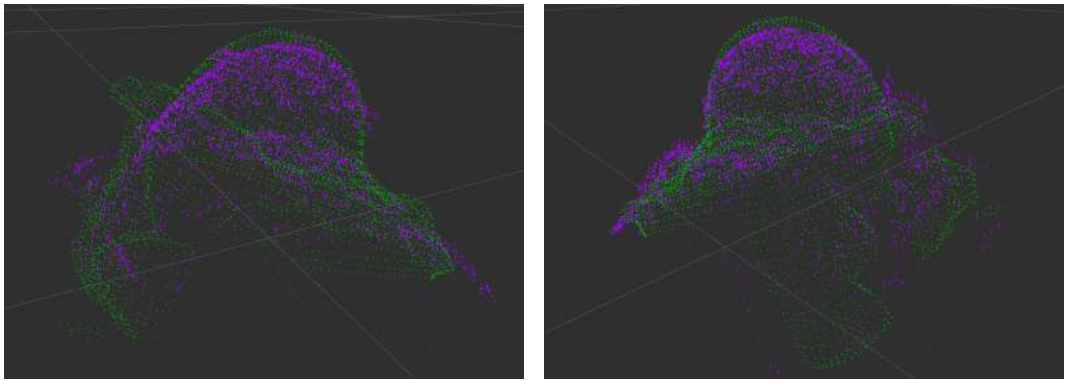
6.2 Results

The method presented in Section 6.1 has been tested in a surgical simulator with the aims of performing autonomously one of the phase of a radical prostatectomy. Figure 6.4 shows the phantom of the lower abdominal used in the experiments. It is composed of the following anatomical structures: bladder, prostate, seminal vesicles, rectum, urethra and fat. The phantom is developed by ACMIT *Gmbh*, Austria.

The 3D model of the phantom obtained from MRI segmentation (Figure 6.3a) is registered to the SLAM sparse 3D reconstruction. In the following, we will show a quantitative and qualitative evaluation of our methodology. We tested the scale estimation method measuring the size of the visible anatomy both in the 3D reconstruction and in the MRI (assumed to be the ground truth). We tested the system accuracy comparing the position of a set of fiducial points in the MRI with the position of the robot end-effector during contact. Finally, we execute the bladder pushing task in an autonomous way.

6.2.1 Scaling evaluation and error bounds

To evaluate the reliability of the autonomous scale estimation we measured the size of the phantom and we compared it with the same measurement obtained from the 3D reconstruction. The scale estimation error is the combination of



(a) (b)

Fig. 6.5: The registration of the pre-operative map. The purple dots are part of the point cloud obtained by SLAM, the initial map is shown in green.

the 3D reconstruction error and the hand-eye calibration T_e^c. The computed scale error is 4 mm.

The system accuracy is evaluated measuring the positioning error ϵ

$$\epsilon = \left\| \begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} - \begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} \right\| \tag{6.5}$$

where (x_m, y_m, z_m) are the fiducial points selected on the surface of the 3D model, and (x_r, y_r, z_r) are the corresponding positions of the end-effector. Since our phantom doesn't have any specific landmarks in it we selected the corners along the boundary of the fat as fiducial points. The arm used to perform this measurement is the patient side manipulator (PSM) of the da Vinci® which has a positioning error of 1 mm with respect the common reference frame and is assumed to be the ground truth. The overall error, is around 6.7 mm, is estimated as the mean of the error computed on the landmark set. It is worth remarking that the measured errors reported in Table 6.1 include all the calibration errors of the system. The 3D reconstruction obtained by the SLAM is referenced to the camera reference frame and in order to be placed in the common reference frame (world) we need to traverse multiple transformations, as shown in Figure 6.6.

The positioning error is the combination of: (i) the shared reference frame calibration errors T_w^{eb}, T_w^{pb} (which affects the ECM and the PSM1 end effector position), (ii) the hand-eye calibration T_e^c (which affects the positioning of the endoscope with respect to the ECM end effector), and (iii) the registration error. An estimation of the registration error is provided by the ICP itself. The Euclidean norm of the misalignment between the source point cloud (pre-operative model) and the target point cloud (SLAM model) is of 0.560 mm.

6.2.2 Bladder pushing

The task of pushing down the bladder is needed to create the space for the main surgeon to resect the prostate during the radical prostatectomy proce-

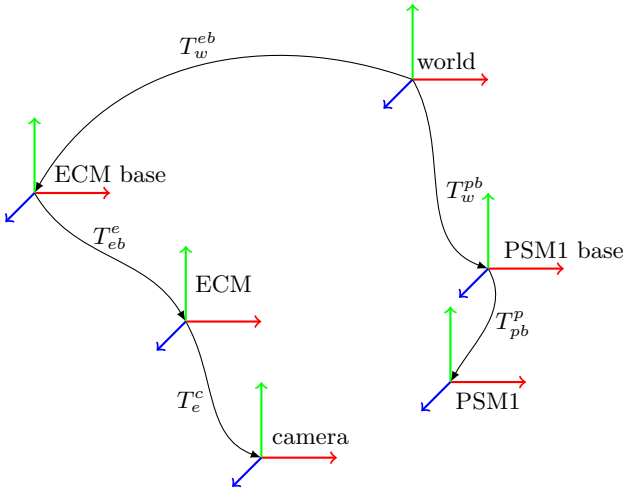
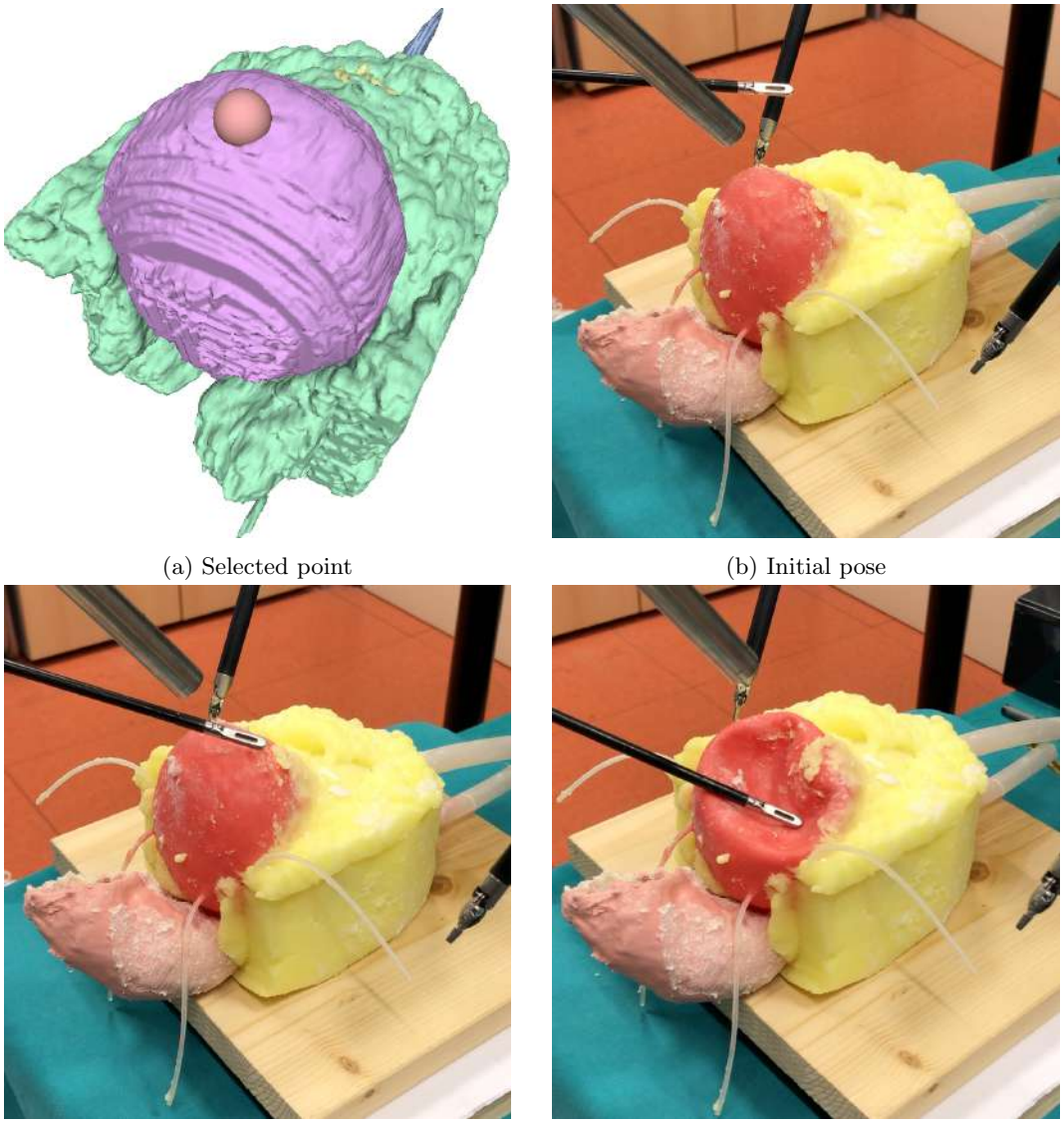


Fig. 6.6: The reference frames involved in the error budget evaluation (the axes direction of the reference frames are only for visualisation purpose).

Table 6.1: Overall system accuracy evaluation, the position of the points are expressed in $\mathbb{R}^3$.

MRI (x_m, y_m, z_m) (mm)	Reference (x_r, y_r, z_r) (mm)	Error ϵ (mm)
6.33, -60.41, 62.99	9.00, -58.00, 59.00	5.4
7.59, 22.05, 48.37	6.00, 29.00, 52.00	8.0

dure. During R-MIS it is performed by the assistant surgeon using standard laparoscopic tools. In this work we want to execute this task in an autonomous way just by providing to the control architecture of the autonomous arm the target points on our 3D model. The task of pushing down the bladder can be modelled by a finite state machine composed of four states: approaching the apex of the bladder (S0), push down the bladder (S1), stay still (S2), and leave movement (S3). The approach point was defined as the highest point of the bladder with respect to the MRI reference frame. The target position, reached during the *push down* motion, is obtained applying a vertical displacement starting from the approach point. Since in the surgical environment there are also the da Vinci® instruments, the movement of the assistant robot must be collision-free. A possible approach tailored to R-MIS has been proposed in [168]. Figure 6.7 shows a few snapshots taken during the *bladder pushing* experiment. The approach point selected by the system is depicted in Figure 6.7a: after the registration phase, the position is transformed in the robot reference frame and sent to the robot. As shown in Figure 6.7c the end effector of the robot is placed nearby the desired point. Once the point on the surface of the bladder has been reached, the next step of the procedure is to reach a point placed inside the bladder in order to push it down. The system calculates that point taking a vertical displacement of 0.045 m along the vertical axis starting from the approach point. Figure 6.7d shows the robot end effector at the desired target position. Supplementary material includes a video with all



(a) Selected point

(b) Initial pose

(c) Bladder approach

(d) Bladder push

Fig. 6.7: An example of the bladder pushing phase. a) the red sphere represents the point selected directly on the MRI, in this case the apex of the bladder, b) the robot initial position, c) the robot end effector once it has reached the approach position over the bladder, d) the robot end effector at the target position.

the phases of the experimental setup and it shows the bladder pushing phase with several approaching points.

6.3 Discussion and conclusions

In this work we proved the feasibility of using point cloud based registration in order to plan the motion of a semi-autonomous system using the pre-operative data of a patient. We tested our solution in a real environment using the

da Vinci® endoscope as the video source for the monocular SLAM. A realistic phantom has been used to test our method during the initial phase of a radical prostatectomy (i.e. the scenario at the beginning of the procedure can be assumed to be motionless). The registration of medical images with the environment is carried out by aligning a subset of vertices of the 3D anatomical model, extracted from pre-operative images, to the outcome of SLAM. A refinement of the initial registration is done using ICP. The registration accuracy shown in Table 6.1 allows to perform semi-autonomous tasks. Even if the experiments are done in a open environment, the movement of the arm is limited by its remote centre of motion and this lead to a mostly linear trajectories. The robot used in the experiments has 4 degrees of freedom: three for the positioning of the end effector and one for the rotation of the grasper along the tool axis. In the future we will investigate the integration of this methodology within a more complex task where the SLAM algorithm has to be extended to account for soft tissues deformations. Moreover we plan to update the pre-operative model with the deformation and perform the registration online with the extended SLAM.

7

Medical SLAM

For Minimally Invasive Surgery (MIS), the use of pre- and intra-operative image guidance has well established benefits. However, its application to procedures with large tissue deformation, such as those encountered in cardiovascular, gastrointestinal and abdominal surgery, is still limited.

In the previous chapters we established that in order to carry out a 3D reconstruction it is necessary to have a common reference frame and to have the pre- and intra-operative information and how we registered this information to the outcome of the SLAM. The use of fiducial markers and optical tracking, as well as intra-operative imaging such as ultrasound, MR and x-ray fluoroscope have been explored extensively. However, the use of vision techniques based on images from laparoscopes/endoscopes during MIS has clear advantages. It does not require the introduction of additional equipment to what is already a very complex surgical setup.

One of the main difficulties to be addressed in soft-tissue MIS is the fast, accurate and robust acquisition of the anatomy during surgery. For Augmented Reality (AR) visualization of subsurface anatomical details overlayed on the laparoscopic video, intra-operative 3D data has to be registered non-rigidly to 3D pre-procedural planning images and models [177]. Tomographic intra-operative imaging modalities, such as ultrasound (US), intra-operative computed tomography (CT) and interventional magnetic resonance imaging (iMRI) have been investigated for acquiring detailed information about the tissue morphology. However, there are significant technological challenges, costs and risks associated with real-time image acquisition in a surgical theatre or interventional radiology suite with traditional instrumentation while providing images with acceptable signal-to-noise ratio (SNR) [93].

In MIS, an increasingly attractive approach involves 3D reconstruction of soft-tissue surfaces using the endoscope itself by interpreting the properties and geometry of light reflecting off the surfaces at the surgical site [123]. Optical techniques for 3D surface reconstruction can roughly be divided into two categories [117]: passive methods that only require images, and active methods that require controlled light to be projected into the environment. Passive methods include stereoscopy, monocular Shapefrom- X (SfX) and Simultaneous Localization and Mapping (SLAM) while the best known active methods are based on structured light and Time-of-Flight (ToF). Both active and pas-

sive technologies have found successful applications in a wide spectrum of fields including domestic and industrial robotics, and the film and games industries.

Reconstruction of the patient anatomy for MIS, however, poses several specific challenges that have not yet been solved. While many applications focus on the 3D reconstruction of static scenes, the methods applied in MIS must be able to cope with a dynamic and deformable environment. Furthermore, tissue may have homogeneous texture making automatic salient feature detection and matching difficult. The critical nature of surgery means that techniques must have high accuracy and robustness in order to ensure patient safety. This is particularly challenging in the presence of specular highlights, smoke, and blood, all of which occur frequently in laparoscopic interventions. New technologies in the operating room also require seamless integration into the clinical workflow with minimum setup and calibration times.

In this chapter, we present how we modified the architecture based on ORB-SLAM2 presented in the previous chapter to manage some simple deformations and how to recognize and separate static from dynamic features, following the scheme shown in the figure 5.4.

7.1 Breathing compensation

One of the most intuitive and natural deformations that the human being has is breath. Respiratory system modeling has been extensively studied in steady-state conditions to simulate sleep disorders, to predict its behavior under ventilatory diseases or stimuli and to simulate its interaction with mechanical ventilation. A typical respiratory cycle is asymmetrically periodic and can be modeled as:

$$Y_i = C + \alpha \sin(\omega T_i + \phi) + E_i \tag{7.1}$$

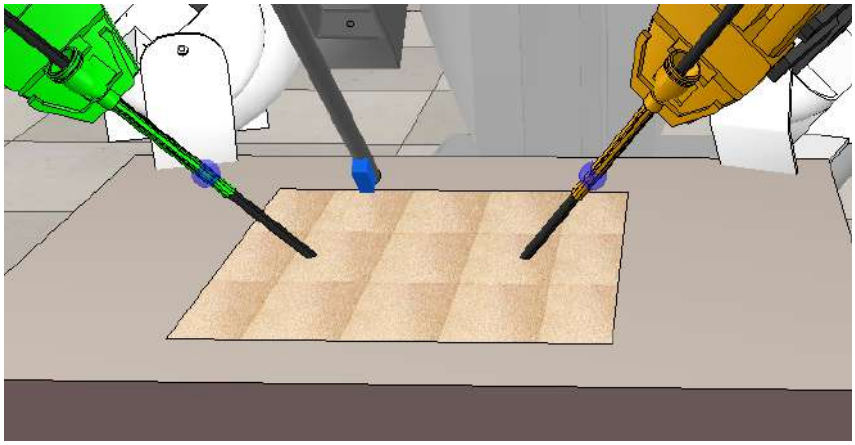
where C is constant defining a mean level, α is an amplitude for the sine wave, ω is the frequency, Ti is a time variable, ϕ is the phase, and E_i is the error sequence in approximating the sequence Y_i by the model. This sinusoidal model can be fit using nonlinear least squares.

To obtain a good fit, nonlinear least squares routines may require good starting values for the constant, the amplitude, and the frequency. The respiration cycle can be estimated using any point of the surface, assuming it can be tracked and the motion is along a single axis.

The transformation from the global coordinate system to the respiration coordinate system is unique to each point. This means that points on the surface of the surface can move and deform in independent directions but share the same respiration model. Given a model of respiration, it is therefore possible to estimate the dynamic tissue motion using the inverse PCA transformation matrix and a given point in the respiration cycle. The parameters of the (7.1) are estimated using Levenberg-Marquardt minimization algorithm where the problem is posed as a least squares curve fitting.

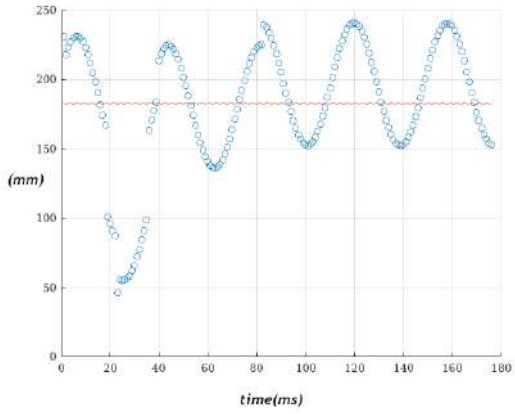
In our framework, we extended the *loop closing* section of the architecture, shown in 5.4, by using the *g2o* [62] library, which provides an implementation

of the functions needed to estimate the breathing model. Finally, we tested the outcome of the ORB-SLAM presented in the previous section in a simulated environment shown in 7.1a.



(a) V-Rep scene

(b) Skin



(c) Setup

Fig. 7.1: Simulation setup of the respiratory cycle during an intervantion: a) DVRK setup inside V-Rep simulator; b) texture of the skin; c) example of a fit. The blue points represent the real points of the map created by the SLAM, the red ones instead the result of the fit.

Tests were done by using V-Rep [162], which is an open source robotic system simulator, made by a company called Coppelia Robotics. It is based on a distributed control architecture: each object/model can be individually controlled via an embedded script, a plug-in, a Robot Operating System (ROS) node, a remote Application Programming Interface (API) client, or a custom solution. V-Rep is a highly customizable simulator, every aspect of a simulation can be customized, also the simulator itself. It can be programmed with seven different programming languages. V-Rep's dynamics module currently supports three different physics engines: the Bullet physics library, the Open Dynamics Engine (ODE) and the Vortex Dynamics engine.

In our experiments we used ODE engine and the simulated scene represents the DVRK operating room. To simulate breathing, we have created a script on a geometric model where we have attached a texture that represents the skin (Fig. 7.1b). Figure7.1c shows the result of the fit, the blue points represent the real points of the map created by the SLAM, the red ones instead the result of the fit. From this graph we can see when the SLAM algorithm closes the loop through the jumps and optimizes the map. The tests were done only in simulation and not on a real setup, due to COVID-19 rules to access the laboratory.

7.2 Dynamic 3D point detection

An important part of this research has been to study dynamic features: it is necessary to identify and separate the areas that deform from the static ones to treat them differently. For each frame, we detect if the scene has changed through the analysis of the histogram between the corresponding frame and select several keyframes which observe the same scene. Then we calculate the correlation coefficient between keyframes and current frame. If the correlation coefficient is less than a threshold K, the scene of current frame has possibily changed. So, we project the map points in the keyframe to the current frame. For a 2D feature point p in keyframe, its corresponding 3D point is denoted as P and its projection in current frame is denoted as p'. We compute the appearance difference of the patch centred at p in the current frame with respect to the patch centered at p' in the keyframe:

$$D(p) = \min_{d} \sum_{a \in B(p)} |I_a - A_a I_{a'+d}| \tag{7.2}$$

where $B(p)$ denotes the image-patch centred at p. We apply an affine warping A_a as[183] to current patch, because the keyframe is typically a little far from current frame. Due to the estimation error, the depth of P may deviate from the true value. So, we violate the epipolar constrains to get a subpixel accurate feature correspondence (directly add a little translation d to projection position p'). If the difference $D(p)$ is larger than a threshold, it is very likely that the point P has changed its position or occluded by other objects.

7.2.1 A method to distinguish static and dynamic features

Various approaches are used to detect dynamic features in the scene and these approaches can be roughly classified into three types:

- dynamic features detection depending solely on geometry information
- dynamic features detection depending solely on semantic information
- dynamic features detection through naive combination of the results from geometry calculation and semantic information in a loosely coupled way.

Most existing works on the geometry of multiple images rely on the assumption that the observed scene is rigid. The rigidity constraint allows to derive matching relations among two or more images, represented by e.g. the fundamental matrix or trifocal tensors [68]. These matching tensors encapsulate the motion and the intrinsic parameters of the cameras which took the underlying images, and thus all the geometric information needed to perform 3D reconstruction. Matching tensors for rigid scenes can also be employed for scenes composed of multiple, independently moving objects, which requires however that enough features be extracted for each object, making segmentation. On the other hand, there is a growing body of literature dealing with the case of independently moving features, often termed as dynamic features. The goal of these works is to provide algorithms for dynamic structure and motion recovery as well as matching tensors for images of dynamic features.

Kundu et al. [90] construct the fundamental matrix from robot odometry to define two geometric constraints, one of which is derived from the epipolar geometry. According to the epipolar geometry constraint, a matched feature in the subsequent frame is most likely to be considered as dynamic if it resides too far from the epipolar line (Figure 7.2a). The key in this kind of method is the estimation of the fundamental matrix, if a relatively reliable fundamental matrix can be acquired, then most of the dynamic features can be easily detected. The fundamental matrix can be acquired using purely visual method, such as the 5-point algorithm or 8-point algorithm [67].

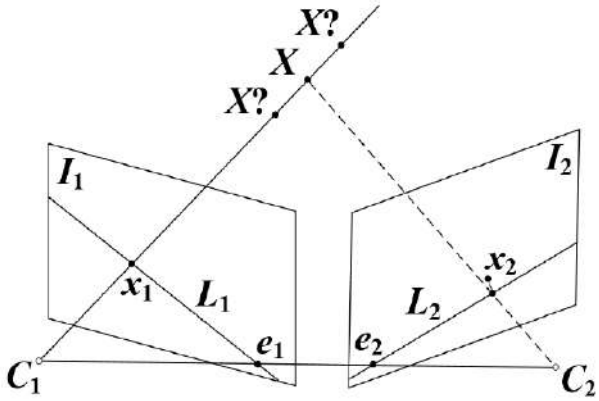
Zou and Tan [227] classify map points as dynamic or static at every frame by analyzing their triangulation consistency. They project features from the previous frame into the current frame and measure the reprojection error of the tracked features. The error should be small if the map point is static, otherwise the map point is classified as dynamic.

Wang et al. [200] take current RGB image, previous image and current depth image as input, they firstly cluster the depth image into several objects, extract features in current RGB image and count the number and percentage of features on each object. Then features correspondences between current RGB image and previous RGB image are used to calculate fundamental matrix, which is subsequently used to filter out outliers, the number and percentage of remaining inliers on each object are counted again. The remaining inliers are used to calculate fundamental matrix one more time and the following procedure is the same as before. At last a moving objects, judgment model is designed based on the statistical characteristics obtained above, and once an object is considered as moving, all features on it are eliminated.

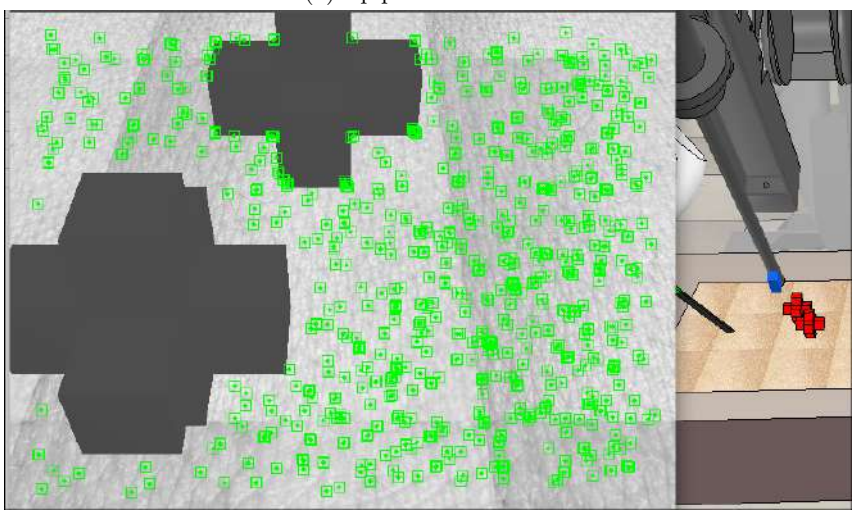
Sun et al. [184] adapt the codebook learning and inference mechanisms form to deal with the SLAM problem in dynamic environments. Their motion removal approach consists of two online parallel process : the learning process that builds and updates the foreground model; the inference process that pixel-wisely segments the foreground with the built model.

Fan et al. [46] construct a camera motion model for the moving platform, then decompose the motion model into two parts: translation and rotation. At last, two constraints are proposed to locate the dynamic regions.

With the quick development of deep learning in recent years, computer vision tasks such as object detection and semantic segmentation can be solved excellently and the accuracy can even outperform human. In SLAM system, when a new frame is coming, by applying advanced CNN architectures like YOLO [155] , SSD [99] , SegNet [6] , Mask-RCNN [69] , the semantic label



(a) Epipolar constraint



(b) Geometry constraint example on V-Rep scene

Fig. 7.2: A static feature should satisfy epipolar constraint in multiple-view geometry, while a dynamic feature will violate the standard epipolar constraint. In the simulated scene (b) we can notice that the closest object has no features on it, this is because movement was imposed by script.

of the extracted features can be acquired. Then features lying on semantically dynamic objects such as people or cars are considered dynamic and removed.

Zhong et al. [221] use object detection network SSD to detect movable objects, such as people, dog, cat and car. For instance, once a person is detected, it is regarded as a potentially moving object whether it is walking or standing and all features belong to this region are removed.

Zhang et al. [219] use YOLO to get semantic message, they consider features which are always located on the moving objects as unstable and filter them out.

Wang et al. [201] propose a step-wise approach that consists of object detection and contour extraction to extract semantic information of dynamic objects in a more computationally efficient way.

Xiao et al. [210] use SSD object detection network running in a separate thread to get prior knowledge about dynamic objects, and the features on dynamic objects are then processed through a selective tracking algorithm in the tracking thread, to significantly reduce the error of pose estimation. Some recent works combine the dynamic detection results from geometry calculation and the semantic information.

Yu et al. [217] proposed DS-SLAM A Semantic Visual SLAM towards Dynamic Environments, based on ORB-SLAM2. They use SegNet to get pixel-wise semantic label in a separate thread. If a feature is segmented to be "person", further moving consistency check is then conducted using epipolar geometry constraint. If the check result is dynamic, then all features with the semantic label "person" will be classified as dynamic and removed. This method actually treats features with label "person" as a whole and takes the intersect of two results: only features that are both semantically and geometrically dynamic are considered as dynamic.

Bescos et al.[8] combine the results of semantic segmentation from Mask R-CNN and multi-view geometry. They actually take the union of the two results: features either semantically dynamic or geometrically dynamic are all considered as dynamic.

Linyan Cui et al.[26] proposed Semantic Optical Flow SLAM (SOF-SLAM), which is built on ORB-SLAM2. SOF-SLAM fully utilizes the complementary characteristic of motion prior information from semantic segmentation and motion detection information from epipolar geometry constraint, while the existing SLAM systems either depend solely on semantic information or geometry information, or naively combine the results of them to remove dynamic features.

These works prove that deep learning algorithms have become much better at modelling complex tasks. This allowed researchers to be able to perform high quality semantic scene segmentation in many challenging applications, such as autonomous driving, social scene analysis, etc. A partial reason for this success is also the huge size of the datasets that have been released in recent years, e.g. ADE20K [225], Mapillary Vistas [131] and Cityscapes [25]. In this chapter, we combine the results from geometry calculation and semantic information extracted from deep learning algorithm to distinguish dynamic features on sur-

gical tasks from static features by segmenting the surgical instruments and the anatomy, rejecting the dynamic features as outliers, and using only the static features to track the endoscope position as well as to complete the subsequent SLAM process.

7.3 Semantic Registration of CT Scan and Intra-Operative Anatomical 3D Reconstruction

Considering the specific task of semantic segmentation *in a surgical environment*, deep learning algorithms have not been able to provide good enough results. One main reason for this performance degradation in the case of medical images is the sheer complexity of the scene. In the scenarios that are normally contemplated by researchers, objects have a fixed color structure and fixed shapes. This rigid structure and the spatial dependency between the different components make them easier to identify. Objects in the images have very clear boundaries and are all of the component are always connected to each other. In a surgical environment instead, organs do not hold their shape and the boundaries between these organs are not very clear. Moreover, the scene in the surgical cavity keeps changing, due to a number of factors such as bleeding, presence of smoke, etc. The fact that semantically different organs have similar color and/or texture makes them even more difficult to differentiate.

All the recent semantic segmentation models are based on encoder-decoder architecture. Each model contains two submodules: encoder and decoder. Encoder is used to compress the image information and extract the low dimensional feature representations. Most of the encoder are adapted from the standard classification CNN architectures, such as, AlexNet [87], Inception architecture [186], ResNet [70]. Decoder submodule upsamples the features into original size using deconvolution operation and then uses softmax layer to produce the class probabilities for each pixel.

Fully Convolutional neural Networks (FCNN) were first introduced in semantic scene segmentation by Long et al. [100]. FCNN networks have a lot of advantages over traditional Convolutional Neural Networks (CNN). Most of the parameters of a CNN model are in the fully connected layers [176]. As FCNNs do not contain any fully connected layer, this type of architectures have much lower number of parameters. Additionally, FCNN architectures are scale-independent, i.e. these models can take input images of any size, scale and aspect ratio in contrast to conventional CNNs which only accept input of one size. This scale independence makes it convenient to use them for domain adaptation and fine tuning. This also makes them easier to train on images or video frames of different size and aspect ratio.

There is a lot of ongoing work on the various aspects of semantic segmentation modeling. Some papers try to improve quality and accuracy by focusing on the model architecture [20, 116, 163], while others focus on the loss function to design a better loss function, which can accurately define the problem of at hand [101, 211].

Recently, in particular, a wide range of approaches have been proposed to improve the performance of FCN models for semantic segmentation. U-Net is one of the most famous such architectures, and was initially proposed specifically with medical images in mind [163]. This model uses skip connections between the encoder and decoder layers. Initial layer of the model are encoder part of the model. These layers reduce the spatial dimension of the features, pooling layers are place to compress the feature representation. the second part of the model is decoder, where these reduced features are upsampled again to have output of same size as input. Deconvolution layer is used to upsample the features. Additionally, skip connection are used in model to provide the spatial context from the deeper layers of the model. Another architecture, termed V-Net, is inspired by U-Net. The V-Net model follows similar a skip connection approach between the layers but, instead of using 2D images as input, it uses volumetric data (the 3D volume collecting multiple frames together).

In this work we propose a semantic registration for a semi-autonomous surgical robotic system by improving the approach presented in the previous chapter to make use of semantically annotated medical images, such as CT, MRI, and PET scans, and to improve the reliability of 3D registration in the non-static conditions of surgical environments. The architecture revised architecture is shown in Figure 7.3.

In our framework, we register the anatomical model extracted from a CT scan, after applying the semantic information, to its 3D semantic reconstruction obtained by an RGB-D sensor. We aim to understand the state of the surgical scene, i.e., the nature of the events taking place there and the current situation in the surgical cavity. Fully understanding the current scene can help the different autonomous components during the procedure. We validate the method proposed within the SARAS project, using the phantom which is used to perform the Robot-Assisted Radical Prostatectomy (RARP).

7.3.1 Pre-operative model

The pre-operative model used in the experimental validation of this work is obtained from a CT scan of the anatomical phantom for RARP. In particular, the semi-automatic segmentation approach provided by Slicer3D is exploited to extract the structures of interest.

Once the model is extracted, we highlight the structures useful for semantic registration by encoding them with the same color map used for the segmentation: for our experiments, the most prominent anatomies are the bladder, the connective tissue, and the prostate, whenever visible (Figure 7.4).

7.3.2 Semantic Segmentation

We applied the model presented in [77], which is a Generative Adversarial Network (GAN) trained on a few examples of the SARAS phantom setup to include all the replicated anatomical structures plus the catheter, the da Vinci®, and the SARAS tools. Compared to other models, this network presents the advantage of requiring very few training samples to obtain quality results. In

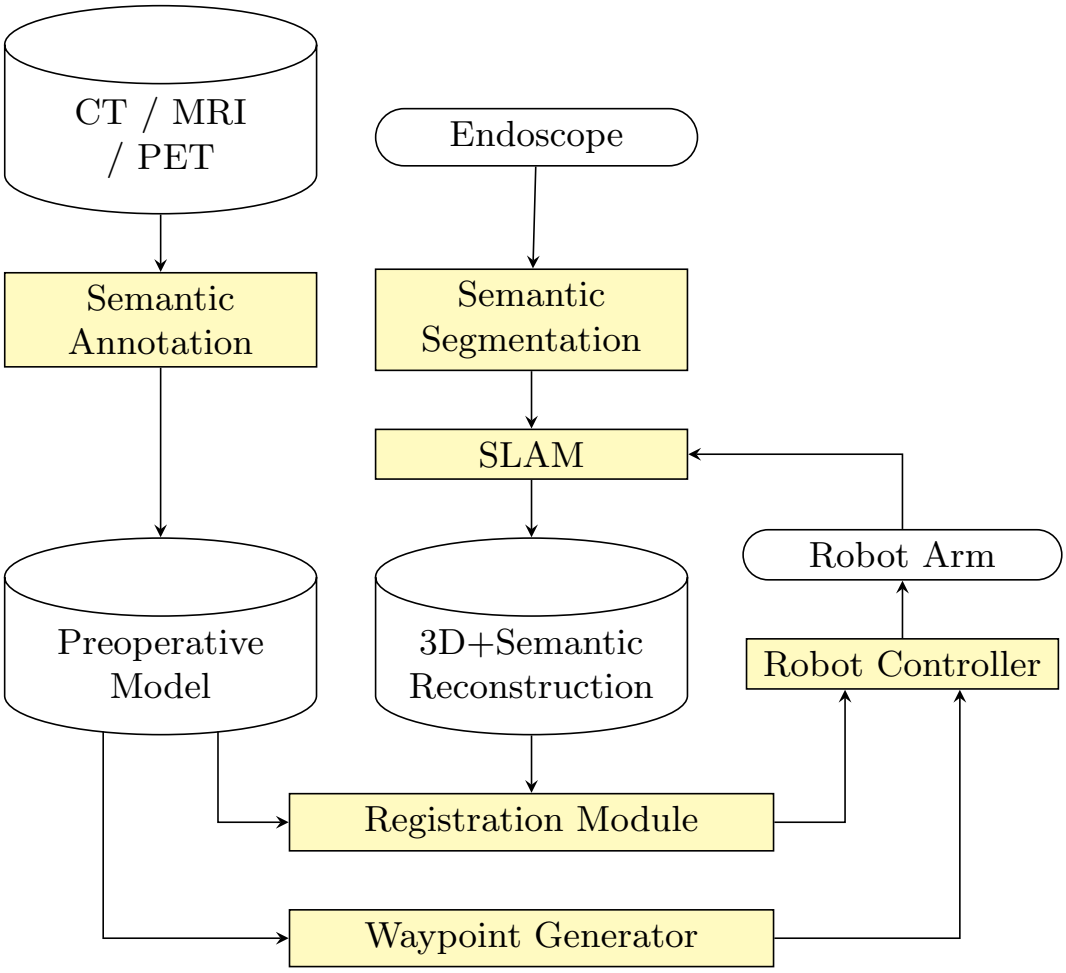


Fig. 7.3: The revised software architecture: the differences from [149] are in the use of medical images and the real time SLAM preceded by a semantic segmentation module.

our tests, only 15 labelled images were ultimately required to obtain the desired segmentation output compared to the thousands usually necessary to train models like *U-NET* [164]. It achieves this by adopting a dual network configuration. A first network, the *Generator* (G) is trained to produce multiple samples of the labelled image starting from the input sample image; for this, the generator is usually a modified u-net structure. The second network, the *Discriminator* (D), operates as a loss function to identify whether the generated image is similar or not to the ground truth (Figure 7.5 presents a simple schematic of the network).

The result is a pixel-to-pixel mapping model (hence the name *pix2pix*) that passes through only the first network that is now capable of "fooling" the best trained discriminator, thus, in this application, it's capable of segmenting the input image. However, given the low granularity, pixel-to-pixel nature of the mapping learned by the model, the output requires some post-processing to clear out noise, namely a median and a clustering filter to reduce scattered patches. The critical structures for the semi-autonomous tasks have been color-

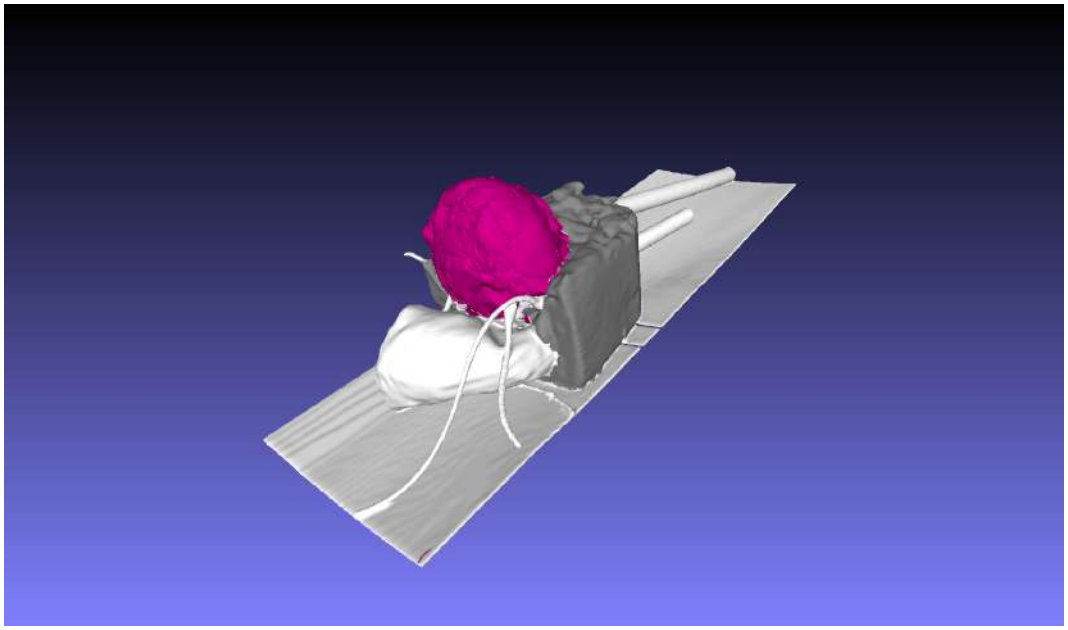


Fig. 7.4: A pre-operative semantically-segmented model

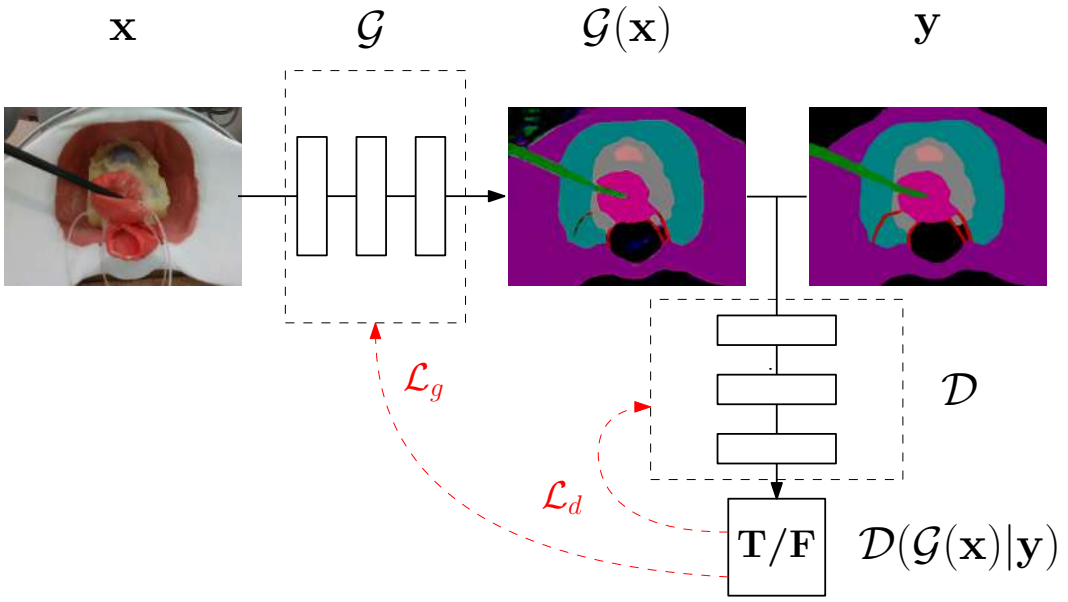
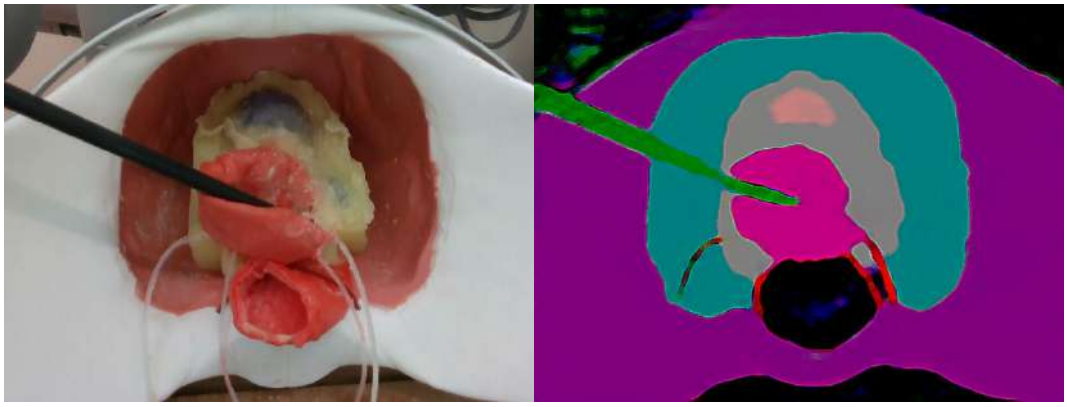


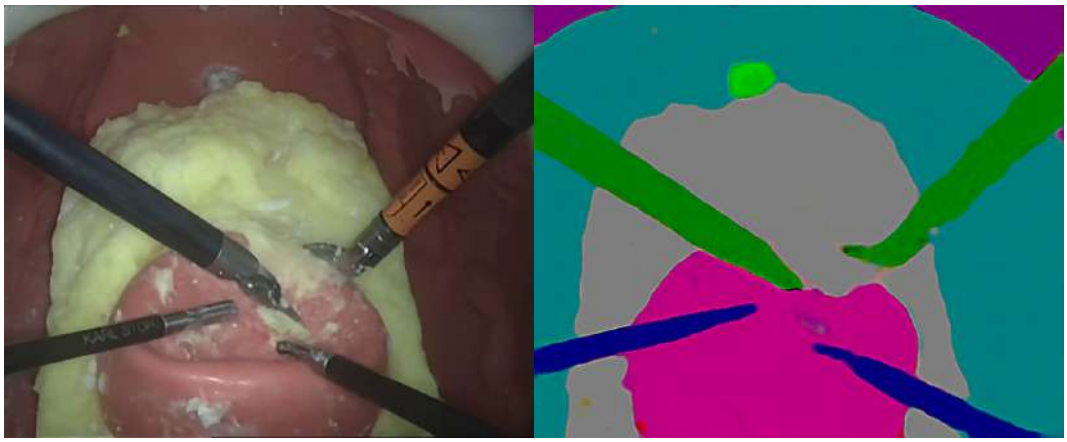
Fig. 7.5: GAN schematics for network training: the discriminator $\mathcal{D}$ operates as a loss function for both itself and the generator $\mathcal{G}$.

coded with the encoding presented in Table 7.1. Figure 7.6 shows two examples of segmentation of scenes acquired from a RealSense® and a da Vinci® endoscope camera perspective (Figures 7.6a and 7.6c respectively). It is possible to immediately verify the good quality of the segmentation, with the da Vinci® tool clearly discernible, apart from a few misidentified spots on the upper-left corner that are outside of the phantom area and, thus, easily ignorable. Most importantly, the bladder is uniformly colored with the rectum clearly separated (we applied the "background" label to the rectum as it is not relevant to

prostatectomy operations). Also important are the vas deferns and the seminal vescicles. As they are semi-transparent they are very difficult to identify but the model does identify them in an acceptable manner. Finally, the prostate is also correctly identified as it is not fully covered in fat at the beginning in this phantom iteration. The main drawback of this family of neural networks is their tendency to generate multiple false positives depending on the amount of distinct semantic mappings used for the training stage. This results in the false identification of anatomical structure in locations where they appear on average in the training dataset. For instance, the prostate could appear around the edges of the instruments even if it is still completely covered in fat material (as it appears slightly in Figure 7.6d) or the ureters appear at the bottom of the bladder even when not visible. This issue can be easily corrected by providing just a single new image to the training dataset and training the model, which is a quick computation given the reduced dataset size.



(a) RealSense® view (b) Segmented image



(c) Endoscope view (d) Segmented image

Fig. 7.6: Examples of real-time semantic segmentation computed over two different cameras.

Color	Semantic Meaning
	Background
	Da Vinci tools
	SARAS tools
	Pelvic Bone
	Pelvic Floor Muscle
	Anterior Prostatic Fat
	Vas Deferns and Seminal Vescicles
	Ureter
	Catheter
	Bladder
	Bladder Neck
	Prostate
	Prostate Neck
	Urethra

Table 7.1: Semantic Scene Color Encoding

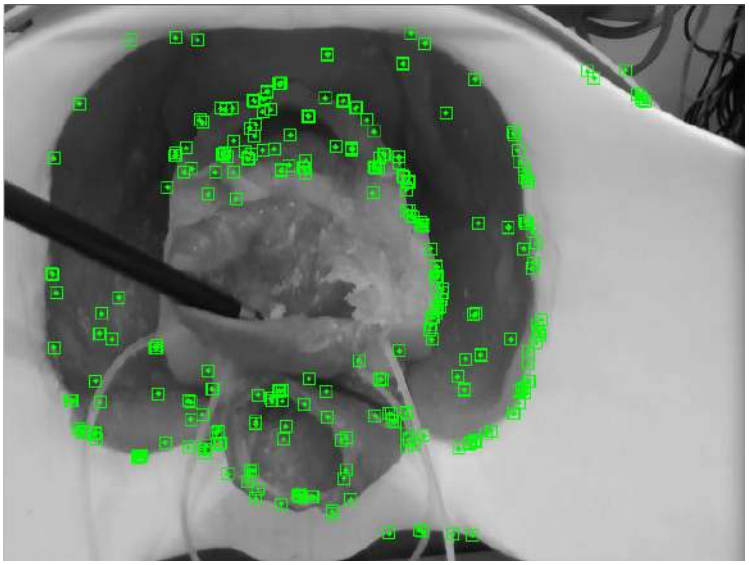
7.3.3 Semantic monocular SLAM

A basic assumption in most current SLAM approaches is that the environment is static. However, active objects exist in many real-world scenes. In the case of laparoscopic surgery, the most typical dynamic objects are surgical tool and the organ or surface being manipulated. To address this issue, identifying the dynamic objects from the static parts, and then discarding them before pose estimation is necessary.

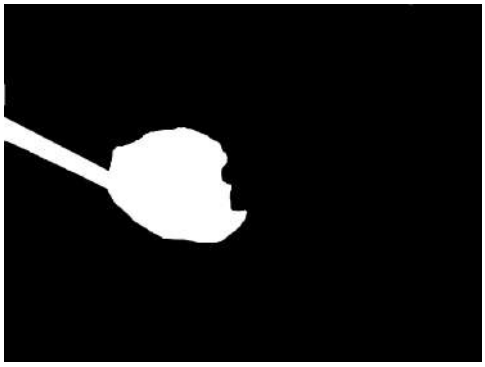
Knowing the surgery to be performed, we know which organs could be moved or the most deformable surfaces and through the use of semantic segmentation we can create a mask around them. Assuming that each input frame contains dynamic features, a mask is obtained by binary semantic segmentation calculation. The features obtained by the ORB feature extraction algorithm in the SLAM system are removed from the feature sequence when they are located in the mask region, while the features in other regions continue to be used for subsequent tracking and mapping. By using the mask to limit the feature detection area and thus prevent the feature points from concentrating on the dynamic objects, false extraction and matching can be avoided.

As shown in Figure 7.7, the feature point detection on the dynamic objects is successfully excluded by using the mask. Fig. 7.7b shows an example of filtering dynamic ORB features by using semantic segmentation results naively as masks. In Fig.7.7a, the ORB features extracted by the SLAM where we subtract the mask, which includes the bladder and the instruments.

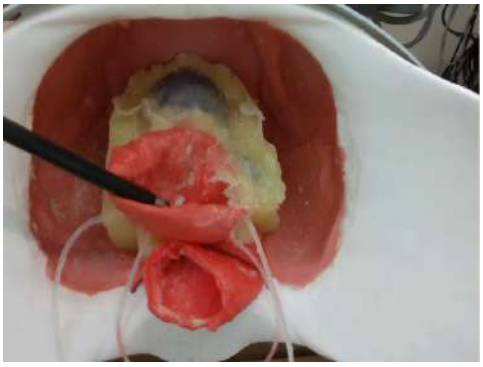
Compared with the original ORB-SLAM2 algorithm, the feature points on the surface of dynamic objects are eliminated using semantic segmentation, and a higher quality of map construction can be obtained. Figure 7.8 shows an example of the sparse pointcloud obtained after removing the instruments, with anatomy mapped with semantic colors.



(a) ORB-SLAM frame



(b) Semantic mask



(c) Original frame

Fig. 7.7: SLAM process frame acquisition: (a) ORB feature detection after applying the semantic mask; (b) the semantic mask of the bladder and instruments; (c) the original RGB frame.

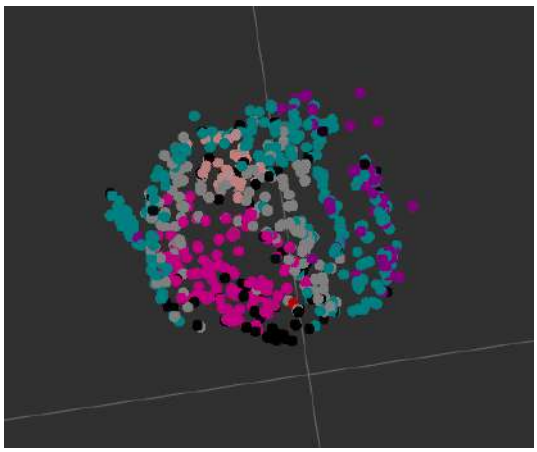


Fig. 7.8: Sparse point cloud encoded with semantic colors

7.4 Discussion and Conclusions

We validate the method proposed within the SARAS project, using the phantom utilized in perform the Robot-Assisted Radical Prostatectomy (RARP).

RARP is a surgical procedure where the surgeon utilises a robotic manipulator to remove the prostate along with, in some cases, the seminal vesicles and the pelvic lymph nodes [143]. The procedure is performed for treatment of prostate cancer. All results of robotic prostatectomy so far indicate the benefits of minimally invasive surgery while also showing encouraging short and long-term outcomes in terms of continence, potency, and cancer control; it is regarded as a major innovation in the surgical treatment of prostate cancer. RARPs are currently performed using either the da Vinci® surgical system or any comparable robotic platform. Surgeons remotely control the instruments of the robotic manipulator using two joysticks available on the console. In the operating room, there must be also an assistant surgeon next to the patient, helping the main surgeon. This work provides a first step towards the full automation of the assistant surgeon's role.

We followed the procedure presented in [138, 171] that bridges the clinical and engineering requirements to improve the effectiveness of both the surgeon, equipped with two da Vinci® instruments and an endoscope, and the semi-autonomous assistant handling two standard laparoscopy instruments. At first, the surgeon identifies the proper plane of dissection to operate on the bladder neck, which is then divided transversely with respect to the urethra, until he/she identifies the urethral catheter pushed through the prostate. At this point, the assistant surgeon, using the right laparoscopic tool, mobilizes the bladder to clear the view, and, with the left laparoscopic tool, raises the prostate. Once the prostate is suspended anteriorly, the main surgeon grasps the tip of the catheter and lifts it upwards to increase access to the lower part of the prostate, including the vas deferens and the seminal vesicles.

After the prostate has been removed, the main surgeon performs the vesicourethral anastomosis. During this phase, the activities of the assistant surgeon consist of avoiding the bladder inflation by keeping it pushed down and, once the suture has been completed, cutting the needle's thread with the scissors [143].

In this work we proved the feasibility of using semantic registration in order to plan the motion of a semi-autonomous system using the pre-operative data of a patient. We modify the ORB-SLAM2 architecture presented in chapter 6 by incorporating semantic segmentation for minimally invasive surgery scenarios. The semantic segmentation network was based on the model presented in [77], which is a Generative Adversarial Network (GAN) trained on the SARAS phantom setup to include all the replicated anatomical structures plus the catheter, the da Vinci®, and the SARAS tools. This network presents the advantage of requiring very few training samples to obtain quality results. Knowing the type of surgery it was possible to remove the dynamic structures, such as the bladder and instruments, and thus improve the SLAM system in order to obtain more accurate mapping results. In the last chapter we will explain how we are going to further improve this framework.

8

Conclusions

This thesis has faced the problem of reaching level 2 of autonomy in robotic surgery, as of the classification of autonomy levels described in [213].

Specifically, in the first part of the thesis the choice of sensors, a new endoscope prototype and the calibration problem has been analyzed, which is fundamental to obtain an accuracy such as to be able to deal with autonomous surgical operations.

The proposed framework for autonomous surgical task execution has been validated through different scenarios. Starting from benchmark training task for surgeons, then to surgical procedure performed on realistic phantoms that simulate the the lower abdominal. One challenge of autonomous robotic surgery is the unpredictability of the anatomical environment and its behavior intra-operatively, depending on the specific patient. To solve the problem of non-rigid body a new SLAM algorithm has been proposed, based on ORB-SLAM2 architecture, which adapts to the anatomical environment by registration of the pre-operative images to the 3D reconstruction and then to distinguish dynamic features from static ones to adapt the algorithm to a dynamic environment. Finally, for the first time (to the best of the knowledge of the author) a semantic registration for a semi-autonomous surgical robotic system has been successfully applied. This proves that it is possible to reconstruct an anatomical environment in an accurate way and to manage the dynamic parts separately from the static environment. The experimental setup provided by the ARS and SARAS projects allowed the entire system to be tested.

8.1 Future works

In the future works, we will focus on upgrading our SLAM algorithm, mainly in two areas:

- Biomechanical properties to handle tissue deformation;
- Instrument marker-less tracking for an automatic hand-eye calibration.

8.1.1 Biomechanical modeling

While rigid registration algorithms allow computing the pose of internal organ structures based on surface information, this sparse sensor information is often

insufficient for compensating soft-tissue deformation inside the organ. In the context of the sparse data exploration problem [139], accurate non-rigid registration can be solved by incorporating a priori knowledge about the mechanical properties of the tissue via biomechanical modeling. Using elasticity theory, the approach can be formulated as a boundary value problem with displacement boundary conditions generated from intra-operative sensor data. In general, the finite element method (FEM) is used to solve the resulting set of partial differential equations. In several neurosurgical applications, this approach has been successfully applied to compensate the brain shift with intra-operative images [19, 115, 179, 208]. In contrast to neurosurgery, there are only a few studies on abdominal or laparoscopic interventions that adapt this concept to date [16, 17, 41, 114, 144, 150, 178, 185].

Using biomechanical models for non-rigid registration is challenging as finite element (FE) models are computationally intensive, but have to be solved in real time for computer assisted surgery (CAS) while still being robust and accurate. The application of fast, GPU-based FE solvers in combination with a reduced model complexity is therefore crucial regarding real-time capability. Various FE algorithms exist which can be used for hyper-, visco-, and poroelastic models in the field of real-time soft tissue simulation [109, 115]. Both methods have drawbacks regarding robustness and numerical complexity, especially in the context of an intra-operative application. Since previous studies have shown that in this context the material law and its parameterization has very little impact on the registration accuracy as long as a geometrically nonlinear model is used [115, 208], more efficient models, e.g. the corotated FE [113, 185], can be used, also taking vascular structures inside the organ into account [144]. Another aspect that has to be considered are morphological changes due to cuts which have to be propagated in real time on the FE mesh. A promising and efficient method for real-time cut simulation is e.g. the extended finite element method (X-FEM). Several approaches based on X-FEM can be found in the literature [79, 197].

8.1.2 Instrument hand-eye calibration

In chapter 3 we have emphasized the importance of calibration in the RMIS procedure and how very low the calibration error needs to be in order to increase the accuracy in the surgical intervention. So we considered introducing an automatic calibration procedure using the surgical instruments by exploiting the robot kinematics. we have already evaluated methods of segmentation of the tools based on standard approaches and on semantic information, but the estimation of the pose given the images has not yet been addressed in this thesis.

An efficient method, for tools belonging to the category of robotic devices, is to use a robot renderer with a CAD model in order to generate tool templates according to specific kinematic joint configurations [157]. This is stated to be desirable because collecting training data becomes easier than if it had to come from videos, thus enabling larger collection with less effort. Advantages of this type of data generation have been shown successfully in [175]. However,

choosing appropriate object parts to model can also prove challenging, particularly due to occlusion from other instruments, tissue and from the field of view. For surgical tools, modelling the tip region is the most viable tactic as it is the most characteristic landmark for tool differentiation and is the most likely component to be in view, relative to the tool end or tool body. However, tool tips can be cumbersome to model when made of many parts, which is the case for articulated surgical instruments

References

[1] T. K. Adebar, A. E. Fletcher, and A. M. Okamura. "3-D ultrasound-guided robotic needle steering in biological tissue". In: *IEEE Transactions on Biomedical Engineering* 61.12 (2014), pp. 2899–2910.

[2] R. Alami, R. Chatila, S. Fleury, M. Ghallab, and F. Ingrand. "An architecture for autonomy". In: *The International Journal of Robotics Research* 17.4 (1998), pp. 315–337.

[3] M. Alsheakhali, M. Yigitsoy, A. Eslami, and N. Navab. "Surgical tool detection and tracking in retinal microsurgery". In: *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 9415. International Society for Optics and Photonics. 2015, p. 941511.

[4] M. Ambai and Y. Yoshida. "CARD: Compact and real-time descriptors". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 97–104.

[5] J. Aulinas, Y. R. Petillot, J. Salvi, and X. Lladó. "The slam problem: a survey." In: *CCIA* 184.1 (2008), pp. 363–371.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[7] H. Bay, T. Tuytelaars, and L. Van Gool. "Surf: Speeded up robust features". In: *European conference on computer vision*. Springer. 2006, pp. 404–417.

[8] B. Bescos, J. M. Fácil, J. Civera, and J. Neira. "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4076–4083.

[9] P. J. Besl and N. D. McKay. "Method for registration of 3-D shapes". In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. International Society for Optics and Photonics. 1992, pp. 586–606.

[10] M. Bonfè, F. Boriero, R. Dodi, P. Fiorini, A. Morandi, R. Muradore, L. Pasquale, A. Sanna, and C. Secchi. "Towards automated surgical robotics: A requirements engineering approach". In: *IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. 2012.

[11] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin. "Detecting surgical tools by modelling local appearance and global shape". In: *IEEE transactions on medical imaging* 34.12 (2015), pp. 2603–2617.

[12] S. Bouraine, T. Fraichard, and H. Salhi. "Provably safe navigation for mobile robots with limited field-of-views in dynamic environments". In: *Autonomous Robots* 32.3 (2012), pp. 267–283.

[13] D. Burschka, M. Li, M. Ishii, R. H. Taylor, and G. D. Hager. "Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery". In: *Medical Image Analysis* 9.5 (2005), pp. 413–426.

[14] B. F. Buxton and H. Buxton. "Computation of optic flow from the motion of edge features in image sequences". In: *Image and Vision Computing* 2.2 (1984), pp. 59–75.

[15] Z. Cai, J. Han, L. Liu, and L. Shao. "RGB-D datasets using microsoft kinect or similar sensors: a survey". In: *Multimedia Tools and Applications* 76 (Feb. 2017).

[16] D. M. Cash, M. I. Miga, S. C. Glasgow, B. M. Dawant, L. W. Clements, Z. Cao, R. L. Galloway, and W. C. Chapman. "Concepts and preliminary data toward the realization of image-guided liver surgery". In: *Journal of Gastrointestinal Surgery* 11.7 (2007), pp. 844–859.

[17] D. Cash, M. Miga, T. Sinha, R. Galloway, and W. Chapman. "Compensating for intraoperative soft-tissue deformations using incomplete surface data and finite elements". In: *IEEE Transactions on Medical Imaging* 24.11 (2005), pp. 1479–1491.

[18] H. H. Chen. "A screw motion approach to uniqueness analysis of head-eye geometry". In: *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society. 1991, pp. 145–146.

[19] I. Chen, A. M. Coffey, S. Ding, P. Dumpuri, B. M. Dawant, R. C. Thompson, and M. I. Miga. "Intraoperative Brain Shift Compensation: Accounting for Dural Septa". In: *IEEE Transactions on Biomedical Engineering* 58.3 (2011), pp. 499–508.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[21] Y. Chen and G. Medioni. "Object modeling by registration of multiple range images". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 1991.

[22] J. C. Chou and M Kamel. "Finding the position and orientation of a sensor on a robot manipulator using quaternions". In: *The international journal of robotics research* 10.3 (1991), pp. 240–254.

[23] T. Collins and A. Bartoli. "Towards live monocular 3D laparoscopy using shading and specularity information". In: *International Conference on Information Processing in Computer-Assisted Interventions (MICCAI)*. Springer. 2012.

[24] N. Conen and T. Luhmann. "Overview of photogrammetric measurement techniques in minimally invasive surgery using endoscopes". In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2017), p. 33.

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[26] L. Cui and C. Ma. "SOF-SLAM: A semantic visual SLAM for dynamic environments". In: *IEEE access* 7 (2019), pp. 166528–166539.

[27] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[28] K. Daniilidis. "Hand-eye calibration using dual quaternions". In: *The International Journal of Robotics Research* 18.3 (1999), pp. 286–298.

[29] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. "MonoSLAM: Real-time single camera SLAM". In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1052–1067.

[30] G. De Rossi, M. Minelli, S. Roin, F. Falezza, A. Sozzi, F. Ferraguti, F. Setti, M. Bonfè, C. Secchi, and R. Muradore. "A First Evaluation of a Multi-Modal Learning System to Control Surgical Assistant Robots via Action Segmentation". In: *IEEE Transactions on Medical Robotics and Bionics* (2021), pp. 1–11. ISSN: 2576-3202.

[31] G. De Rossi, M. Minelli, A. Sozzi, N. Piccinelli, F. Ferraguti, F. Setti, M. Bonfè, C. Secchi, and R. Muradore. "Cognitive Robotic Architecture for Semi-Autonomous Execution of Manipulation Tasks in a Surgical Environment". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019.

[32] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. "Formal verification of ethical choices in autonomous systems". In: *Robotics and Autonomous Systems* 77 (2016), pp. 1–14.

[33] L. A. Dennis, M. Fisher, and A. Winfield. "Towards verifiably ethical robot behaviour". In: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[34] A. Desai, T. Dreossi, and S. A. Seshia. "Combining model checking and runtime verification for safe robotics". In: *International Conference on Runtime Verification*. Springer. 2017, pp. 172–189.

[35] A. Diodato, M. Brancadoro, G. De Rossi, H. Abidi, D. Dall'Alba, R. Muradore, G. Ciuti, P. Fiorini, A. Menciassi, and M. Cianchetti. "Soft Robotic Manipulator for Improving Dexterity in Minimally Invasive Surgery". In: *Surgical Innovation* (2018).

[36] C. Dixon, M. Webster, J. Saunders, M. Fisher, and K. Dautenhahn. ""The fridge door is open"–Temporal Verification of a Robotic Assistant's Behaviours". In: *Conference Towards Autonomous Robotic Systems*. Springer. 2014, pp. 97–108.

[37] Z. Dogmus, G. Gezici, V. Patoglu, and E. Erdem. "Developing and Maintaining an Ontology for Rehabilitation Robotics." In: *KEOD*. 2012, pp. 389–395.

[38] C. Doignon, P. Graebling, and M. De Mathelin. "Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature". In: *Real-Time Imaging* 11.5-6 (2005), pp. 429–442.

[39] C. Doignon, F. Nageotte, and M. De Mathelin. "Detection of grey regions in color images: application to the segmentation of a surgical instrument in robotized laparoscopy". In: *2004 IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). Vol. 4. IEEE. 2004, pp. 3394–3399.

[40] X. Du, M. Allan, A. Dore, S. Ourselin, D. Hawkes, J. D. Kelly, and D. Stoyanov. "Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery". In: *International journal of computer assisted radiology and surgery* 11.6 (2016), pp. 1109–1119.

[41] P. Dumpuri, L. W. Clements, B. M. Dawant, and M. I. Miga. "Model-updated image-guided liver surgery: preliminary results using surface characterization". In: *Progress in biophysics and molecular biology* 103.2-3 (2010), pp. 197–207.

[42] H. Durrant-Whyte and T. Bailey. "Simultaneous localization and mapping: part I". In: *IEEE robotics & automation magazine* 13.2 (2006), pp. 99–110.

[43] R. Elfring, M. de la Fuente, and K. Radermacher. "Assessment of optical localizer accuracy for computer aided surgery systems". In: *Computer Aided Surgery* 15.1-3 (2010), pp. 1–12.

[44] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. "An evaluation of the RGB-D SLAM system". In: *2012 IEEE international conference on robotics and automation*. IEEE. 2012, pp. 1691–1696.

[45] J. Engel, T. Schöps, and D. Cremers. "LSD-SLAM: Large-scale direct monocular SLAM". In: *European conference on computer vision*. Springer. 2014, pp. 834–849.

[46] Y. Fan, H. Han, Y. Tang, and T. Zhi. "Dynamic objects elimination in SLAM based on image fusion". In: *Pattern Recognition Letters* 127 (2019), pp. 191–201.

[47] L. E. Fernandes, V. Custodio, G. V. Alves, and M. Fisher. "A rational agent controlling an autonomous vehicle: Implementation and formal verification". In: *arXiv preprint arXiv:1709.02557* (2017).

[48] F. Ferraguti, N. Preda, G. De Rossi, M. Bonfè, R. Muradore, P. Fiorini, and C. Secchi. "A two-layer approach for shared control in semi-autonomous robotic surgery". In: *2015 European Control Conference (ECC)*. IEEE. 2015, pp. 747–752.

[49] F. Ferraguti, N. Preda, A. Manurung, M. Bonfe, O. Lambercy, R. Gassert, R. Muradore, P. Fiorini, and C. Secchi. "An energy tank-based interactive control architecture for autonomous and teleoperated robotic surgery". In: *IEEE Transactions on Robotics* 31.5 (2015), pp. 1073–1088.

[50] P. Fiorini, D. Dall'Alba, M. Ginesi, B. Maris, D. Meli, H. Nakawala, and A. Roberti. "Challenges of Autonomous Robotic Surgery". In: *Hamlyn Symposium on Medical Robotics (HSMR)*. 2019.

[51] M. A. Fischler and R. C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[52] M. Fisher, L. Dennis, and M. Webster. "Verifying autonomous systems". In: *Communications of the ACM* 56.9 (2013), pp. 84–93.

[53] M. P. Fried, J. Kleefield, H. Gopal, E. Reardon, B. T. Ho, and F. A. Kuhn. "Image-guided endoscopic surgery: results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system". In: *The Laryngoscope* 107.5 (1997), pp. 594–601.

[54] S. Fuchs. "Calibration and multipath mitigation for increased accuracy of time-of-flight camera measurements in robotic applications". PhD thesis. Universitätsbibliothek der Technischen Universität Berlin, 2012.

[55] R. Furukawa, Y. Sanomura, S. Tanaka, S. Yoshida, R. Sagawa, M. Visentini-Scarzanella, and H. Kawasaki. "3D endoscope system using DOE projector". In: *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016.

[56] P. Gainer, C. Dixon, K. Dautenhahn, M. Fisher, U. Hustadt, J. Saunders, and M. Webster. "CRutoN: Automatic verification of a robotic assistant's behaviours". In: *Critical Systems: Formal Methods and Automated Verification*. Springer, 2017, pp. 119–133.

[57] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. "Variable baseline/resolution stereo". In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[58] W. Gander and J. Hrebicek. *Solving problems in scientific computing using Maple and Matlab*. Springer Science & Business Media, 2011.

[59] M. Ginesi, D. Meli, H. Nakawala, A. Roberti, and P. Fiorini. "A knowledge-based framework for task automation in surgery". In: *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE. 2019, pp. 37–42.

[60] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini. "Autonomous task planning and situation awareness in robotic surgery*". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 3144–3150.

[61] O. G. Grasa, J. Civera, and J. M. M. Montiel. "EKF monocular SLAM with relocalization for laparoscopic sequences". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2011.

[62] G. Grisetti, R. Kümmerle, H. Strasdat, and K. Konolige. "g2o: A general framework for (hyper) graph optimization". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2011, pp. 9–13.

[63] J. Guiochet, M. Machin, and H. Waeselynck. "Safety-critical advanced robots: A survey". In: *Robotics and Autonomous Systems* 94 (2017), pp. 43–52.

[64] S. Haase, J. Wasza, T. Kilgus, and J. Hornegger. "Laparoscopic instrument localization using a 3-d time-of-flight/rgb endoscope". In: *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE. 2013, pp. 449–454.

[65] T. Haidegger, P Kazanzides, I. Rudas, B Benyó, and Z. Benyó. "The importance of accuracy measurement standards for computer-integrated

interventional systems". In: *EURON GEM Sig Workshop on The Role of Experiments in Robotics Research at IEEE ICRA*. 2010, pp. 1–6.

[66] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[67] R. I. Hartley. "In defense of the eight-point algorithm". In: *IEEE Transactions on pattern analysis and machine intelligence* 19.6 (1997), pp. 580–593.

[68] R. I. Hartley. "Lines and points in three views and the trifocal tensor". In: *International Journal of Computer Vision* 22.2 (1997), pp. 125–140.

[69] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[70] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[71] D. Heß, M. Althoff, and T. Sattel. "Formal verification of maneuver automata for parameterized motion primitives". In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, pp. 1474–1481.

[72] R. Hoffmann, M. Ireland, A. Miller, G. Norman, and S. Veres. "Autonomous agent behaviour modelled in PRISM–A case study". In: *International Symposium on Model Checking Software*. Springer. 2016, pp. 104–110.

[73] B. K. Horn and B. G. Schunck. "Determining optical flow". In: *Artificial intelligence* 17.1-3 (1981), pp. 185–203.

[74] M. Hu, G. Penney, P. Edwards, M. Figl, and D. J. Hawkes. "3D reconstruction of internal organ surfaces for minimal invasive surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2007, pp. 68–77.

[75] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes. "Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes". In: *Medical image analysis* 16.3 (2012), pp. 597–611.

[76] Y. Hu, H. U. Ahmed, C. Allen, D. Pendsé, M. Sahu, M. Emberton, D. Hawkes, and D. Barratt. "MR to ultrasound image registration for guiding prostate biopsy and interventions". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2009, pp. 787–794.

[77] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. 2017.

[78] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera".

In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 2011, pp. 559–568.

[79] L. Jeřábková and T. Kuhlen. "Stable cutting of deformable objects in virtual environments using xfem". In: *IEEE computer graphics and applications* 29.2 (2009), pp. 61–71.

[80] S. Kahn, D. Haumann, and V. Willert. "Hand-eye calibration with a depth camera: 2D or 3D?" In: *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*. Vol. 3. IEEE. 2014, pp. 481–489.

[81] M. Kazhdan, M. Bolitho, and H. Hoppe. "Poisson Surface Reconstruction". In: *Eurographics Symposium on Geometry Processing (SGP)*. 2006.

[82] D. W. Kim and J. E. Ha. "Hand/eye calibration using 3d-3d correspondences". In: *Applied Mechanics and Materials*. Vol. 319. Trans Tech Publ. 2013, pp. 532–535.

[83] G. Klein and D. Murray. "Parallel tracking and mapping for small AR workspaces". In: *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society. 2007, pp. 1–10.

[84] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, H. Feußner, and J. Hornegger. "ToF meets RGB: novel multi-sensor super-resolution for hybrid 3-D endoscopy". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2013.

[85] S. Konur, C. Dixon, and M. Fisher. "Formal verification of probabilistic swarm behaviours". In: *International Conference on Swarm Intelligence*. Springer. 2010, pp. 440–447.

[86] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas. "Where's waldo? sensor-based temporal logic motion planning". In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE. 2007, pp. 3116–3121.

[87] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[88] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. "g 2 o: A general framework for graph optimization". In: *2011 IEEE International Conference on Robotics and Automation*. IEEE. 2011, pp. 3607–3613.

[89] D. Kundrat, R. Graesslin, A. Schoob, D. T. Friedrich, M. O. Scheithauer, T. K. Hoffmann, T. Ortmaier, L. A. Kahrs, and P. J. Schuler. "Preclinical Performance Evaluation of a Robotic Endoscope for Non-Contact Laser Surgery". In: *Annals of Biomedical Engineering* 49.2 (2021), pp. 585–600. ISSN: 0090-6964.

[90] A. Kundu, K. M. Krishna, and J. Sivaswamy. "Moving object detection by multi-view geometric techniques from a single camera mounted

robot". In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, pp. 4306–4312.

[91] D. M. Kwartowitz, S. D. Herrell, and R. L. Galloway. "Toward image-guided robotic surgery: determining intrinsic accuracy of the da Vinci robot". In: *International Journal of Computer Assisted Radiology and Surgery* 1.3 (2006), pp. 157–165.

[92] V. Lahanas, C. Loukas, and E. Georgiou. "A simple sensor calibration technique for estimating the 3D pose of endoscopic instruments". In: *Surgical endoscopy* 30.3 (2016), pp. 1198–1204.

[93] R. Lange. "3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology". In: (2000).

[94] H. N. D. Le, J. D. Opfermann, M. Kam, S. Raghunathan, H. Saeidi, S. Leonard, J. U. Kang, and A. Krieger. "Semi-autonomous laparoscopic robotic electro-surgery with a novel 3D endoscope". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018.

[95] B. Lee and D. D. Lee. "Learning anisotropic ICP (LA-ICP) for robust and efficient 3D registration". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016.

[96] C. Lee, Y.-F. Wang, D. R. Uecker, and Y. Wang. "Image analysis for automated tracking in robot-assisted endoscopic surgery". In: *Proceedings of 12th International Conference on Pattern Recognition*. Vol. 1. IEEE. 1994, pp. 88–92.

[97] M. Li and D. Betsis. "Head-eye calibration". In: *Proceedings of IEEE International Conference on Computer Vision*. IEEE. 1995, pp. 40–45.

[98] R.-h. Liang and J.-f. Mao. "Hand-eye calibration with a new linear decomposition algorithm". In: *Journal of Zhejiang University-SCIENCE A* 9.10 (2008), pp. 1363–1368.

[99] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[100] J. Long, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[101] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[102] H. C. Longuet-Higgins. "The reconstruction of a plane surface from two perspective projections". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 227.1249 (1986), pp. 399–410.

[103] Y. K. Lopes, S. M. Trenkwalder, A. B. Leal, T. J. Dodd, and R. Groß. "Supervisory control theory applied to swarm robotics". In: *Swarm Intelligence* 10.1 (2016), pp. 65–97.

[104] D. G. Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[105] B. D. Lucas, T. Kanade, et al. "An iterative image registration technique with an application to stereo vision". In: Vancouver. 1981.

[106] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel. "ORBSLAM-Based Endoscope Tracking and 3D Reconstruction". In: *Computer-Assisted and Robotic Endoscopy*. Ed. by T. Peters, G.-Z. Yang, N. Navab, K. Mori, X. Luo, T. Reichl, and J. McLeod. Cham: Springer International Publishing, 2017, pp. 72–83. ISBN: 978-3-319-54057-3.

[107] N. Mahmoud, A. Hostettler, T. Collins, L. Soler, C. Doignon, and J. Montiel. "SLAM based quasi dense reconstruction for minimally invasive surgery scenes". In: *arXiv preprint arXiv:1705.09107* (2017).

[108] J. Malik, S. Belongie, T. Leung, and J. Shi. "Contour and texture analysis for image segmentation". In: *International journal of computer vision* 43.1 (2001), pp. 7–27.

[109] S. Marchesseau, T. Heimann, S. Chatelin, R. Willinger, and H. Delingette. "Fast porous visco-hyperelastic soft tissue model for surgery simulation: application to liver surgery". In: *Progress in biophysics and molecular biology* 103.2-3 (2010), pp. 185–196.

[110] S. Matthias, M. Kästner, and E. Reithmeier. "A 3D measuring endoscope for hand-guided operation". In: *Measurement Science and Technology* 29.9 (2018), p. 094001.

[111] S. McKenna, H. N. Charif, and T Frank. "Towards video understanding of laparoscopic surgery: Instrument tracking". In: *Proc. of Image and Vision Computing, New Zealand*. 2005.

[112] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Vol. 544. John Wiley & Sons, 2004.

[113] J. Mezger, B. Thomaszewski, S. Pabst, and W. Straßer. "Interactive physically-based shape editing". In: *Computer Aided Geometric Design* 26.6 (2009), pp. 680–694.

[114] M. I. Miga, P. Dumpuri, A. L. Simpson, J. A. Weis, and W. R. Jarnagin. "The sparse data extrapolation problem: strategies for soft-tissue correction for image-guided liver surgery". In: *Medical Imaging 2011: Visualization, Image-Guided Procedures, and Modeling*. Vol. 7964. International Society for Optics and Photonics. 2011, p. 79640C.

[115] K. Miller, G. Joldes, D. Lance, and A. Wittek. "Total Lagrangian explicit dynamics finite element algorithm for computing soft tissue deformation". In: *Communications in numerical methods in engineering* 23.2 (2007), pp. 121–134.

[116] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE. 2016, pp. 565–571.

[117] D. J. Mirota, M. Ishii, and G. D. Hager. "Vision-based navigation in image-guided interventions". In: *Annual review of biomedical engineering* 13 (2011), pp. 297–319.

[118] S. Mitsch, K. Ghorbal, and A. Platzer. "On provably safe obstacle avoidance for autonomous robotic ground vehicles". In: *Robotics: Science and Systems IX, Technische Universität Berlin, Berlin, Germany, June 24-June 28, 2013*. 2013.

[119] S. Mitsch, K. Ghorbal, D. Vogelbacher, and A. Platzer. "Formal verification of obstacle avoidance and navigation of ground robots". In: *The International Journal of Robotics Research* 36.12 (2017), pp. 1312–1340.

[120] S. Moarref and H. Kress-Gazit. "Decentralized control of robotic swarms from high-level temporal logic specifications". In: *2017 international symposium on multi-robot and multi-agent systems (MRS)*. IEEE. 2017, pp. 17–23.

[121] J. Morse, D. Araiza-Illan, J. Lawry, A. Richards, and K. Eder. "Formal specification and analysis of autonomous systems under partial compliance". In: *arXiv preprint arXiv:1603.01082* (2016).

[122] P. Mountney, S. Giannarou, D. Elson, and G.-Z. Yang. "Optical biopsy mapping for minimally invasive cancer screening". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2009.

[123] P. Mountney, D. Stoyanov, A. Davison, and G.-Z. Yang. "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2006.

[124] P. Mountney and G.-Z. Yang. "Motion compensated SLAM for image guided surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2010.

[125] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System". In: *IEEE Transactions on Robotics* 31.5 (2015), pp. 1147–1163. ISSN: 1552-3098.

[126] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163.

[127] R. Mur-Artal and J. D. Tardós. "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras". In: *IEEE Transactions on Robotics* 33.5 (2017), pp. 1255–1262.

[128] R. Muradore, P. Fiorini, G. Akgun, D. E. Barkana, M. Bonfe, F. Boriero, A. Caprara, G. De Rossi, R. Dodi, O. J. Elle, et al. "Development of a cognitive robotic system for simple surgical tasks". In: *International Journal of Advanced Robotic Systems* 12.4 (2015), p. 37.

[129] R. Muradore, P. Fiorini, G. Akgun, D. E. Barkana, M. Bonfe, F. Boriero, A. Caprara, G. D. Rossi, R. Dodi, O. J. Elle, F. Ferraguti, L. Gasperotti, R. Gassert, K. Mathiassen, D. Handini, O. Lambercy, L. Li, M. Kruusmaa, A. O. Manurung, G. Meruzzi, H. Q. P. Nguyen, N. Preda, G. Riolfo, A. Ristolainen, A. Sanna, C. Secchi, M. Torsello, and A. E. Yantac. "Development of a Cognitive Robotic System for Simple Surgical Tasks". In: *International Journal of Advanced Robotic Systems* 12.4 (2015), p. 37.

[130] M. Najafi, N. Afsham, P. Abolmaesumi, and R. Rohling. "A closed-form differential formulation for ultrasound spatial calibration: multi-wedge phantom". In: *Ultrasound in medicine & biology* 40.9 (2014), pp. 2231–2243.

[131] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder. "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes". In: *International Conference on Computer Vision (ICCV)*. 2017.

[132] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. "DTAM: Dense tracking and mapping in real-time". In: *2011 international conference on computer vision*. IEEE. 2011, pp. 2320–2327.

[133] D. Nistér. "An efficient solution to the five-point relative pose problem". In: *IEEE transactions on pattern analysis and machine intelligence* 26.6 (2004), pp. 756–770.

[134] D. Nistér and H. Stewénius. "A minimal solution to the generalised 3-point pose problem". In: *Journal of Mathematical Imaging and Vision* 27.1 (2007), pp. 67–79.

[135] D. P. Noonan, P. Mountney, D. S. Elson, A. Darzi, and G.-Z. Yang. "A stereoscopic fibroscope for camera motion and 3D depth recovery during minimally invasive surgery". In: *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, pp. 4463–4468.

[136] T. Ojala, M. Pietikainen, and T. Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), pp. 971–987.

[137] E. Oleari, A. Leporini, D. Trojaniello, A. Sanna, U. Capitanio, F. Dehó, A. Larcher, F. Montorsi, A. Salonia, F. Setti, and R. Muradore. "Enhancing Surgical Process Modeling for Artificial Intelligence development in robotics: the SARAS case study for Minimally Invasive Procedures". In: *International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019.

[138] E. Oleari, A. Leporini, D. Trojaniello, A. Sanna, U. Capitanio, F. Dehó, A. Larcher, F. Montorsi, A. Salonia, F. Setti, et al. "Enhancing surgical process modeling for artificial intelligence development in robotics: the saras case study for minimally invasive procedures". In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. IEEE. 2019, pp. 1–6.

[139] R. E. Ong, J. J. Ou, and M. I. Miga. "Non-rigid registration of breast surfaces using the laplace and diffusion equations". In: *Biomedical engineering online* 9.1 (2010), pp. 1–14.

[140] K. Pachtrachai, M. Allan, V. Pawar, S. Hailes, and D. Stoyanov. "Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 2485–2491.

[141] J. H. Palep. "Robotic assisted minimally invasive surgery". In: *Journal of minimal access surgery* 5.1 (2009), p. 1.

[142] H. Pan, N. L. Wang, and Y. S. Qin. "A closed-form solution to eye-to-hand calibration towards visual grasping". In: *Industrial Robot: An International Journal* (2014).

[143] V. R. Patel, K. K. Shah, R. K. Thaly, and H. Lavery. "Robotic-assisted laparoscopic radical prostatectomy: the Ohio State University technique". In: *Journal of robotic surgery* 1.1 (2007), pp. 51–59.

[144] I. Peterlík, C. Duriez, and S. Cotin. "Modeling and real-time simulation of a vascularized liver tissue". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 50–57.

[145] Z. Pezzementi, S. Voros, and G. D. Hager. "Articulated object tracking by rendering consistent appearance parts". In: *2009 IEEE International Conference on Robotics and Automation*. 2009, pp. 3940–3947.

[146] D. Phan, J. Yang, R. Grosu, S. A. Smolka, and S. D. Stoller. "Collision avoidance for mobile robots with limited sensing and limited information about moving obstacles". In: *Formal Methods in System Design* 51.1 (2017), pp. 62–86.

[147] D. Phan, J. Yang, D. Ratasich, R. Grosu, S. A. Smolka, and S. D. Stoller. "Collision avoidance for mobile robots with limited sensing and limited information about the environment". In: *Runtime Verification*. Springer. 2015, pp. 201–215.

[148] N Piccinelli, A Roberti, E Tagliabue, F Setti, G Kronreif, R Muradore, and P Fiorini. "Rigid 3D Registration of Pre-operative Information for Semi-Autonomous Surgery". In: *2020 International Symposium on Medical Robotics (2020) - Atlanta (USA)*. Ed. by I. S. on Medical Robotics. International Symposium on Medical Robotics. Apr. 22, 2020.

[149] N. Piccinelli, A. Roberti, E. Tagliabue, F. Setti, G. Kronreif, R. Muradore, and P. Fiorini. "Rigid 3D Registration of Pre-operative Information for Semi-Autonomous Surgery". In: *International Symposium on Medical Robotics ISMR 2020, Atlanta, USA, 18–20 Nov.* 2020, pp. 1–6.

[150] P. Pratt, D. Stoyanov, M. Visentini-Scarzanella, and G.-Z. Yang. "Dynamic guidance for robotic surgery using image-constrained biomechanical models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 77–85.

[151] N. Preda, F. Ferraguti, G. De Rossi, C. Secchi, R. Muradore, P. Fiorini, and M. Bonfè. "A cognitive robot control architecture for autonomous execution of surgical tasks". In: *Journal of Medical Robotics Research* 1.04 (2016), p. 1650008.

[152] N. Preda, A. Manurung, O. Lambercy, R. Gassert, and M. Bonfè. "Motion planning for a multi-arm surgical robot using both sampling-based algorithms and motion primitives". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 1422–1427.

[153] M. Proetzsch, K. Berns, T. Schuele, and K. Schneider. "Formal verification of safety behaviours of the outdoor robot ravon." In: *ICINCO-RA (1)*. 2007, pp. 157–164.

[154] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, et al. "ROS: an open-source Robot Operating System". In: *ICRA workshop on open source software*. Vol. 3. 3.2. Kobe, Japan. 2009, p. 5.

[155] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[156] K. Reinhard, S. Karsten, and K. Andreas. "Computer vision—three dimensional data from images". In: *Springer-Verlag* (1998).

[157] A. Reiter and P. K. Allen. "An online learning approach to in-vivo tracking using synergistic features". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 3441–3446.

[158] A. Reiter, P. K. Allen, and T. Zhao. "Feature classification for tracking articulated surgical tools". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 592–600.

[159] A. Reiter, P. K. Allen, and T. Zhao. "Marker-less articulated surgical tool detection". In: *Computer assisted radiology and surgery*. 2012.

[160] M. I. Ribeiro. "Kalman and extended kalman filters: Concept, derivation and properties". In: *Institute for Systems and Robotics* 43 (2004), p. 46.

[161] A. Roberti, N. Piccinelli, D. Meli, R. Muradore, and P. Fiorini. "Improving Rigid 3-D Calibration for Robotic Surgery". In: *IEEE Transactions on Medical Robotics and Bionics* 2.4 (2020), pp. 569–573.

[162] E. Rohmer, S. P. Singh, and M. Freese. "V-REP: A versatile and scalable robot simulation framework". In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 1321–1326.

[163] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[164] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[165] R. B. Rusu, N. Blodow, and M. Beetz. "Fast point feature histograms (FPFH) for 3D registration". In: *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, pp. 3212–3217.

[166] Y. Sahillioğlu and L. Kavan. "Skuller: a volumetric shape registration algorithm for modeling skull deformities". In: *Medical image analysis* 23.1 (2015), pp. 15–27.

[167] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. "Slam++: Simultaneous localisation and mapping at the

level of objects". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 1352–1359.

[168] N. Sayols, A. Sozzi, N. Piccinelli, A. Hernansanz, A. Casals, M. Bonfè, and R. Muradore. "Global/local motion planning based on Dynamic Trajectory Reconfiguration and Dynamical Systems for autonomous surgical robots". In: *IEEE International Conference on Robotics and Automation (ICRA)*. To Appear. 2020.

[169] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou. "An endoscopic 3D scanner based on structured light". In: *Medical Image Analysis* 16.5 (2012), pp. 1063–1072.

[170] D. Seto, B. Krogh, L. Sha, and A. Chutinan. "The Simplex architecture for safe online control system upgrades". In: *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No. 98CH36207)*. Vol. 6. IEEE. 1998, pp. 3504–3508.

[171] F. Setti, E. Oleari, A. Leporini, D. Trojaniello, A. Sanna, U. Capitanio, F. Montorsi, A. Salonia, and R. Muradore. "A multirobots teleoperated platform for artificial intelligence training data collection in minimally invasive surgery". In: *International Symposium on Medical Robotics (ISMR)*. IEEE. 2019.

[172] M. Shah, R. D. Eastman, and T. Hong. "An overview of robot-sensor calibration methods for evaluation of perception systems". In: *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*. 2012, pp. 15–20.

[173] J. Shi et al. "Good features to track". In: *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE. 1994, pp. 593–600.

[174] Y. C. Shiu and S. Ahmad. "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX= XB." In: *IEEE Transactions on robotics and automation* 5.1 (1989), pp. 16–29.

[175] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. "Real-time human pose recognition in parts from single depth images". In: *CVPR 2011*. Ieee. 2011, pp. 1297–1304.

[176] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[177] T. Simpfendorfer, M. Baumhauer, M. Muller, C. N. Gutt, H.-P. Meinzer, J. J. Rassweiler, S. Guven, and D. Teber. "Augmented reality visualization during laparoscopic radical prostatectomy". In: *Journal of endourology* 25.12 (2011), pp. 1841–1845.

[178] A. L. Simpson, P. Dumpuri, W. R. Jarnagin, and M. I. Miga. "Model-assisted image-guided liver surgery using sparse intraoperative data". In: *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*. Springer, 2012, pp. 7–40.

[179] O. Škrinjar, C. Studholme, A. Nabavi, and J. Duncan. "Steps toward a stereo-camera-guided biomechanical model for brain shift compensa-

tion". In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 2001, pp. 183–189.

[180] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake. "MIS-SLAM: Real-Time Large-Scale Dense Deformable SLAM System in Minimal Invasive Surgery Based on Heterogeneous Computing". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4068–4075.

[181] T. Sotiropoulos, H. Waeselynck, J. Guiochet, and F. Ingrand. "Can robot navigation bugs be found in simulation? an exploratory study". In: *2017 IEEE International conference on software quality, reliability and security (QRS)*. IEEE. 2017, pp. 150–159.

[182] S. Speidel, J. Benzko, S. Krappe, G. Sudra, P. Azad, B. P. Müller-Stich, C. Gutt, and R. Dillmann. "Automatic classification of minimally invasive instruments based on endoscopic image sequences". In: *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*. Vol. 7261. International Society for Optics and Photonics. 2009, 72610A.

[183] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. "A benchmark for the evaluation of RGB-D SLAM systems". In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 573–580.

[184] Y. Sun, M. Liu, and M. Q.-H. Meng. "Motion removal for reliable RGB-D SLAM in dynamic environments". In: *Robotics and Autonomous Systems* 108 (2018), pp. 115–128.

[185] S. Suwelack, H. Talbot, S. Röhl, R. Dillmann, and S. Speidel. "A biomechanical liver model for intraoperative soft tissue registration". In: *Medical Imaging 2011: Visualization, Image-Guided Procedures, and Modeling*. Vol. 7964. International Society for Optics and Photonics. 2011, p. 79642I.

[186] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[187] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua. "Data-driven visual tracking in retinal microsurgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 568–575.

[188] R. Sznitman, R. Richa, R. H. Taylor, B. Jedynak, and G. D. Hager. "Unified detection and tracking of instruments during retinal microsurgery". In: *IEEE transactions on pattern analysis and machine intelligence* 35.5 (2012), pp. 1263–1273.

[189] P. Tabuada. *Verification and control of hybrid systems: a symbolic approach*. Springer Science & Business Media, 2009.

[190] N. Taffinder, S. G. T. Smith, J. Huber, R. C. G. Russell, and A. Darzi. "The effect of a second-generation 3D endoscope on the laparoscopic precision of novices and experienced surgeons". In: *Surgical Endoscopy* 13.11 (1999), pp. 1087–1092. ISSN: 1432-2218.

[191] S. Thrun and J. J. Leonard. "Simultaneous localization and mapping". In: *Springer handbook of robotics* (2008), pp. 871–889.

[192] A. Tiwari. "Abstractions for hybrid systems". In: *Formal Methods in System Design* 32.1 (2008), pp. 57–83.

[193] J. Totz, K. Fujii, P. Mountney, and G.-Z. Yang. "Enhanced visualisation for minimally invasive surgery". In: *International Journal of Computer Assisted Radiology and Surgery* 7.3 (2012), pp. 423–432.

[194] J. Totz, P. Mountney, D. Stoyanov, and G.-Z. Yang. "Dense surface reconstruction for enhanced navigation in MIS". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2011.

[195] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. "Bundle adjustment—a modern synthesis". In: *International workshop on vision algorithms*. Springer. 1999, pp. 298–372.

[196] R. Y. Tsai, R. K. Lenz, et al. "A new technique for fully autonomous and efficient 3 D robotics hand/eye calibration". In: *IEEE Transactions on robotics and automation* 5.3 (1989), pp. 345–358.

[197] L. M. Vigneron, S. K. Warfield, P. A. Robe, and J. G. Verly. "3D XFEM-based modeling of retraction for preoperative image update". In: *Computer Aided Surgery* 16.3 (2011), pp. 121–134.

[198] S. Voros, J.-A. Long, and P. Cinquin. "Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders". In: *The International Journal of Robotics Research* 26.11-12 (2007), pp. 1173–1190.

[199] D. Walter, H. Täubig, and C. Lüth. "Experiences in applying formal verification in robotics". In: *International Conference on Computer Safety, Reliability, and Security*. Springer. 2010, pp. 347–360.

[200] R. Wang, W. Wan, Y. Wang, and K. Di. "A new RGB-D SLAM method with moving object detection for dynamic indoor scenes". In: *Remote Sensing* 11.10 (2019), p. 1143.

[201] Z. Wang, Q. Zhang, J. Li, S. Zhang, and J. Liu. "A computationally efficient semantic slam solution for dynamic scenes". In: *Remote Sensing* 11.11 (2019), p. 1363.

[202] Z. Wang, Z. Liu, Q. Ma, A. Cheng, Y.-h. Liu, S. Kim, A. Deguet, A. Reiter, P. Kazanzides, and R. H. Taylor. "Vision-based calibration of dual RCM-based robot arms in human-robot collaborative minimally invasive surgery". In: *IEEE Robotics and Automation Letters* 3.2 (2017), pp. 672–679.

[203] M. O. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition - 360 Degree Business*. 2nd. USA: A. K. Peters, Ltd., 2015. ISBN: 1482257378.

[204] M. Webster, N. Cameron, M. Fisher, and M. Jump. "Generating certification evidence for autonomous unmanned aircraft using model checking and simulation". In: *Journal of Aerospace Information Systems* 11.5 (2014), pp. 258–279.

[205] M. Webster, C. Dixon, M. Fisher, M. Salem, J. Saunders, K. L. Koay, and K. Dautenhahn. "Formal verification of an autonomous personal robotic assistant". In: *2014 AAAI Spring Symposium Series*. 2014.

[206] M. Webster, M. Fisher, N. Cameron, and M. Jump. "Formal methods for the certification of autonomous unmanned aircraft systems". In: *International Conference on Computer Safety, Reliability, and Security*. Springer. 2011, pp. 228–242.

[207] J. Weese, G. P. Penney, P. Desmedt, T. M. Buzug, D. L. Hill, and D. J. Hawkes. "Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery". In: *IEEE transactions on information technology in biomedicine* 1.4 (1997), pp. 284–293.

[208] A. Wittek, R. Kikinis, S. K. Warfield, and K. Miller. "Brain shift computation using a fully nonlinear biomechanical model". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2005, pp. 583–590.

[209] R. Wolf, J. Duchateau, P. Cinquin, and S. Voros. "3D tracking of laparoscopic instruments using statistical and geometric modeling". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 203–210.

[210] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou. "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment". In: *Robotics and Autonomous Systems* 117 (2019), pp. 1–16.

[211] S. Xie and Z. Tu. "Holistically-nested edge detection". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1395–1403.

[212] M. Xu. "3D Endoscope Based on the Controlled Aberration Method". MA thesis. The University of Arizona., 2018.

[213] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, et al. *Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy*. 2017.

[214] J. Yang, H. Li, and Y. Jia. "GO-ICP: Solving 3D registration efficiently and globally optimally". In: *IEEE International Conference on Computer Vision (ICCV)*. 2013.

[215] M. Yang, K. Kpalma, and J. Ronsin. *A survey of shape feature extraction techniques. 2008*. 2008.

[216] Z. Yaniv and K. Cleary. "Image-guided procedures: A review". In: *Computer Aided Interventions and Medical Robotics* 3 (2006), pp. 1–63.

[217] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei. "DS-SLAM: A semantic visual SLAM towards dynamic environments". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1168–1174.

[218] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability". In: *Neuroimage* 31.3 (2006), pp. 1116–1128.

[219] L. Zhang, L. Wei, P. Shen, W. Wei, G. Zhu, and J. Song. "Semantic SLAM based on object detection and improved octomap". In: *IEEE Access* 6 (2018), pp. 75545–75559.

[220] Z. Zhang. "A flexible new technique for camera calibration". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.11 (2000), pp. 1330–1334.

[221] F. Zhong, S. Wang, Z. Zhang, and Y. Wang. "Detect-SLAM: Making object detection and SLAM mutually beneficial". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 1001–1010.

[222] F. Zhong, Z. Wang, W. Chen, K. He, Y. Wang, and Y.-H. Liu. "Hand-Eye Calibration of Surgical Instrument for Robotic Surgery Using Interactive Manipulation". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1540–1547.

[223] Y. Zhong. "Intrinsic shape signatures: A shape descriptor for 3D object recognition". In: *IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2009.

[224] Y. Zhong. "Intrinsic shape signatures: A shape descriptor for 3d object recognition". In: *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE. 2009, pp. 689–696.

[225] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. "Scene Parsing through ADE20K Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[226] J. Zhou and S. Payandeh. "Visual tracking of laparoscopic instruments". In: *Journal of Automation and Control Engineering Vol* 2.3 (2014), pp. 234–241.

[227] D. Zou and P. Tan. "Coslam: Collaborative visual slam in dynamic environments". In: *IEEE transactions on pattern analysis and machine intelligence* 35.2 (2012), pp. 354–366.