TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2014

34

Óscar García Grasa

# Visual SLAM for Measurement and Augmented Reality in Laparoscopic Surgery

Departamento
Instituto de Investigación en Ingeniería [I3A]

Director/es
Martínez Montiel, José María



ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

Universidad Zaragoza
1542

Tesis Doctoral

# VISUAL SLAM FOR MEASUREMENT AND AUGMENTED REALITY IN LAPAROSCOPIC SURGERY


Autor

Óscar García Grasa

Director/es

Martínez Montiel, José María

**UNIVERSIDAD DE ZARAGOZA**

Instituto de Investigación en Ingeniería [I3A]


2014

Universidad de Zaragoza

# PHD THESIS

PhD on Biomedical Engineering

# Visual SLAM for Measurement and Augmented Reality in Laparoscopic Surgery.

**SLAM visual para Mediciones y Realidad Aumentada en Cirugía Laparoscópica.**

Óscar García Grasa

Thesis Advisor: José María Martínez Montiel

Instituto Universitario de Investigación de Ingeniería de Aragón
Universidad Zaragoza

Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza

January 2014

En memoria de mi tío Eduardo.

# Agradecimientos

Durante el desarrollo de esta tesis muchas personas han estado a mi lado ayundándome, apoyándome y animándome en los momentos más difíciles tanto personales como profesionales. Aquí va la lista de todos ellos. Espero no olvidarme de nadie.

José María Martínez Montiel. Confiaste en mi para dirigirme la tesis sin conocerme de nada. Al principio tu querías una tesis robótica, pero finalmente te redirigí al mundo clínico. Gracias por permitirlo y por aportar todos tus conocimientos y tu experiencia en este nuevo campo. Espero que esta nueva línea de investigación, que hemos abierto en el departamento, te sea muy fructífera.

Ernesto Bernal, Santiago Casado e Ismael Gil. Sin vuestro trabajo los datos experimentales nunca se habrían podido obtener. Gracias por aceptarme como uno más y por iluminarme con vuestro conocimiento tanto dentro como fuera de quirófano. Me habéis demostrado lo gran profesionales y magníficas personas que sois. Ojalá todo el mundo fuese igual de responsable y competente.

Javier Civera. Me ayudaste a entender el funcionamiento del EKF-SLAM soportándome durante los primeros años de doctorado. Gran parte de los resultados de esta tesis te los debo a ti.

Víctor F. Muñoz. Gracias por aportar la primera secuencia laparoscópica con la cual se comenzó a trabajar.

Antonio Güemes, Miguel Ángel Bielsa y Félix Lamata. Vosotros me facilitásteis mi primer acceso a un quirófano. Gracias a ello pude realizar mi primera captura de una intervención laparoscópica completa.

Andrew Davison, Ian Reid y Brian Williams. Gracias tanto por el software inicial de SLAM monocular como por el de relocalización.

A todos mis compañeros del grupo de robótica. Vosotros me habéis hecho más llevaderos estos años gracias a los ratos vividos durante las

comidas y a las “sesiones de terapia” en los cafés.

Marta Salas. Dicen que un amigo es un tesoro. Espero que yo forme parte de tu fortuna igual que tu formas parte de la mía. Ojalá no te pierda nunca como amiga aunque el tiempo y la distancia se interpongan en nuestras vidas. Gracias por ayudarme siempre que te lo he pedido y por contagiarme tu afición a la fotografía.

Estíbaliz Fraca. Si la riqueza en el mundo se midiese por el tamaño del corazón y el buen hacer de la gente, tu formarías parte de ese 1% de la población que controla el 40% de la riqueza mundial. Gracias por permitirme conocerte. Tú siempre has estado ahí, disponible para lo que yo necesitase.

Belén Masiá. Pusiste en marcha el algoritmo de relocalización durante tu PFC y has revisado mi inglés siempre que te lo he pedido sin poner ninguna objeción. Más aún, te ofrecías voluntaria para hacerlo. Tu PFC me permitió descubrir a una maravillosa persona que hizo que me prendara de su forma de ser y de pensar. Sigue siendo como eres, igual de madura, formal y responsable. Eres una de esas personas que me dolería perder como amiga.

A mis amigos de la infancia, de la universidad y del dotorado. Nombraros a todos me llevaría un par de páginas. A todos vosotros gracias por los buenos ratos vividos que hicieron que me evadiese del doctorado y de mis problemas personales. Ahora que voy a ser libre, espero vivir muchos más.

A toda mi familia, por estar siempre ahí aguantando y apoyando, y sobre todo a mis padres por todos los duros momentos que nos han tocado vivir durante el desarrollo de esta tesis.

Finalmente, al proyecto español FIT-360005-2007-9, al proyecto europeo RAWSEEDS (FP6-IST-045144), y al proyecto del ministerio de ciencia e innovación DPI2009-07130 por aportar la financiación necesaria para el desarrollo de este trabajo.

# List of Acronyms

| | |
|---|---|
| **1-PR** | 1-Point RANSAC |
| **5-PR** | 5-Point RANSAC |
| **AM** | Active Matching |
| **AR** | Augmented Reality |
| **BA** | Bundle Adjustment |
| **CT** | Computed Tomography |
| **DoF** | Degrees of Freedom |
| **DoV** | Direction of View |
| **EKF** | Extended Kalman Filter |
| **FAST** | Features from Accelerated Segment Test |
| **FEM** | Finite Elements Method |
| **FoV** | Field of View |
| **IC** | Individual Compatible matches |
| **ID** | Inverse Depth |
| **JCBB** | Joint Compatibility Branch and Bound |
| **LVHR** | Laparoscopic Ventral Hernia Repair |
| **MIS** | Minimally Invasive Surgery |
| **MRI** | Magnetic Resonance Imaging |
| **NRSfM** | Non-Rigid Structure from Motion |
| **PnP** | Perspective-n-Point Problem |
| **PTAM** | Parallel Tracking And Mapping |

**RANSAC** RANdom SAmple Consensus

**RLR** Randomised List Relocalization

**SfM** Structure from Motion

**SLAM** Simultaneous Localization And Mapping

**TAC** Tomografía Axial Computarizada

**VSM** Visual SLAM Measurement

## Resumen

A pesar de los grandes avances que ha supuesto la cirugía laparoscópica en el ámbito quirúrgico, esta aún presenta dificultades en su realización causadas principalmente por la complejidad de sus maniobras y sobre todo por la pérdida de la percepción de profundidad (se ha pasado de una cirugía completamente 3D –cirugía abierta o laparotomía–, donde el cirujano tenía un contacto directo con los órganos, a una cirugía realizada a través de una cámara, donde la escena 3D se proyecta a un plano 2D –imagen–, y un instrumental especial).

El principal objetivo de esta tesis es hacer frente, en la medida de lo posible, a esa pérdida de percepción 3D haciendo uso de algoritmos de Simultaneous Localization and Mapping (SLAM por sus siglas en inglés) desarrollados en los campos de la robótica móvil y la visión por computador a lo largo de estos últimos años. Estos algoritmos permiten localizar, en tiempo real (25 $\sim$ 30 imágenes por segundo), una cámara que se mueve libremente dentro de un entorno rígido desconocido y, al mismo tiempo, construir un mapa de ese entorno únicamente haciendo uso de las imágenes capturadas por dicha cámara.

Este tipo de algoritmos ha sido ampliamente validado tanto en entornos de construcción humana (edificios, habitaciones, ...) como en entornos de exteriores mostrando una gran robustez ante oclusiones, movimientos bruscos de la cámara o entornos atestados de objetos. En esta tesis se pretende extender un poco más el uso de este tipo de algoritmos mediante su aplicación a la cirugía laparoscópica. Debido a la naturaleza de las imágenes del interior del cuerpo (escenas deformables, brillos, condiciones de iluminación variable, limitaciones en los movimientos, ...), aplicar esta clase de algoritmos a laparoscopia supone un completo desafío.

El conocimiento de la localización de la cámara (laparoscopio en cirugía laparoscópica) con respecto a la escena (cavidad abdominal) y el mapa 3D de esta abren varias posibilidades de gran interés dentro del campo quirúrgico. Este conocimiento permite: realizar inserciones en realidad aumentada sobre las imágenes del laparoscopio (ej. alineamiento de modelos TAC 3D preoperatorios); mediciones de distancias 3D intracavitarias; o reconstrucciones 3D fotorrealistas de la cavidad abdominal recuperando sintéticamente la profundidad perdida. Estas nuevas cualidades aportan seguridad y rapidez a los procedimientos quirúrgicos sin perturbar el flujo de trabajo clásico. Por lo tanto, estas nuevas herramientas están disponibles en el arsenal del cirujano siendo este quien decide si usarlas o no. Además, el conocimiento de la localización de la cámara con respecto a la cavidad abdominal del paciente es

fundamental para el futuro desarrollo de robots que operen automáticamente ya que, gracias a esa localización, el robot sería capaz de localizar, con respecto al paciente, cualquier otra herramienta que fuese controlada por él mismo.

De forma detallada, las contribuciones de esta tesis han sido:

1. Demostrar la viabilidad de aplicar algoritmos de SLAM en cirugía laparoscópica mostrando experimentalmente la obligatoriedad de emplear un emparejamiento de datos robusto.

2. Robustecer uno de estos algoritmos, en concreto el EKF-SLAM, adaptando un sistema de relocalización y mejorando la asociación de datos mediante un algortimo de emparejamiento robusto.

3. Desarrollo de un método de emparejamiento robusto (algoritmo 1-Point RANSAC).

4. Desarrollo de un procedimiento quirúrgico que facilita el uso de SLAM visual en laparoscopia.

5. Validar extensivamente el algoritmo de EKF-SLAM robusto (EKF + relocalización + 1-Point RANSAC) obteniendo errores milimétricos y funcionando en tiempo real sobre simulaciones y cirugías humanas reales. La cirugía seleccionada ha sido la eventroplastia (reparación de hernias ventrales).

6. Demostrar el potencial que tienen estos algoritmos en laparoscopia: permiten hacer inserciones en realidad aumentada, recuperar sintéticamente la profundidad del campo operativo perdida por usar cámaras monoculares, y realizar medidas de distancias únicamente con una herramienta de laparoscopia y las imágenes obtenidas por el laparoscopio.

7. Hacer una validación clínica mostrando que estos algoritmos permiten acortar los tiempos quirúrgicos de las operaciones y además aportar seguridad a estas.

## Abstract

In spite of the great advances in laparoscopic surgery, this type of surgery still shows some difficulties during its realization, mainly caused by its complex maneuvers and, above all, by the loss of the depth perception. Unlike classical open surgery –laparotomy– where surgeons have direct contact with organs and a complete 3D perception, laparoscopy is carried out by means of specialized instruments, and a monocular camera (laparoscope) in which the 3D scene is projected into a 2D plane –image.

The main goal of this thesis is to face with this loss of depth perception by making use of Simultaneous Localization and Mapping (SLAM) algorithms developed in the fields of robotics and computer vision during the last years. These algorithms allow to localize, in real time ($25 \sim 30$ frames per second), a camera that moves freely inside an unknown rigid environment while, at the same time, they build a map of this environment by exploiting images gathered by that camera.

These algorithms have been extensively validated both in man-made environments (buildings, rooms, ...) and in outdoor environments, showing robustness to occlusions, sudden camera motions, or clutter. This thesis tries to extend the use of these algorithms to laparoscopic surgery. Due to the intrinsic nature of internal body images (they suffer from deformations, specularities, variable illumination conditions, limited movements, ...), applying this type of algorithms to laparoscopy supposes a real challenge.

Knowing the camera (laparoscope) location with respect to the scene (abdominal cavity) and the 3D map of that scene opens new interesting possibilities inside the surgical field. This knowledge enables to do augmented reality annotations directly on the laparoscopic images (e.g. alignment of preoperative 3D CT models); intracavity 3D distance measurements; or photorealistic 3D reconstructions of the abdominal cavity recovering synthetically the lost depth. These new facilities provide security and rapidity to surgical procedures without disturbing the classical procedure workflow. Hence, these tools are available inside the surgeon's armory, being the surgeon who decides to use them or not. Additionally, knowledge of the camera location with respect to the patient's abdominal cavity is fundamental for future development of robots that can operate automatically since, knowing this location, the robot will be able to localize other tools controlled by itself with respect to the patient.

In detail, the contributions of this thesis are:

1. To demonstrate the feasibility of applying SLAM algorithms to laparoscopy showing experimentally that using robust data association is a must.

2. To robustify one of these algorithms, in particular the monocular EKF-SLAM algorithm, by adapting a relocalization system and improving data association with a robust matching algorithm.

3. To develop of a robust matching method (1-Point RANSAC algorithm).

4. To develop a new surgical procedure to ease the use of visual SLAM in laparoscopy.

5. To make an extensive validation of the robust EKF-SLAM (EKF + relocalization + 1-Point RANSAC) obtaining millimetric errors and working in real time both on simulation and real human surgeries. The selected surgery has been the ventral hernia repair.

6. To demonstrate the potential of these algorithms in laparoscopy: they recover synthetically the depth of the operative field which is lost by using monocular laparoscopes, enable the insertion of augmented reality annotations, and allow to perform distance measurements using only a laparoscopic tool (to define the real scale) and laparoscopic images.

7. To make a clinical validation showing that these algorithms allow to shorten surgical times of operations and provide more security to the surgical procedures.

---

# Contents

# List of Figures

# 1 Introduction

## 1.1 Laparoscopy

Laparoscopic surgery is a modern technique of minimally invasive surgery (MIS). In this technique, operations inside the abdominal cavity are performed with small incisions (5-10 mm) through the abdominal wall in contrast with large incisions of classical open surgery (laparotomy).

Laparoscopy requires of a telescopic rod lens system connected to a videocamera (endoscope or laparoscope) that gathers images of the abdominal cavity. These images are displayed on some monitor (TV, computer screen) and used by surgeons during the operation in order to see the interior of the abdomen. Since the abdominal cavity is dark, attached to the lens is a fiber optic cable system connected to a "cold" light source (halogen or xenon) that illuminates the operative field. The abdominal wall has to be separated from the internal organs, and then the abdomen is insufflated with carbon dioxide gas ($CO_2$) and blown up like a balloon creating a workspace called pneumoperitoneum. $CO_2$ is used because it is a gas produced by the human body, therefore, it is easily absorbed by tissues and removed by the respiratory system. Additionally, $CO_2$ is non-flammable, which is important because electrosurgical devices are commonly used in laparoscopic procedures. Finally, both camera and tools are inserted in the abdominal cavity through 5-10 mm cannula-shaped input ports called trocars.

Despite the incisions of the input ports being small, laparoscopy has its own risks. Precisely the most important ones are caused by trocars. During their insertion, they can damage internal organs (small or large bowel) causing perforations and peritonitis, or penetrate blood vessels causing vascular injuries such as hematomas or hemorrhages that may be life-threatening.

Electrosurgical tools can cause electrical burns that can lead to organ perforations and even peritonitis. Patients with existing pulmonary or heart disorders may not tolerate pneumoperitoneum resulting in a need for conversion to open surgery after the initial attempt at the laparoscopic approach. Besides, the pressure exerted by $CO_2$ may cause difficulties in the venous return and increase the cardiac output making this surgery more dangerous for this type of patients. Finally, since not all of the $CO_2$ introduced into the abdominal cavity is removed through the incisions, it tends to rise and push the diaphragm, muscle that separates the abdominal cavity from the thoracic cavity and facilitates breathing, putting pressure on the phrenic nerve and causing sensation of pain that disappears as $CO_2$ is eliminated through respiration.

Nevertheless, these risks are thoroughly compensated with the advantages of this kind of surgery versus laparotomy. Although there exists a minimal risk of hemorrhages, this is much lower than in laparotomy, reducing the chance of needing a blood transfusion. Smaller incisions reduce muscular injuries in the abdominal wall resulting in a lesser post-operative aesthetic impact, and a lower risk of wound infections and pain; therefore, less antibiotics and pain medication are needed. Smaller incisions also shorten recovery time, often with a same day discharge, which leads to a faster return to everyday living. Besides, reduced exposure of internal organs to possible external contaminants decreases the risk of acquiring infections.

From the surgeons' perspective, laparoscopy presents several disadvantages with respect to laparotomy. The limited range of motion of surgical tools, that results in a loss of dexterity; the tool tip that moves in the opposite direction to the surgeon's hands, due to it pivoting around the fulcrum (entering point in the abdominal cavity), making laparoscopic surgery non-intuitive; and the indirect manipulation of tissues through the laparoscopic tools, that results in a reduction of tactile sensation making diagnosis of tissues tissues by feeling more difficult (e.g. to detect tumors), and perform delicate procedures such as suturing, make the learning curve be complex and require to make a great effort to learn this technique. Additionally, since the operations are usually performed through monocular images (the majority of endoscopes are monocular), surgeons are faced with the loss of depth perception. Besides, the endoscope Field of View (FoV) is limited and not all the operative field is visible at each moment, as a consequence, surgeons require a deep knowledge of the patient's anatomy. This knowledge is also required when surgeons need to remove critical structures such as tumors or work in critical areas such as areas near vital blood vessels.

The first use of a laparoscope was performed in a dog by Georg Kelling

in 1901. In 1910, Hans Christian Jacobaeus, based on Kelling's works, reported the first human laparoscopic human intervention [Hat+06]. However, the use of laparoscopy was very limited, only diagnosis and performance of simple gynecologic procedures were performed, until 1975 when Tarasconi performed the first organ resection (salpingectomy) first reported in the Third AAGL (American Association of Gynecologic Laparoscopist) Meeting in 1976, and later published in [Tar81]. Nowadays, laparoscopy is very extended and practically any abdominal or pelvic surgery can be performed with this technique (cholecystectomies, hepatectomies, bowel resections, hernia repairs, ...). The present and future of laparoscopy is strongly tied to the developments in computer vision and robotics which ease the dexterity and improve the depth perception, the FoV, and the tactile sensation by means of robots, stereo endoscopes, SLAM algorithms or haptic interfaces. It is worth to mention the DaVinci System as an example of these developments.

## 1.2 SLAM

Simultaneous Localization And Mapping (SLAM) is a classical problem and one of the most researched topics in mobile robotics. Given a mobile sensor moving along an unknown trajectory in an unknown environment, SLAM is able to simultaneously estimate both the environment structure (a 3D map of the environment) and the sensor location with respect to that map. This estimation process is carried out taking the information gathered by the sensor as the sole input data to the SLAM algorithm. Additionally, SLAM is usually required to work in real-time at frame rate.

In the most typical SLAM problem, sensory information comes from proprioceptive sensors –odometry or inertial measurement units– and exteroceptive sensors, that measure entities external to the robot. Traditionally, laser has been the predominant exteroceptive sensor used in SLAM, although other sensors, like sonars, have also been used. It is only very recently that cameras have been adopted massively by the robotic community as the main SLAM sensor. In this thesis, a monocular camera (laparoscope) is used as the unique sensor and full 3D SLAM is performed. This configuration is usually named monocular SLAM.

The monocular SLAM problem is particularly challenging because only a sequence of 2D projections of a 3D scene is available; in any case, 30 Hz real-time systems estimating up-to-scale 3D camera motions and maps of 3D points using commodity cameras and computers are widely available for mobile robotics environments nowadays. All these systems provide extensive experimental validation of the SLAM algorithms and real-time performance

for man-made, mainly rigid, scenes which are typical in mobile robotics.

Monocular SLAM has been possible thanks to intense research on salient feature detection and description [Can86; HS88; ST94; Low04; RD05] and spurious rejection [FB81; NT01] which has provided with an automated way of robustly matching features along images. Finally, it has been tackled from two radically different approaches. On the one hand, algorithms based on keyframes (*keyframe methods*) which try to adapt traditional pairwise offline Stucture form Motion (SfM) + Bundle Adjustment (BA) methods to achieve sequential online estimation. In order to do so, a sparse set of keyframes is chosen and SfM methods and BA are applied over a subset of them that are close to the current frame. One of the best performers in this vein is the Parallel Tracking And Mapping (PTAM) algorithm proposed by Klein and Murray [KM07]. On the other hand, *filtering methods* which apply Bayesian filtering techniques that, at each step, integrate the information from the current frame into a multidimensional probability distribution that summarizes the information gathered for all previous frames along the sequence. Davison was the first to demonstrate real-time performance with this approach using an Extended Kalman Filter (EKF) [Dav03; Dav+07].

## 1.3 The aim: Laparoscopy as a monocular SLAM problem

The main goal of this thesis is to demonstrate the feasibility of applying SLAM algorithms, that come from robotics and computer vision and that recover the 3D of a scene and the camera motion in real time, to laparoscopic environments.

Laparoscopy can be posed as a monocular SLAM problem. In laparoscopy, a surgeon explores the abdominal cavity by pivoting the laparoscope around the fulcrum. Then, the tip of the laparoscope moves inside the abdomen gathering images from this cavity. SLAM algorithms enable to estimate an up-to-scale 3D map of the observed cavity from these images without resorting to any additional sensor such as optical or magnetic trackers, accelerometers, structured light, or artificial landmarks. Furthermore, it is worth noting that SLAM not only recovers the 3D model, but also the actual trajectory followed by the laparoscope.

SLAM opens new interesting possibilities inside the surgical field. The knowledge of the laparoscope location with respect to the abdominal cavity enables to do augmented reality annotations directly on laparoscopic images (e.g. alignment with preoperative 3D CT models, signal critical regions like blood vessels, or provide other additional information); intracavity 3D dis-

tance measurements, provided that the absolute scale of the map is recovered from the observation of a known-size surgical tool; or photorealistic 3D reconstructions of the adominal cavity recovering synthetically the lost depth and improving the FoV. These new facilities provide security and rapidity to surgical procedures.

Furthermore, the camera location knowledge with respect to the patient's abdominal cavity is fundamental for future development of robots that can operate automatically since, thanks to this location, the robot will be able to localize other tools, controlled by itself, with respect to the patient.

## 1.4 Related Work

The usage of SLAM-like methods in MIS can be traced back to the seminal work in providing 3D models from body monocular image sequences proposed by Burschka *et al.* [Bur+05]. Assuming scene rigidity, the system produces a map for registering preoperative CT scans with the endoscopic images. Its main limitations are map size and the lack of robustness with respect to outlier matches. Computer vision methods based on a discrete set of views have been applied to medical images, assuming scene rigidity, in order to just compute the 3D structure of the cavity. In [WSC07], the classical two view RANSAC structure from motion is applied to mannequin images to determine the 3D structure; a constraint-based factorization 3D modelling method produces a dense 3D reconstruction in near real time. In [Dan+07], structure from motion is used to build a photorealistic 3D reconstruction of the colon; in a first stage, images are processed pairwise to produce an initial 3D map; in a second stage, all the maps are joined in a unique photorealistic 3D cavity model. In [Mir+12], these methods have been refined and extended to deal with multiple views with a significant boost in performance in rigid medical scenes. Thanks to careful feature selection and a quite robust spurious tracking ASKC [Wan+08] (Adaptive Scale Kernel Consensus), they are able to estimate both 3D models of the cavity and the location of the camera with respect to this cavity up to submillimeter accuracy in a cadaver for endonasal skull surgery. Hu *et al.* in [Hu+12], in a similar vein, propose a 3D structure estimation from multiple images. They deal with outliers by means of the trifocal tensor, then a bundle adjustment optimization is performed reporting accuracies slightly over a millimeter.

Cavity 3D reconstruction from medical sequences of a non-moving stereo endoscope has been proposed in [MDCM01] and [SDY05]. Visual SLAM methods have proven to be valid to process medical images coming from a moving stereo endoscope in [Mou+06], where an EKF stereo SLAM, as-

suming smooth camera motion and scene rigidity, is validated over synthetic sequences and qualitatively over in-vivo animal sequences; no usage of algorithms robust to spurious data is reported. In [MY10], the scene non-rigidity is considered: EKF visual SLAM is combined with a dynamic periodic model, learnt online, to estimate the respiration cycle from stereo images.

Intensive research is being done in designing medical miniaturised devices that can provide depth map as stereo endoscopes while avoiding the correspondence problem. A monoport structured light device based on a stereo scope is presented in [Mau+12], preliminary but promising results are reported. In [Sch+12], a catadioptric structured light prototype specifically designed to recover the lumen of a tubular cavity is described, reporting 0.1 mm accuracy tested on a phantom and ex-vivo animal. In [Haa+13], a monoport prototype combining time-of-flight (ToF) and RGB is proposed; despite the low resolution of the depth map, promising results are reported. All these previous devices are still under development, but in any case the rich 3D information that they can provide suggests a promising venue of research for SLAM algorithms.

Our proposal is also based on EKF SLAM, however, we deal with monocular sequences, our method is robust to outliers, and we provide extensive validation over both synthetic data and real human in-vivo sequences.

Malti *et al.* [MBC11] propose a two-phase 3D monocular reconstruction of the abdominal cavity based on NRSfM (Non-Rigid Structure from Motion). The first phase consists in exploring the abdominal cavity in order to obtain an initial 3D rigid reconstruction using two views and the essential matrix + camera resection + bundle adjustment combination. Afterwards, in the second phase, this reconstruction is exploited to infer 3D scene deformations during operation. The algorithm is one of the first to deal with the scene non-rigidity under general deformation. However, the correspondences are assumed known, computing time is not reported, and only a qualitative validation over one sequence is provided.

Recently, methods based on photometric properties are being used in endoscopic sequences. In [MBC12], the results of [MBC11] are taken as input to provide a dense 3D model based on shape from shading; only a quantitative validation for synthetic data, and qualitative validation for one in-vivo sequence of the uterus are provided. Collins *et al.* [CB12b] propose shape from shading in real time at 23Hz for medical rigid scenes thanks to a GPGPU implementation; in-vivo and ex-vivo experimental validation is provided but the authors acknowledge poor conditioning and the strong assumption of a constant albedo as prior data. The same authors propose in [CB12a] a preliminary work based on photometric stereo with learnt reflectance models

in order to estimate a 3D reconstruction of an organ from one image using three different color light sources; for this, the tip of the endoscope has to be modified to include three color filters. The method is able to compute the absolute depth without detecting image features, although it is sensitive to illumination changes. They provide preliminary experiments over one in-vivo pig liver sequence, including comparison with respect to ground-truth. The main advantage of photometric methods with respect to feature-based ones is their ability to deal with textureless images. However, they are still sensitive to illumination changes.

All the aforementioned methods are able to yield camera location with respect to the observed scene, a basic requirement for augmented reality insertions, navigation, or multimodal image fusion that have proven to be useful in medical applications (e.g. [Oku+11; Nic+11]). In [Tot+11], EKF stereo SLAM is also used to artificially expand the intraoperative field-of-view of the laparoscope (dynamic view expansion).

Finally, it is worth mentioning the recent review about optical 3D reconstruction from medical image laparoscopic sequences provided by Maier-Hein *et al.* [MH+13].

## 1.5 Contributions of this Thesis

Due to the specific characteristics of laparoscopic scenes (sudden endoscope motions, endoscope extraction and reinsertion, tissue deformations, ...), the use of SLAM algorithms is possible provided that a robust and efficient spurious detector and a good relocalization system are available.

Without loss of generality, the monocular EKF-SLAM proposed in [Dav03; CDM08] has been chosen as a basic demonstrator of the feasibility of this type of algorithms in laparoscopy because it is well known, mature, and performs well in small environments like rooms (abdominal cavity is even smaller) and in real time. Nevertheless, other methods, like the proposal of Klein and Murray [KM07], would perform equally well provided that they comply with the two previous conditions (relocalization system and a robust-to-spurious policy).

The first version of the EKF-SLAM integrated the Joint Compatibility Branch and Bound (JCBB) algorithm [NT01] for robust data association. JCBB performs well when low spurious rates are present, however, its exponential computational cost in the number of spurious violates the real-time restrictions when the number of outliers increases. Additionally, this version lacked of a relocalization system what hindered to process laparoscopic sequences that included sudden laparoscope motions, occlusions or laparo-

scope extractions and reinsertions into the abdominal cavity. Nevertheless, this version allowed to prove the potential use of monocular SLAM in laparoscopy.

This initial version has been robustified by integrating the Randomised List Relocalisation (RLR) system [WKR07] and substituting JCBB by the new 1-Point RANSAC algorithm (1-PR) developed in this thesis.

Finally, the combination of EKF-SLAM + RLR + 1-PR has been extensively validated with real human laparoscopic sequences of ventral hernia repairs. This validation has been carried out in terms of accuracy, obtaining millimetric reconstructions, and clinical utility.

In detail, the main contributions of this thesis are:

1. **A proof of concept demonstrating the potential use of visual SLAM applied to laparoscopy**. In Chapter 2, it has been proposed to use EKF + JCBB SLAM to process real hand-held monocular laparoscopic sequences in order to improve the FoV by means of photorealistic reconstructions; to measure distances; and to insert augmented reality annotations. However, JCBB data association violates the real-time constraints when several outliers are present in the image (a very common situation in laparoscopy), due to its exponential computational complexity in the number of spurious. Furthermore, this combination lacks of a relocalization system. It makes impossible to run laparoscopic sequences if the laparoscope is suddenly moved or extracted and reinserted into the abdominal cavity, or if there are large occlusions. This contribution was reported in [GG+09b; GG+09a].

2. **Developing the 1-Point RANSAC (1-PR) algorithm for robust data association**. This RANSAC-based algorithm, detailed in Chapter 3, exploits the probabilistic prediction of the EKF filter to generate hypotheses of the measurements using only one point (measurement) as dataset. The hypotheses are voted by the other non-integrated measurements. The most voted hypothesis is integrated with an EKF update in order to detect the spurious matches. This algorithm overcomes JCBB, since it can cope with high spurious rates in real time, and works after integrating a subset of points (one point), what corrects part of the system errors allowing to obtain more accurate estimations. This contribution was reported in [Civ+09a; Civ+10].

3. **Integrating the Randomised List Relocalisation system (RLR) along with EKF + 1-PR**. RLR, proposed in [WKR07], is a relocalization system that detects the losses of tracking and freezes the recovered map in order to avoid a possible map corruption. Then, it tries

to relocalize the camera searching for putative matches between the current image and the frozen map and applying RANSAC + 3-Point-Pose algorithm. When it finds a good camera location, it unfreezes the map and reactivates the normal system behavior. In Chapter 3, it is integrated with EKF + 1-PR in order to obtain a possible robust EKF-SLAM to be used in laparoscopy. This contribution was reported in [GGCM11].

4. **A surgical protocol for using visual SLAM in ventral hernia repair surgeries**. In order to use visual SLAM in an operating room, a surgical protocol easily integrable into surgical procedures has been developed. This protocol is presented in Chapter 4. This contribution was reported in [Gil+11a; Gil+11b].

5. **An extensive accuracy validation of the EKF + 1-PR + RLR**. The validation has been performed both with real laparoscopic sequences corresponding to fifteen ventral hernia repairs, and with simulations. The ventral hernia repair operation has been chosen because surgeons need to measure the hernia dimensions, and hence these measurements are used as ground truth to validate the system. Finally, SLAM reconstructions are compared with measurements and with simulation ground truth obtaining millimetric errors and working in real time in both cases (real surgeries and simulations). Chapter 4 thoroughly details the ventral hernia repair based on visual SLAM and the validation. This contribution was reported in [GG+14].

6. **A clinical validation of the EKF + 1-PR + RLR**. A validation of the system is not complete without a clinical validation. The system has demonstrated not to disturb the classical ventral hernia repair workflow. Besides, it has allowed to shorten surgical times of operations and provide more security to the surgical procedures. Thus, the system may be incorporated as an additional tool inside the surgeon's armory. This validation is also detailed in Chapter 4, and the contribution was reported in [Ber+].

It is worth noting that the works [Gil+11a], [Gil+11b] and [Ber+] correspond with clinical publications, and, for that reason, the first authors are surgeons. However, these publications would not exist without the engineering contribution by the author of this thesis that has been essential in all of them.

# 2 Monocular EKF-SLAM

In the last years, SLAM research has focused on monocular cameras as the unique sensorial input, giving origin to monocular SLAM methods. 30 Hz real-time systems estimating full 3D camera motions and maps of 3D points using commodity cameras and computers have been reported [Dav03; KM07; ED08]. Traditionally, there have been two different approaches to monocular SLAM: keyframe methods and Bayesian filtering methods. They are briefly summarized in Section 2.1.

The main goal of this thesis is to demonstrate that these methods can be applied in laparoscopic surgery. Thus, without loss of generality, the monocular Extended Kalman Filter SLAM (EKF-SLAM) proposed in [Dav03; CDM08] has been chosen as a basic demonstrator of the feasibility of this type of algorithms in laparoscopy.

The original implementation of monocular EKF-SLAM, proposed by Davison [Dav03], encodes the map features in a 3D vector which represents the world point localization (Euclidean parametrization). However, this parametrization suffers from large linearization errors when there are points that have been seen with low parallax. Since EKF estimation strongly depends on measurement model linearity, bad linearizations will produce a degradation of this estimation. The inverse depth (ID) parametrization, proposed by Civera *et al.* [CDM08], improves this situation, however, it increases the size map, and hence increments the computational cost of the estimation. Therefore, the best way to operate with monocular EKF-SLAM is to encode low parallax features in ID and, as the estimation evolves and they are seen with enough parallax, convert them to Euclidean. Section 2.2 details the monocular EKF-SLAM, the Euclidean parametrization and the ID parametrization and its conversion to Euclidean.

Additionally, the EKF update stage assumes a perfect data association. This, however, is not true and just one spurious match may wreck the estimation. Joint Compatibility Branch and Bound (JCBB), detailed in Section 2.3, is a state-of-the-art algorithm for data association in EKF-SLAM. This algorithm detects and removes the spurious measurements before the EKF update, improving the robustness of SLAM.

Finally, monocular SLAM recovers a sparse 3D map of the scene and the camera motion along with the corresponding covariances. Both the map and the motion can be used as a geometrical backbone to support useful information such as 3D distance measurements, augmented reality (AR) insertions or photorealistic reconstructions (Section 2.4).

The first contribution of this thesis is a proof-of-concept, reported in [GG+09b; GG+09a], that proves the feasibility of using monocular visual SLAM algorithms in laparoscopic surgery showing its potential in this surgical field (Section 2.5). It is worth noting that this contribution, up to the author's knowledge, is the first one which applies monocular visual SLAM algorithms in real laparoscopic human surgeries.

## 2.1 SLAM Methods

Monocular SLAM has been tackled by means of methods based on keyframes, which try to adapt traditional pairwise offline Stucture form Motion (SfM) + Bundle Adjustment (BA) in order to achieve sequential online estimation, or by means of methods based on filtering, which apply Bayesian filtering techniques that, at each step, integrate the information from the current frame into a multidimensional probability distribution which summarizes the information gathered for all previous frames along the sequence.

### 2.1.1 Keyframe Methods

Keyframe methods are strongly related to SfM methods, whose origins can be traced back to the so-called Photogrammetry (second half of $19^{th}$ century). Photogrammetry aims to extract geometric information from images. Initially, it started with a set of features manually identified by the user, and then applied non-linear optimization techniques, known as BA [MBM01], to minimize the reprojection error. Nowadays, research on computer vision has allowed to achieve complete automation by assuming rigidity in the scene and by automatizing the feature extraction, matching, and spurious detection between images.

SfM has been usually processed by pairwise geometry algorithms [HZ04]

and refined by global optimization procedures (BA) [Tri+00] in order to minimize the reprojection error and refine the estimation into a globally consistent one. One of the main drawbacks of the traditional SfM + BA combination is its incapacity to deal with long sequences of images, since it was initially thought for the processing of sparse sets of images (SfM relates pairs –at most triplets– of images, but lacks a global formulation for an image stream). Recently, there has been significant novel research in the field aiming to extend the capabilities of SfM methods to sequentially process large image sequences, the estimation being carried out in real-time [KM07; Mou+09]. In this vein, one of the best performers in visual SLAM is the Parallel Tracking And Mapping (PTAM) algorithm proposed by Klein and Murray [KM07]. This keyframe-based algorithm makes use of two parallel processing threads. The first one performs camera tracking assuming a known map of natural features at 30 fps. The second constructs a consistent map performing global BA over selected frames of the sequence which summarize the whole sequence.

### 2.1.2 Filtering Methods

Filtering methods are based on Bayes filters [TBF05]. They estimate recursively a probability distribution function over the unknown parameters (camera location and 3D map of the scene) of a state vector from measurements gathered by a sensor (camera) and the dynamic and measurement models in two steps: prediction, and update. In the first one, the probability distribution function for the frame at time instant $k-1$ is projected into the frame at time instant $k$ based on the probabilistic dynamic model of the system. In the update stage, measurements are collected, according to the measurement model known in advance, and fused with the probability distribution function from the prediction step using Bayes' rule.

The key difference between filtering methods and keyframe ones is that they do not operate in a pairwise manner –estimating location from one frame with respect to another– nor do they pile up measurements waiting for a BA optimization. Instead of that, the overall state of the system is summarized into a multidimensional probability distribution, which is updated as new measurements arrive and their information is integrated in it. Therefore, its computational complexity scales with the size of the state and not with the number of frames, being naturally suited for the processing of large streams of measurements.

Historically, the Extended Kalman Filter (EKF), the non-linear version of the Kalman Filter, was the first filtering technique to offer a solution to the SLAM task (EKF-SLAM) [SSC87; Dis+01; Cas+99], and also the first one to demonstrate real-time performance in visual SLAM [Dav03; Dav+07].

In addition to EKF, other different filters have been proposed trying to relax the EKF assumptions, but generally incurring in higher computational cost: the Unscented Kalman Filter [JU97], Particle Filters [Mon+02], and Sum of Gaussians Filter [DW+03]. Particle Filters and the Unscented Kalman Filter have been used for visual SLAM in [ED06] and [HKM09], respectively.

One of the drawbacks of EKF-SLAM is its quadratic computational cost with respect to the state size. SLAM research has pursued to reduce this computational cost, leading to interesting results. Information filters have taken advantage of the sparsity of the problem when presented in information form –dual of the covariance form–, e.g. [Thr+04; ESL05]. Submapping-based techniques have also been developed to cope with the complexity of the filtering-based SLAM estimation [ENT05]. EKF-based submapping has been applied to the visual estimation case in [Cle+07; Paz+08; PT08].

In spite of the quadratic computational cost, EKF filtering has been chosen in this thesis because it is well known, mature, and performs well in small environments where it reaches real time. Thus, an improved version of Davison's work [Dav03; Dav+07] will be used as a starting point. This version will be robustified by combining EKF with RANSAC and will be adapted and validated to work over real human laparoscopic sequences in the next chapters.

## 2.2 Monocular EKF-SLAM

EKF, the first Bayesian filtering solution successfully applied to the SLAM estimation problem, was proposed for monocular visual SLAM in [Dav03]. This algorithm estimates recursively a probability distribution function over the unknown parameters of a state vector $\mathbf{x}$ from measurements gathered by a sensor and the dynamic and measurement models in two steps: prediction, and update. In the first one, the probability distribution function $p\left(\mathbf{x}_{k-1}\right)$ from step $k-1$ is projected into step $k$ based on the probabilistic dynamic model of the system $p\left(\mathbf{x}_{k|k-1}|\mathbf{x}_{k-1|k-1},\mathbf{u}_k\right)$ –equation (2.1). In the update stage, measurements $\mathbf{z}_k$ are collected, according to the measurement model $p\left(\mathbf{z}_k|\mathbf{x}_{k|k-1}\right)$, known in advance, and fused with the probability distribution function from prediction step using the Bayes' rule –equation (2.2).

$$p\left(\mathbf{x}_{k|k-1}\right) = \int p\left(\mathbf{x}_{k|k-1}|\mathbf{x}_{k-1|k-1},\mathbf{u}_k\right) p\left(\mathbf{x}_{k-1}\right) d\mathbf{x}_{k-1} \tag{2.1}$$

$$p\left(\mathbf{x}_{k|k}\right) = \eta\, p\left(\mathbf{z}_k|\mathbf{x}_{k|k-1}\right) p\left(\mathbf{x}_{k|k-1}\right). \tag{2.2}$$

$\eta$ corresponds to the normalization constant that converts $p\left(\mathbf{x}_{k|k}\right)$ into a probability distribution function. It is worth noting that in order to com-

pute the probability distribution recursively, the algorithm requires the initial probability distribution over $\mathbf{x}$, $p(\mathbf{x}_0)$.

In the case of the Kalman Filter, it is assumed that dynamic and measurement models are linear and represented by multivariate normal distributions –equation (2.3). Under these assumptions and knowing the initial Gaussian probability distribution at time $k = 0$, the a posteriori distribution over the estimated parameters is Gaussian and thus, it may be represented by its mean and covariance $\mathbf{x} \sim \mathcal{N}(\hat{\mathbf{x}}, \mathbf{P})$. However, most of the real systems are not linear but show some degree of linearity. The EKF relaxes the linearity assumption by linearizing the dynamic and measurement models in the mean value at every step of the estimation. Hence, the more linear both models are, the better the EKF estimation is. It is worth noting that the EKF does not give the real a posteriori probability distribution function, but only a Gaussian approximation.

$$p(\mathbf{x}) = det\,(2\pi\Sigma)^{-\frac{1}{2}}\, exp\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\}\ , \tag{2.3}$$

The EKF prediction and update stages involve working with the mean $\hat{\mathbf{x}}$ and the covariance $\mathbf{P}$ of the state. The prediction equations are:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{f}_k(\hat{\mathbf{x}}_{k-1|k-1}, \hat{\mathbf{u}}_k) \tag{2.4}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k\mathbf{P}_{k-1|k-1}\mathbf{F}_k^\top + \mathbf{G}_k\mathbf{Q}_k\mathbf{G}_k^\top \tag{2.5}$$

being $\mathbf{f}\left(\hat{\mathbf{x}}_{k|k-1}, \hat{\mathbf{u}}_k\right)$ the non-linear equation modeling the dynamic evolution of the system; $\mathbf{u}_k$ the input given to the system ($\mathbf{u}_k = \mathbf{0}$ in monocular SLAM); $\mathbf{F}_k$ the derivatives of the dynamic model with respect to the state vector ($\mathbf{F}_k = \frac{\partial \mathbf{f}_k}{\partial \mathbf{x}}$); $\mathbf{Q}_k$ the state noise covariance; and $\mathbf{G}_k$ the derivatives of the dynamic model with respect to such noise ($\mathbf{G}_k = \frac{\partial \mathbf{f}_k}{\partial \mathbf{n}_k}$), being $\mathbf{n}_k$ the state noise.

The equations of the update state are:

$$\nu_k = \mathbf{z}_k - \mathbf{h}(\hat{\mathbf{x}}_{k|k-1}) \tag{2.6}$$

$$\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k \tag{2.7}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^\top\mathbf{S}_k^{-1}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\nu_k \tag{2.8}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1} \tag{2.9}$$

where $\mathbf{z}_k$ are measurements gathered at step $k$; $\mathbf{h}\left(\hat{\mathbf{x}}_{k|k-1}\right)$ the function that defines the sensor measurement model; $\mathbf{H}_k$ the derivatives of the measurement function by the state vector ($\mathbf{H}_k = \frac{\partial \mathbf{h}_k}{\partial \mathbf{x}}$); $\mathbf{R}_k$ the covariance of the

measurement noise; $\nu_k$ and $\mathbf{S}_k$ the filter innovation and its covariance; and $\mathbf{K}_k$ the filter gain.

Therefore, to compute the EKF estimation, it is mandatory to define the state vector ($\mathbf{x}$), the dynamic model or state transition equations ($\mathbf{f}$), and the measurement model ($\mathbf{h}$). These definitions, for the case of the visual SLAM problem, are detailed in Sections 2.2.1, 2.2.2 and 2.2.3, respectively.

### 2.2.1 State Vector Definition

In visual SLAM, the world map and the camera location are represented in a stochastic framework. This probabilistic representation at step $k$ is coded in a unique state vector modeled as a multivariate Gaussian, $\mathbf{x}_k$:

$$\mathbf{x}_k = \left(\mathbf{x}_v^\top, \mathbf{y}_1^\top, \mathbf{y}_2^\top, \cdots, \mathbf{y}_n^\top\right)^\top \quad . \tag{2.10}$$

It is composed of the camera state, $\mathbf{x}_v$, and the map defined by the location of every point, $\mathbf{y}_i$. See Section 2.2.4 for map point management details.

The camera state, $\mathbf{x}_v$, is formed by position, $\mathbf{r}^{WC}$, orientation encoded in a quaternion, $\mathbf{q}^{WC}$, and linear and angular velocities, $\mathbf{v}^W$ and $\omega^C$.

The map is composed of $n$ point features, $(\mathbf{y}_1^\top, \cdots, \mathbf{y}_n^\top)^\top$, whose locations are encoded either in Euclidean coordinates, $\mathbf{y}_i = (X_i, Y_i, Z_i)^\top$, or in inverse depth (ID), $\mathbf{y}_i = (x_i, y_i, z_i, \theta_i, \phi_i, \rho_i)^\top$.

The original monocular EKF-SLAM by Davison uses only Euclidean parametrization, which suffers from large linearization errors in the measurement model at low parallax. In order not to degrade the EKF estimation, low parallax features (features whose depth is much bigger than camera translation or recently initialized features which, even if close to the camera, produce low parallax) must be treated separately from the main map until there is enough information to insert them into the filter (delayed initialization). In the presence of this situation, the system initialization requires an initial known map obtained from a pattern.

Low parallax features are important because, although they cannot be used to estimate camera translation, they contribute to the estimation of orientation, and hence to improve the EKF performance.

Unlike Euclidean point coding, ID point coding [CDM08] improves the measurement linearity at low parallax. As a result, ID improves EKF performance, even for maps only composed of close features, by taking into account low parallax features. Besides, it avoids the use of a pattern during system initialization and the delayed feature initialization by immediate insertion of new features into the main map.

An ID feature is a 6 parameter vector:

$$\mathbf{y}_i = (x_i, y_i, z_i, \theta_i, \phi_i, \rho_i)^\top \tag{2.11}$$

where $x_i,\ y_i,\ z_i$ correspond to camera location when the point was observed for the first time, and $\theta_i$ and $\phi_i$ are azimuth and elevation angles which define the ray unit vector $\mathbf{m}(\theta_i, \phi_i)$. Point depth is coded by its inverse $\rho_i = 1/d_i$, so a point world location $\mathbf{x}_i$ is (Figure 2.1):

$$\mathbf{x}_i = \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} + \frac{1}{\rho_i}\mathbf{m}(\theta_i, \phi_i) \tag{2.12}$$

$$\mathbf{m}(\theta_i, \phi_i) = (\cos\phi_i \sin\theta_i, -\sin\phi_i, \cos\phi_i \cos\theta_i)^\top \tag{2.13}$$

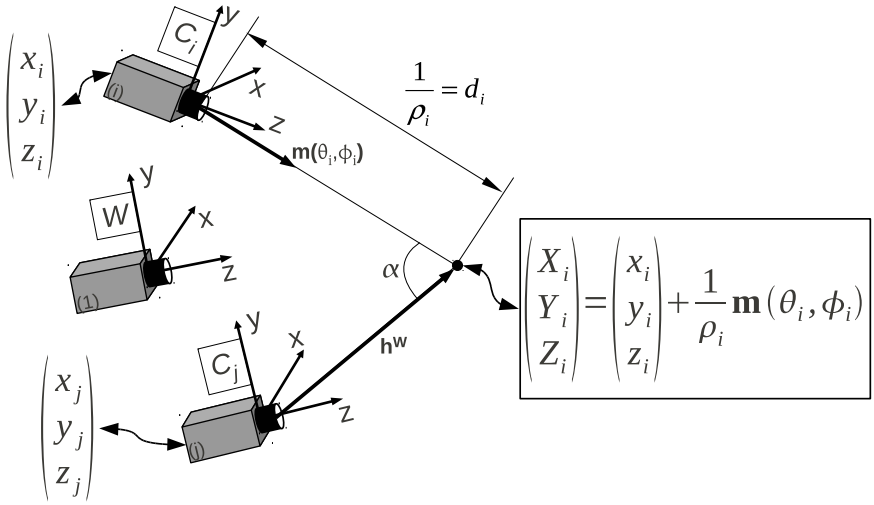



Figure 2.1: Camera-(1) defines the world frame, $W$. A feature is observed for the first time by camera-(i), the feature world location is defined with respect to the camera-(i) pose, $(x_i, y_i, z_i)^\top$, using the distance between camera-(i) and the feature, $d_i = 1/\rho_i$, and a unit directional vector, $\mathbf{m}(\theta_i, \phi_i)$, defined by its azimuth and elevation angles. The $\alpha$ angle stands for the feature parallax between camera-(i) and camera-(j) which can be computed with $\mathbf{m}(\theta_i, \phi_i)$ and $\mathbf{h}^W$ directional vectors, both of them defined in the world frame.

The only drawback of ID is that it doubles the size of each feature vector (it needs a 6D vector versus the 3D vector required by Euclidean parametrization) affecting directly to the computational cost of the EKF update. In

order to obtain accuracy without compromising the computational cost, the use of ID is restricted to low parallax map features converting them to Euclidean when they are seen with enough parallax. The parallax of a feature is defined as the angle $\alpha$ between the directional vector $\mathbf{m}(\theta_i, \phi_i)$ when the feature was initialized and the vector $\mathbf{h}^W$, in the world frame, that joins the current camera with the feature (Figure 2.1). The criterion for conversion is determined by the next linearity index ($L_d$):

$$L_d = \frac{4\sigma_d}{\|\mathbf{h}^W\|}|\cos\alpha| \tag{2.14}$$

$$\sigma_d = \frac{\sigma_\rho}{\rho_i^2},\ \cos\alpha = \frac{\mathbf{m}^\top \mathbf{h}^W}{\|\mathbf{h}^W\|} \tag{2.15}$$

$$\sigma_\rho = \sqrt{\mathbf{P}_{\mathbf{y}_i}(6,6)},\ \mathbf{h}^W = \mathbf{x}_i - \mathbf{r}_j^{WC} \tag{2.16}$$

where $\rho_i$ and $\sigma_\rho$ are the feature inverse depth and its standard deviation obtained from the state vector and the feature covariance matrix; $\mathbf{m}$ corresponds to Equation 2.13; and $\mathbf{h}^W$ is obtained from the 3D current camera ($\mathbf{r}_j^{WC}$) and feature (2.12) world positions. After each EKF estimation, $L_d$ is computed for all ID features and those whose $L_d < 10\%$ are converted to Euclidean encoding.

### 2.2.2 Dynamic Model

Regarding the state transition equation for the camera, a dynamic constant velocity model that encodes its smooth motion is proposed:

$$\mathbf{f}_v = \begin{pmatrix} \mathbf{r}_{k+1}^{WC} \\ \mathbf{q}_{k+1}^{WC} \\ \mathbf{v}_{k+1}^{W} \\ \omega_{k+1}^{C} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_k^{WC} + \left(\mathbf{v}_k^W + \mathbf{V}_k^W\right)\Delta t \\ \mathbf{q}_k^{WC} \times \mathbf{q}\left(\left(\omega_k^C + \Omega^C\right)\Delta t\right) \\ \mathbf{v}_k^W + \mathbf{V}^W \\ \omega_k^C + \Omega^C \end{pmatrix} \tag{2.17}$$

where $\mathbf{q}\left(\left(\omega_k^C + \Omega^C\right)\Delta t\right)$ is the quaternion defined by the rotation vector $\left(\omega_k^C + \Omega^C\right)\Delta t$.

The state noise vector ($\mathbf{n}$) is assumed to be composed of linear, $\mathbf{a}^W$, and angular acceleration, $\alpha^C$, acting as inputs producing, at each step, an impulse of linear velocity, $\mathbf{V}^W = \mathbf{a}^W \Delta t$, and angular velocity $\Omega^C = \alpha^C \Delta t$. Both of them are modeled as zero mean Gaussian processes with known covariance, $\mathrm{diag}\left(\mathbf{Q}_{\mathbf{a}^W}, \mathbf{Q}_{\alpha^C}\right)$.

Regarding the state transition equation for the scene points, a static model with zero state noise to encode the scene as perfectly rigid is proposed:

$$\mathbf{y}_{i_{k+1}} = \mathbf{y}_{i_k}. \tag{2.18}$$

The complete dynamic model ($\mathbf{f}_k$) is the stacking of (2.17) and an instance of (2.18) for each map point. The final state noise ($\mathbf{n}_k$) is assumed to be a zero mean multivariate normal distribution with known covariance $\mathbf{Q}_k = \text{diag}\left(\mathbf{Q}_{\mathbf{a}^W}, \mathbf{Q}_{\alpha^C}, 0^1, \cdots, 0^n\right)$, where each $0^i$ corresponds to the i-th map point.

### 2.2.3 Measurement Model

The measurements, $\mathbf{z}_k = \mathbf{h}\left(\mathbf{x}_k\right)$, are provided by a pinhole camera:

$$\mathbf{h} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 - \frac{f}{d_x}\frac{h_x^C}{h_z^C} \\ v_0 - \frac{f}{d_y}\frac{h_y^C}{h_z^C} \end{pmatrix} \tag{2.19}$$

where $u$, $v$ are the pixel coordinates of the observation in the image. $u_0$, $v_0$, $f$, $d_x$, $d_y$ are the camera intrinsic parameters corresponding to the principal point, the focal length, and the pixel size. $\mathbf{h}^C = \left(h_x^C, h_y^C, h_z^C\right)^\top$ is the vector joining the current camera location with the observed map feature, expressed in the camera frame which. For Euclidean parametrization it is:

$$\mathbf{h}^C = \mathbf{R}^{CW} \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} - \mathbf{r}^{WC} \Bigg) . \tag{2.20}$$

In the case of ID, $(X_i, Y_i, Z_i)^\top$ in Equation (2.20) are replaced by Equation (2.12).

Equation 2.19 gives the 2D image coordinates assuming a pure projective model. Therefore, in order to compensate the lens radial distortion, a two-parameter distortion model [MBM01] is applied. In this model, the ideal projective coordinates $\mathbf{h} = (u, v)^\top$ are recovered from the real distorted ones $\mathbf{h}_d = (u_d, v_d)^\top$:

$$\mathbf{h} = \begin{pmatrix} u_0 + (u_d - u_0)(1 + \kappa_1 r_d^2 + \kappa_2 r_d^4) \\ v_0 + (v_d - v_0)(1 + \kappa_1 r_d^2 + \kappa_2 r_d^4) \end{pmatrix} \tag{2.21}$$

$$r_d = \sqrt{\left(d_x\left(u_d - u_0\right)\right)^2 + \left(d_y\left(v_d - v_0\right)\right)^2} \tag{2.22}$$

where $\kappa_1$ and $\kappa_2$ are the radial distortion coefficients. The distorted coordinates are computed from the ideal ones as follows:

$$\mathbf{h}_d = \begin{pmatrix} u_0 + \frac{(u-u_0)}{(1+\kappa_1 r_d^2+\kappa_2 r_d^4)} \\ v_0 + \frac{(v-v_0)}{(1+\kappa_1 r_d^2+\kappa_2 r_d^4)} \end{pmatrix} \tag{2.23}$$

$$r = r_d(1+\kappa_1 r_d^2+\kappa_2 r_d^4) \tag{2.24}$$

$$r = \sqrt{(d_x\,(u-u_0))^2 + (d_y\,(v-v_0))^2} \tag{2.25}$$

Notice that $r$ is available from (2.19, 2.25), but $r_d$ must be numerically solved from (2.24). Finally, Equation (2.23) is used to compute the distorted point.

Regarding the covariance of the measurement noise ($\mathbf{R}_k$), it corresponds to the image measurement error covariance and is assumed to be a diagonal matrix.

### 2.2.4 Data Association & Map Management

The EKF prediction $\hat{\mathbf{x}}_{k|k-1}$ (2.4, 2.5), provides a prior over the current pose which is used to restrict the search of visual feature correspondences. This is known as active search and has two advantages due to the limited search area. First, it allows the system to run in real time, and second, it reduces the chance of spurious matches.

Additionally, the EKF prediction also provides an estimate for the relative pose of every map point with respect to the camera. This prediction is accurate enough to synthesize in a patch the point image appearance, compensating for rotation and scale variations along the sequence. Therefore, the combination of the FAST feature extractor [RD05] and simple patch correlation is used to extract and recognize the map features because it is cheap and performs satisfactorily. Besides, this combination is favored in the particular case of laparoscopy where, due to the small depth variation of the abdominal cavity and the limited laparoscope movements (it only pivots and slides over the fulcrum), features do not undergo severe perspective changes. According with this, each map point is identified by an $11 \times 11$ pixel planar texture patch extracted when the point is first observed, being unnecessary to resort to expensive invariant descriptors and extractors which would be overkill in SLAM.

Data association is performed by means of active search and synthesized point patches. Then, every map point is exhaustively searched inside the elliptical region defined by its innovation (2.6, 2.7) on the current image by means of normalized correlation with its synthesized patch (Figure 2.2). The pixel scoring highest, $\mathbf{z}_i$, if over a threshold, is selected as the match in the

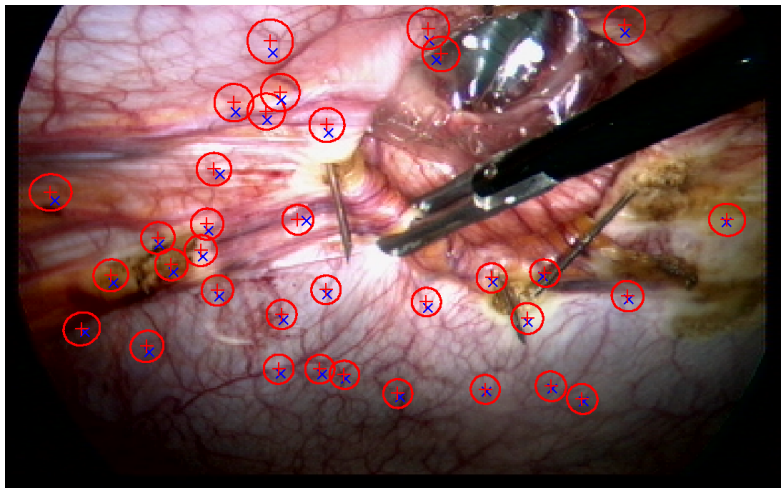


Figure 2.2: Individually compatible –IC– matches. Each map point is predicted in a location on the image (+) along with its innovation (elliptical region). Its corresponding measurement (×) it sought inside the elliptical region by means of correlation with the synthesized patch of the point.

new image. This stage produces the set of putative individual compatible (IC) matches:

$$\mathbf{z}_k^{IC} = (\mathbf{z}_1, \cdots, \mathbf{z}_{m_k})^\top \tag{2.26}$$

corresponding to some of the visible map points.

These IC matches are assumed to be correct and used in the EKF update stage (2.8, 2.9) where they feed the estimation. This is the most expensive step in terms of computational cost with a quadratic computational cost in the map size ($\mathcal{O}(n^2)$).

Regarding map management, the feature initialization criterion is targeted to keep in the field of view a predetermined number of visible features. When the number of visible features in the camera field of view is less than a threshold, features are initialized within a randomly located window favoring less populated areas (image regions with few or no map features). Each new feature is extracted from a new window and initialized in the map. New features are encoded in ID and, as the estimation improves, converted to Euclidean.

A feature is removed from the map if it is repeatedly predicted to be in the image but it is not successfully matched. The reobservation rate is predefined to be higher than 40% for the case of laparoscopic sequences.

## 2.3 Robust Data Association: JCBB

The computation of reliable correspondences from sensor data is at the core of most estimation algorithms in robotics, and in EKF in particular. The search for correspondences, or data association, is usually based, in a first stage, on comparing local descriptors of salient features in the measured data. The ambiguity of such local description usually produces incorrect correspondences (spurious) at this stage. The next stage assumes that the previous one has produced a perfect set of putative matches; however, if one or more of the matches are spurious, the whole estimation process might become wrecked. In order to avoid this situation, robust data association is necessary.

Robust data association is basically a search problem in the space of observation-feature correspondences. Given $m$ observations and a map with $n$ features, the problem consists in traversing a $m$-height $(n+1)$-ary tree (it includes the possibility that a measurement can be spurious; the size of this space is $(n+1)^m$) looking for the best set of correspondences between the measurements and the map features. In order to reduce the size of the tree, the correspondences are selected by means of a test of compatibility between the measurements and the map points, and a selection criterion choosing the best pair among the set of all compatible matches for each measurement. Finally, robust methods analyze the consistency of these pairings against a global model assumed to be generating the data, and discarding as spurious any pair that does not fit into it.

In the case of EKF-SLAM, the quality of the SLAM reconstruction strongly depends on the data association accuracy. Active search along with normalized patch correlation extract a set of IC matches; however, it is not guaranteed that this set is free of spurious matches. Nevertheless, doing so, the space of observation-feature correspondences is bounded to the matches that are the best IC matches. This dramatically decreases the size of the space to $2^m$ (the IC and the spurious possibilities), although the computational complexity continues being exponential. The analysis of consistency is carried out by Joint Compatibility Branch and Bound (JCBB) [NT01], which is a state-of-the-art robust data association method within the EKF-SLAM and has already been successfully used in visual [Cle+07; WKR07] and non-visual [FNL02] SLAM.

JCBB traverses the *2*-ary tree looking for the maximum set of jointly compatible matches. Given the set of jointly compatible matches $\mathcal{H}_{i-1} = \{m_1, \cdots, m_{i-1}\}$, a new match $m_i$ is jointly compatible if the set $\mathcal{H}_i = \{\mathcal{H}_{i-1}, m_i\}$ agrees with Equation (2.30). The JCBB consistency equations

are:

$$\nu_{\mathcal{H}_i} = (\mathbf{z}_{m_1} - \mathbf{h}_{m_1}(\mathbf{y}_{m_1}), \cdots, \mathbf{z}_{m_i} - \mathbf{h}_{m_i}(\mathbf{y}_{m_i}))^\top \quad (2.27)$$

$$\mathbf{S}_{\mathcal{H}_i} = \begin{pmatrix} \mathbf{S}_{m_1} & \cdots & \mathbf{S}_{m_1,m_i} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{m_1,m_i} & \cdots & \mathbf{S}_{m_i} \end{pmatrix} \quad (2.28)$$

$$\mathcal{D}_i = \nu_{\mathcal{H}_i}^\top \mathbf{S}_{\mathcal{H}_i}^{-1} \nu_{\mathcal{H}_i} \quad (2.29)$$

$$\mathcal{JC}_i = \mathcal{D}_i < \chi^2_{\alpha,2i} \quad (2.30)$$

where $\mathbf{z}_m$ and $\mathbf{h}_m(\mathbf{y}_m)$ are the point observation and the measurement model applied to the point $\mathbf{y}_m$. $\nu_{\mathcal{H}_i}$ and $\mathbf{S}_{\mathcal{H}_i}$ are the innovation and its covariance for the possible jointly compatible set $\mathcal{H}_i$. Finally, $\mathcal{D}_i$ and $\mathcal{JC}_i$ are the joint compatibility score (innovation Mahalanobis distance) and the joint compatibility test. In this test, $\alpha$ is the desired confidence level (typically $\alpha = 95\%$), and $2i$ are the degrees of freedom because each monocular SLAM match has 2 measurements $(u,\ v)$.

JCBB exploits the maximum-set criterion to bound the search inside the tree. When a node is reached, the maximum number of non-spurious matches that can be established from this node is counted. If this number is lower than the best pre-selected maximum set, the node is not explored. Figure 2.3 shows a small example with three matches. JCBB explores the green branch discovering that the maximum number of spurious matches is one. Automatically, all red branches are pruned because they contain two or more spurious matches, and hence they will not be explored. Blue branches have to be analyzed because they, theoretically, also have one spurious match. However, some of them may not be jointly compatible. In the case of one or more being jointly compatible –i.e. passing the joint compatibility test (2.30)–, they compete with the green branch. The final selected branch is that with the lowest joint compatibility score (2.29).

In conclusion, JCBB goes across the bounded tree detecting spurious matches based on a predicted probability distribution over the measurements. It does so by extracting from all the possible matches the maximum set that is jointly compatible with the multivariate Gaussian prediction. Consequently, JCBB entails two weaknesses: 1) its exponential computational cost in the number of spurious matches caused by traversing the tree, that causes JCBB to work only in real time for moderate spurious rates; and 2) its accuracy, which is questioned because JCBB operates on the linearized predicted state which, presumably, does not correspond to the real state of the system. Both limitations are overcome in Chapter 3.

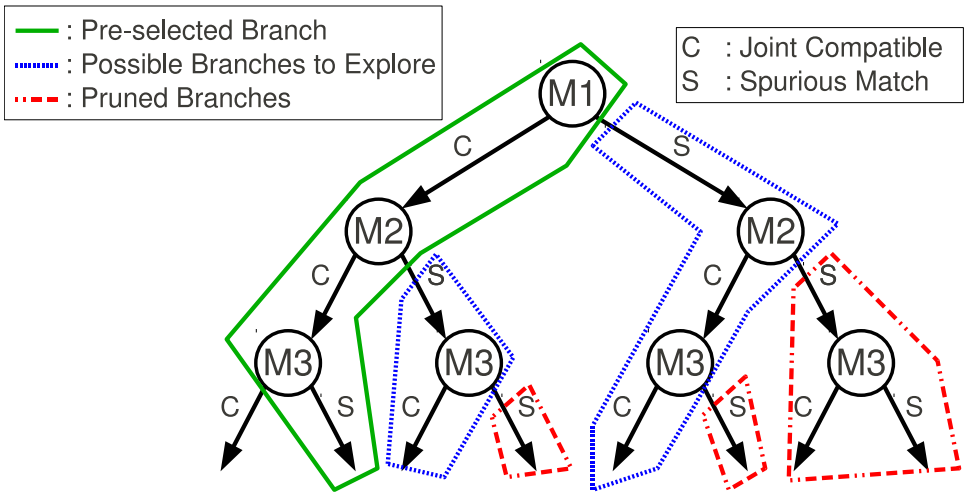



Figure 2.3: Example of JCBB working for 3 IC matches (each level of the tree corresponds to an IC match). The green branch is the pre-selected best branch with one spurious match. Red branches are pruned because they contain two or three spurious matches (they are worse than green branch). Blue branches theoretically contain one spurious match and they should be explored if they pass the joint compatibility test (2.30). In case of two or more branches being jointly compatible and have the same number of spurious than the pre-selected branch, the final selected branch is that with the lowest joint compatibility score (2.29).

## 2.4 SLAM Capabilities

Visual SLAM is an ideal environment to be used as a geometrical backbone to support useful information, above all for medical applications. The sparse 3D map of the scene and the camera motion provided by SLAM enable to do 3D distance measurements, insertions in augmented reality (AR), and to obtain photorealistic reconstructions.

### 2.4.1 Distance Measurement

Distance measuring will be a fundamental pillar to validate the SLAM accuracy in laparoscopy. This validation will be performed over ventral hernia repairs in Chapter 4. In this type of intervention it is mandatory to determine the hernia size in order to apply an adequate prosthetic mesh. Hence, this hernia size will be considered as ground truth and will be used to contrast

the dimensions provided by SLAM.

Monocular SLAM methods recover the map up to an unknown scale factor, implying that only relative distances can be measured. However, in practice, a known dimension can provide the unknown scale factor, and hence real distances may be recovered. Given the probabilistic nature of the SLAM map, the distance estimates are accompanied by an error estimate. Relative distances, along with the corresponding error estimates, can be computed in real time while exploring a scene.

From the 3D map, up to a scale factor, and an element, e.g. a laparoscopic tool, with two points $(r_1, r_2)$ inside the map whose relative distance is known, $s$, the real distance between two other map points $(i, j)$ is:

$$d(i,j) = s\frac{d_m(i,j)}{d_m(r_1,r_2)} \tag{2.31}$$

where $d_m(i,j)$ and $d_m(r_1,r_2)$ are the Euclidean distances between points $(i,j)$ and reference points $(r_1,r_2)$, respectively, measured in the SLAM map.

As the distance is a function of the SLAM state vector, $\mathbf{x}$, the covariance of the distance estimation can be propagated linearly from the SLAM covariance by means of the corresponding Jacobian matrix, $\mathbf{J}$:

$$\mathbf{J} = \frac{\partial d(i,j)}{\partial \mathbf{x}} \tag{2.32}$$

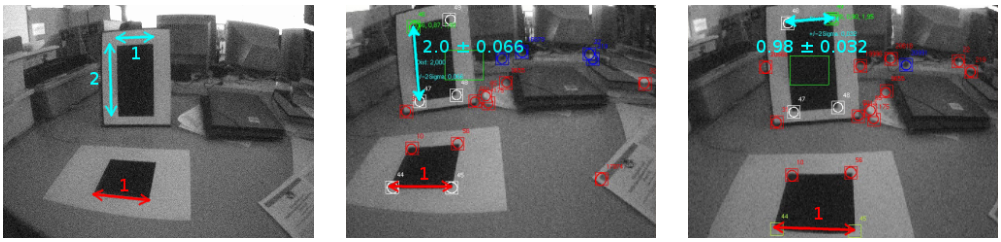
$$\mathbf{x} = \left(\mathbf{x}_v^\top, \mathbf{y}_1^\top, \cdots, \mathbf{y}_{r1}^\top, \cdots, \mathbf{y}_{r2}^\top, \cdots, \mathbf{y}_i^\top, \cdots, \mathbf{y}_j^\top, \cdots\right)^\top. \tag{2.33}$$

Since $d(i,j)$ only depends on $i,\ j,\ r_1$ and $r_2$, and $\mathbf{J}$ is sparse, reduced Jacobian $(\mathbf{J}_r)$ and covariance $(\mathbf{P}_r)$ matrices are used instead of the full matrices to compute the measurement error estimate $(\sigma_d^2)$:

$$\sigma_d^2 = \mathbf{J_r}\mathbf{P_r}\mathbf{J_r}^\top \tag{2.34}$$

$$\mathbf{P}_r = \begin{pmatrix} \mathbf{P}_{\mathbf{y}_{r1}\mathbf{y}_{r1}} & \mathbf{P}_{\mathbf{y}_{r1}\mathbf{y}_{r2}} & \mathbf{P}_{\mathbf{y}_{r1}\mathbf{y}_i} & \mathbf{P}_{\mathbf{y}_{r1}\mathbf{y}_j} \\ \mathbf{P}_{\mathbf{y}_{r2}\mathbf{y}_{r1}} & \mathbf{P}_{\mathbf{y}_{r2}\mathbf{y}_{r2}} & \mathbf{P}_{\mathbf{y}_{r2}\mathbf{y}_i} & \mathbf{P}_{\mathbf{y}_{r2}\mathbf{y}_j} \\ \mathbf{P}_{\mathbf{y}_i\mathbf{y}_{r1}} & \mathbf{P}_{\mathbf{y}_i\mathbf{y}_{r2}} & \mathbf{P}_{\mathbf{y}_i\mathbf{y}_i} & \mathbf{P}_{\mathbf{y}_i\mathbf{y}_j} \\ \mathbf{P}_{\mathbf{y}_j\mathbf{y}_{r1}} & \mathbf{P}_{\mathbf{y}_j\mathbf{y}_{r2}} & \mathbf{P}_{\mathbf{y}_j\mathbf{y}_i} & \mathbf{P}_{\mathbf{y}_j\mathbf{y}_j} \end{pmatrix} \tag{2.35}$$

Figure 2.4 shows a measurement experiment over two planar patterns. The first pattern is a black square which is used as reference and in which one of its edges defines the reconstruction scale considered to be the unit –the two edge corners are the reference points $(r_1, r_2)$– (the red double arrow in Figure 2.4a). The second pattern is a black rectangle whose dimensions, relative to the defined scale, are $2 \times 1$ (cyan double arrows in Figure 2.4a). The




(a) Pattern measurement ground-truth (cyan) and reconstruction scale (red).

(b) Vertical estimated measurement along with $2\sigma$ error (cyan).

(c) Horizontal estimated measurement along with $2\sigma$ error (cyan).

Figure 2.4: Pattern measurement. Red arrow corresponds with the reconstruction scale. Cyan arrows correspond with the dimensions to be measured.

experiment consists in estimating the dimensions of the rectangular pattern, relative to the scale, along with their error by means of SLAM. Therefore, both patterns are located in two different planes and a sequence is gathered and processed with monocular SLAM. Figures 2.4b and 2.4c show that the estimated dimensions are $2.0 \pm 0.066$ and $0.98 \pm 0.032$. Hence, it can be concluded that both estimations are accurate and precise. This experiment, which can be found in the video [GGc], was fundamental in order to successfully communicate to the surgeons the potential of the visual SLAM methods for in-body laparoscopic imagery.

### 2.4.2 Augmented Reality

AR annotations in endoscopic images need accurate real-time estimates for the live camera motion with respect to the observed scene. Monocular SLAM based only on images gathered by a camera has proven capable of providing camera motion in real-time at 30 Hz for rigid scenes [Dav+07; KM07]. AR is useful in laparoscopic surgery because it enables to visualize notations and to fuse other modal images, such as 3D models of CT or MR, with laparoscopic images live during surgery.

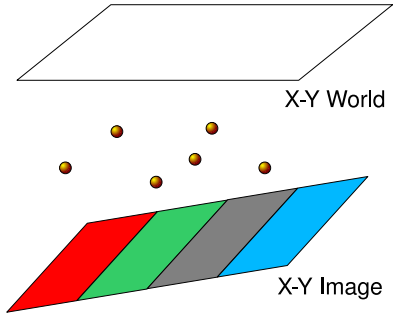
### 2.4.3 Photorealistic Reconstruction

The SLAM map also allows to build a mesh of triangular elastic textured tiles on it. This is a generalization for 3D scenes of the mosaic method proposed in [Civ+09b]. The tiles are defined by a standard 2D Delaunay triangulation over a projection of the 3D map on the absolute XY plane (XY plane in the absolute reference $W$). Each 3D triangle texture is gathered from the images

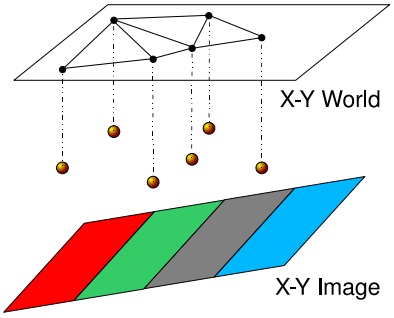
that observe the complete corresponding triangle. Figure 2.5 sketches the photorealistic modeling process.

Since triangulation is a live process –map points, and consequently triangles, are continuously created, erased and their estimates changed–, maintenance operations are performed to deal with new and deleted triangles as the SLAM estimation evolves, and to take textures from the images for the triangles.

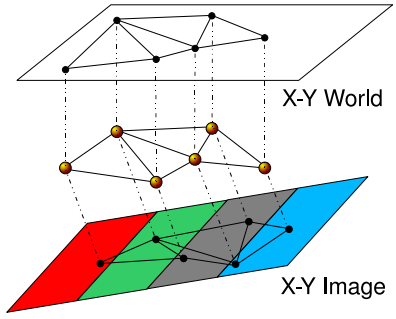
In the case of laparoscopy, this real-time photorealistic modeling process eases the 3D cavity visualization. The textured 3D model allows the synthesis of a panorama that expands the limited field of view (FoV) of the laparoscope.



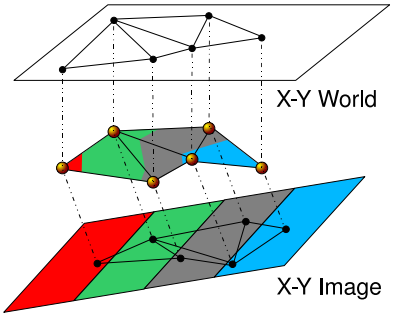

(a) Features in 3D, the current image and the world X-Y plane.




(b) Features projected onto a X-Y plane and triangulation on this plane.




(c) Triangulated image backprojection to obtain textures.




(d) Final photorealistic reconstruction.

Figure 2.5: Steps of the photorealistic reconstruction.

## 2.5 EKF-SLAM in Laparoscopy. A Proof of Concept

The first contribution of this thesis is to prove the feasibility of using monocular visual SLAM algorithms with real monocular laparoscopic sequences. In order to use SLAM, two assumptions have been made: the abdominal cavity is rigid, and the laparoscope undergoes a smooth and non-pure rotational motion. These conditions are fulfilled by a number of medical applications, such as laparoscopic ventral hernia repairs. The chosen algorithm is one of the leading-edge monocular SLAM algorithms (EKF + JCBB).

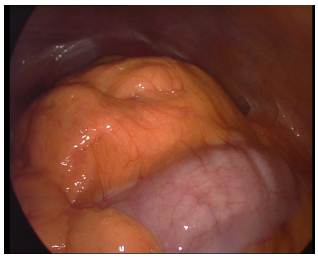
SLAM has been applied over two laparoscopic abdominal exploration sequences. The primary result has been a sparse up-to-scale 3D map composed of salient points –features– of the observed cavity for each sequence. This SLAM map has shown to be adequate to support 3D distance measurements along with the measurement error.

Additionally, the map has been used as a backbone for real-time photorealistic modeling to ease the 3D cavity visualization. The textured 3D model allows to synthesize a panorama that expands the laparoscope FoV. Finally, since the camera motion with respect to the 3D map is accurately known in real time, AR annotations can be supported live in medical sequences.

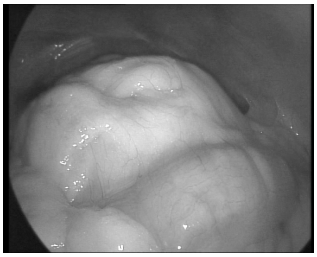
### 2.5.1 Image Processing

Monocular SLAM in robotics uses a correlation score based on luminance, neglecting color information. As laparoscope captures color images, a procedure to convert color images to B&W must be applied. One of the quickest procedures is to select one out of the three channels that compose a color image. The channel selected must preserve high frequencies in order to ease the performance of the feature extractor (features are high frequency components). As can be seen in Figure 2.6, in contrast to the red channel, green and blue channels preserve high frequencies, and hence both channels are possible candidates. Finally, the green channel has been the preferred one because visually it seems to contain a richer contrast than the blue one and a nice texture to produce distinctive patches for recognition.

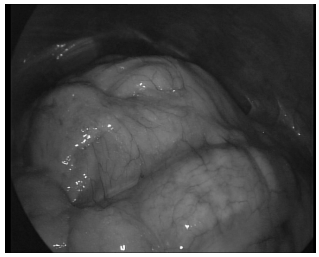
On the other hand, human tissues are prone to produce reflections caused by illumination. These reflections can erroneously fire the feature extractor and then incorrect map points could be initialized. Additionally, as in laparoscopic scenes the light source is fixed to the laparoscope, when it is moved, the light source is also moved and reflections change producing erroneous data associations. In order to avoid both situations, reflections are removed



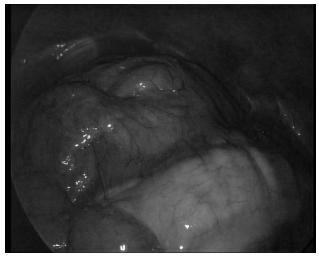
(a) Color



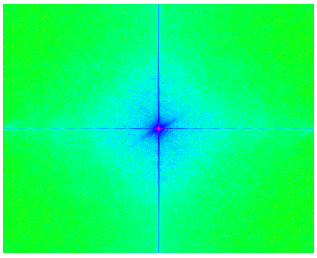
(b) Red channel



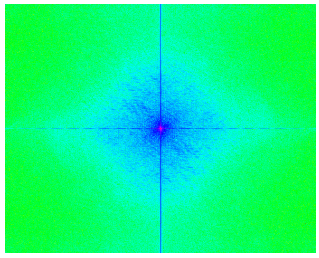
(c) Green channel



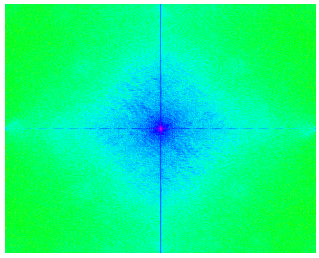
(d) Blue channel



(e) Red channel Fourier spectrum



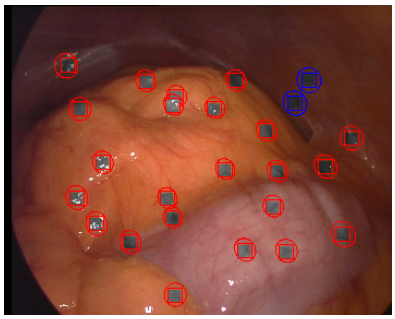
(f) Green channel Fourier spectrum



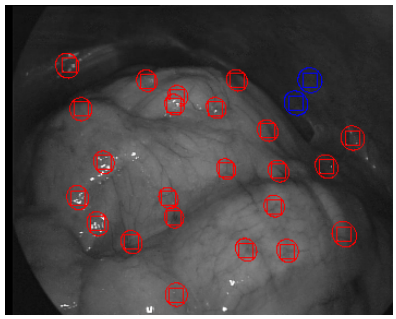
(g) Blue channel Fourier spectrum

Figure 2.6: Color frame and its decomposition in color channels along with their corresponding Fourier spectrums. The red channel (2.6b) has very light areas with small contrast (no high frequency details), as shown by its Fourier spectrum (2.6e). On the contrary, green (2.6c) and blue channels preserve more high frequency details (2.6f, 2.6g). Visually, the green channel seems to have more contrast than the blue.

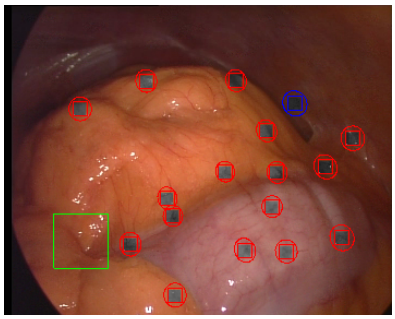
assuming that they produce pixels with a high luminance. If any pixel in a patch around a detected feature is over a threshold (200 over 255), this feature is rejected. Figure 2.7 shows a frame of a laparoscopic sequence. Figures 2.7a and 2.7b, show how specularities have been detected and initialized as map points. However, these specularity features have been rejected in Figures 2.7c and 2.7d.



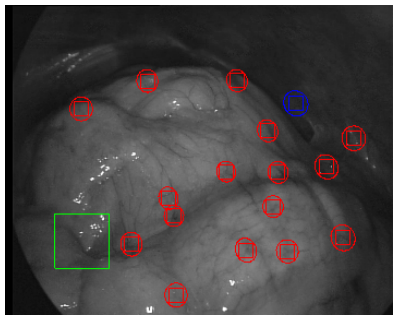
(a) Color frame with specularities.



(b) B&W frame with specularities.



(c) Color frame without specularities.



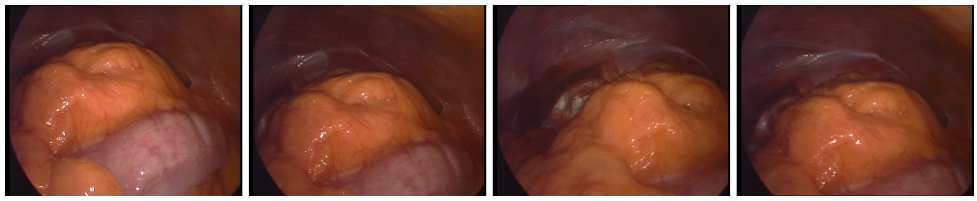
(d) B&W frame without specularities.

Figure 2.7: Four figures corresponding to the same frame from the same sequence processed without reflection filtering (2.7a, 2.7b) and with reflection filtering (2.7c, 2.7d). In the unfiltered case, the feature detector has detected some specularities (white points) as valid features. These specularities are not present in the filtered case.

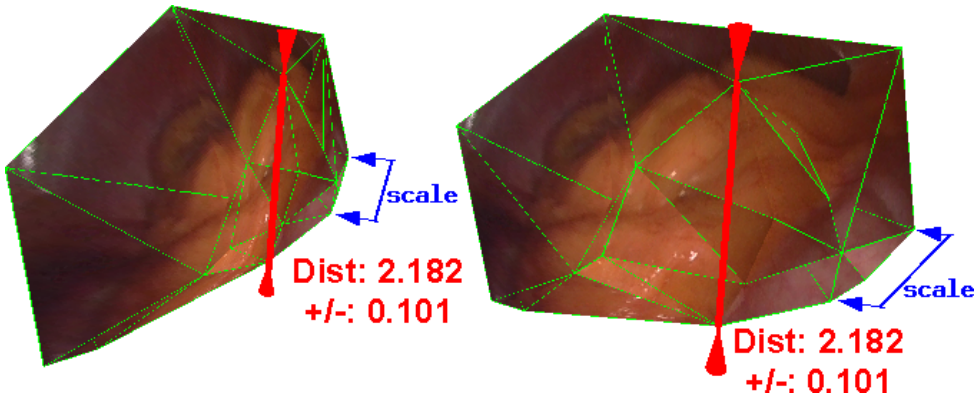
### 2.5.2 Experimental Results

Experimental validation is performed on real images 360x288@25 Hz gathered from a hand-held monocular laparoscope observing two abdominal cavities (341 and 186 frames respectively). The goal of the validation is to prove the feasibility of using SLAM in laparoscopy. To this end, three experiments over these two sequences were carried out showing SLAM performing photorealistic reconstructions, distance measurements, and AR annotations.

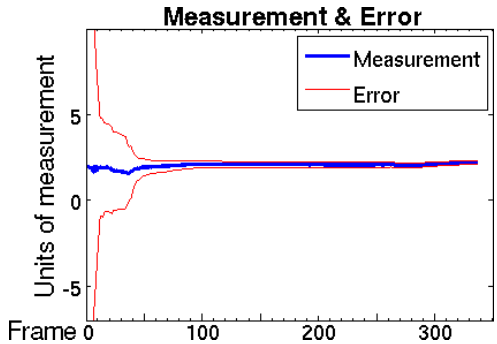
The sequences were the only data input to the algorithm, achieving real-time performance in all the experiments measuring up to 25 features. Laparoscope intrinsic parameters were calibrated using a standard planar pattern calibration method, based on Zhang's initial solution [Zha00], followed by Bundle Adjustment.



(a) Some frames of a laparoscopic sequence of an abdominal cavity exploration (341 frames).




(b) Photorealistic reconstruction with a measurement between two points of the organ. The scale is defined with two other organ points.




(c) Historical evolution of the measurement and its error. Notice that the error reduction as the camera moves and gathers information from different points of view providing higher parallax.

Figure 2.8: Hand-held laparoscope sequence of an abdominal cavity exploration (341 frames).

In the case of the 341-frame sequence (Figure 2.8), from which several frames are shown in Figure 2.8a, a 3D distance measurement experiment was performed (Figure 2.8b). The sequence corresponds to a laparoscope

exploration inside the abdominal cavity. No tool was inserted; therefore, in order to make the distance measurement, two arbitrary points were marked as reference which define the unity of the scale factor (blue arrows in Figure 2.8b). Another two arbitrary points were selected and their relative distance was measured (the red arrow in Figure 2.8b). The distance along with its error were computed relative to the defined scale. Assuming the scale was the real scale, the conversion to real distances is immediate according to (2.31, 2.35). Figure 2.8c shows the estimate history both for the distance and the error. Initially, error uncertainty is large, but as the camera translates and the scene is seen with parallax, point location error decreases and consequently the distance error decreases too. Since the uncertainty is computed in real time, visual feedback gives the surgeon information on how to move the camera in order to reduce the distance error. This experiment can be found in the video [GGb].

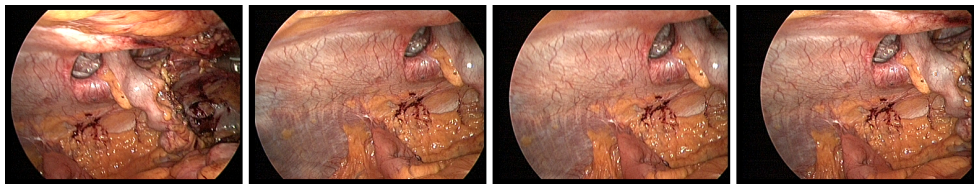
In the case of the 186-frame sequence (Figure 2.9), where several frames are shown in Figure 2.9a, an AR reality experiment was carried out. The sequence corresponds to a laparoscopic abdominal exploration during a human ventral hernia repair. Since the 3D map and the camera location with respect to the map are available in real time, it is possible to anchor AR annotations to map points. Figure 2.9b shows an AR cylinder both in 3D and superimposed on the live laparoscopic image. As the virtual insertions are fixed to the map, they can be observed at their real location even when they are out of the camera FoV. This experiment can be found in the video [GGa].

For both sequences a textured triangular mesh model was obtained (Figures 2.8b and 2.9b). Despite the sparse map being composed of a reduced number of points, the 3D live photorealistic models provide an easy understanding of the 3D cavity structure. Videos [GGb] and [GGa] show both photorealistic reconstruction processes. Besides, video [GGd] shows in more detail the photorealistic reconstruction for the 186-frame sequence.

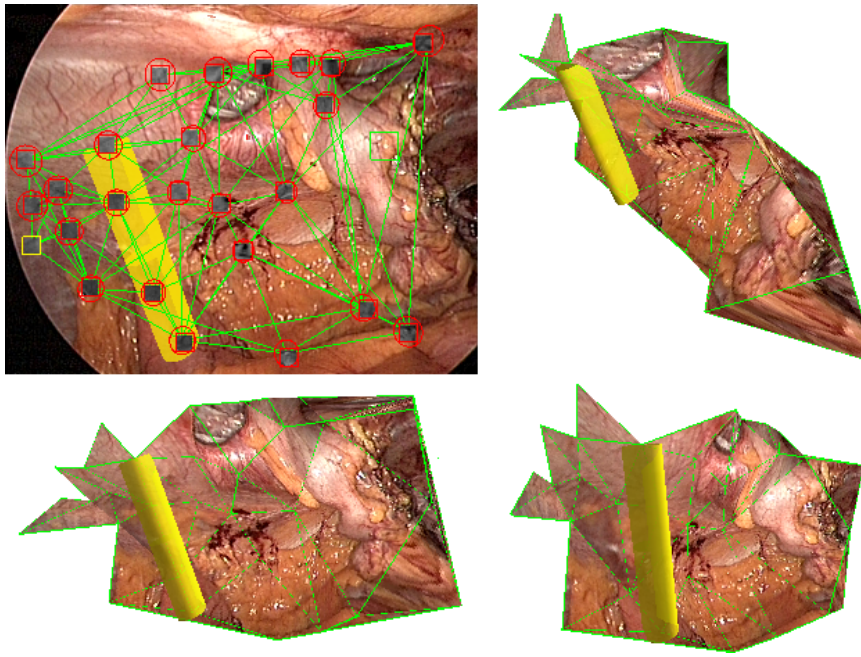
It is worth noting that these experiments, corresponding with the first monocular SLAM experiments over real human laparoscopic sequences in the literature, along with the experiment of Figure 2.4, were crucial to speak with surgeons and to prepare an intervention where SLAM could be relevant and easily validated.

## 2.6 Conclusions

Unlike previous works where human cavity reconstructions are obtained using fixed or moving stereo endoscopes, or monocular endoscopes are used on



(a) Some frames of a laparoscopic sequence of an abdominal cavity exploration (186 frames).



(b) AR cylindrical insertion back projected in live laparoscopic video and the photorealistic 3D model recovered.

Figure 2.9: Hand-held laparoscope sequence of an abdominal cavity exploration (186 frames).

phantoms, this chapter presents the first results of using monocular SLAM in real human laparosocopic surgeries. The proof of concept with real laparoscopic imagery has shown the potential of this robotics technique in the medical field, however, although the combination of EKF + JCBB has proven to be very promising, some problems must still be overcome.

Monocular SLAM recovers a 3D map of the cavity and the trajectory followed by the laparoscope, in real time at 25 fps, using the laparoscopic sequence as only input, opening new venues for the surgery of the future. SLAM may be exploited to increase synthetically the FoV by means of photorealistic reconstructions computed in real time, to do internal distance measurements

along with their error, and to insert AR notations that facilitate the operation. All these capabilities have been shown on real images gathered from a monocular laparoscope observing the abdominal cavity.

After testing the feasibility of EKF + JCBB in laparoscopic imagery, several issues are still open. The current algorithm assumes: 1) scene rigidity; 2) smooth laparoscope motion; 3) that the laparoscope is always inside the cavity; and 4) low motion clutter and occlusions. These assumptions do not hold in general medical scenes: non rigidity is almost prevalent, sudden motions are frequent, the laparoscope is extracted and reinserted inside the cavity, and tools cause a significant motion clutter and occlusions. They produce tracking failures and an increment of the number of spurious matches. Figure 2.10 shows these drawbacks extracted from a real laparoscopic sequence; it depicts the size map, the inlier and the outliers matches, and a tracking failure caused by laparoscope extraction and reinsertion (blue dashed rectangle); it can be seen how in some frames the number of spurious matches is similar to the number of inliers. Regarding tracking failure, relocation algorithms such as [WKR07; CN08] recover the track of the system when it is lost providing robustness to the whole system. Regarding spurious matches, JCBB works fine in man-made environments (mobile robotics scenes) where the scene is completely static and thus the number of spurious low. However, its exponential computational complexity causes this algorithm to not run in real time when several spurious are present, as shown in Section 3.2. Additionally, JCBB uses the EKF prediction to detect spurious matches, which entails linearization errors. If the linearization is not a good approximation, the reconstruction error will degrade as the estimation evolves. Therefore, a more efficient data association is mandatory for using monocular SLAM in laparoscopic surgery. These questions are addressed in Chapter 3.

After the system improvement of Chapter 3, an extensive validation of the accuracy is necessary for showing the system feasibility in medical imagery. This validation is performed in Chapter 4.

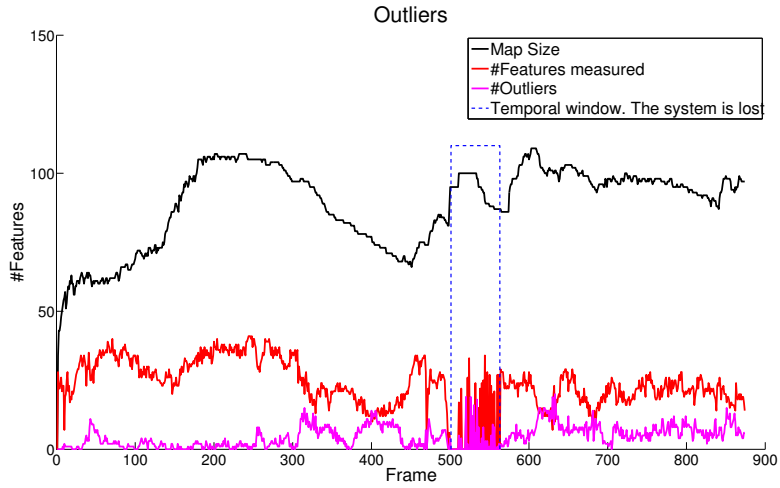



Figure 2.10: Map size –black–: the total number of map features. Inlier matches –red–: measured features. Spurious matches –magenta–: matches found inside the active search region (IC matches) but marked as spurious by robust data association. The blue dashed rectangle corresponds to frames where tracking was lost.

# 3 Robust Monocular SLAM

The EKF + JCBB combination shows several weaknesses that make it unfeasible in real laparoscopic surgeries.

On one hand, when the laparoscope suffers sudden motions, the laparoscope is extracted and reinserted into the abdominal cavity, the image is blurred, there are large occlusions, or the scene is deformed changing its appearance, the tracking will fail because no features will be matched in several consecutive frames (Figure 3.3). In order to avoid this situation, a robust relocalization algorithm is mandatory. The relocalization algorithm must detect loss of tracking and stop EKF integration to avoid a possible map corruption due to incorrect data associations, and then enable a recovery procedure. If tracking is lost, the relocalization must find matches between the current image and the already estimated map in a data-driven manner without assuming priors about the camera location with respect to the map. Randomized List Relocalization (RLR) ([WKR07]) is one of the best relocalization performers in visual SLAM and has been chosen and integrated in the system (Section 3.1).

On the other hand, JCBB presents a limitation concerning computational cost that makes this algorithm inappropriate to be used in laparoscopy. The Branch and Bound search, that JCBB uses for extracting the largest jointly compatible set of matches, has exponential complexity in the number of spurious matches. This complexity does not present a problem for small numbers of matches, but very large computation times arise when the number of spurious grows. In the case of laparoscopic sequences, this is a very common situation due to small tissue deformations, where matches are found inside the active search region but they are not jointly compatible. In addition to computational cost, JCBB entails a problem of accuracy. JCBB oper-

ates over the prediction for the measurements before fusing them. Such a probabilistic prediction comes from the linearization of the dynamic and measurement models and the assumption of Gaussian noise, so it will presumably not correspond to the real state of the system. Both limitations are greatly overcome with the replacement of JCBB in favor of the 1-Point RANSAC (1-PR) algorithm (Section 3.2). The computational complexity of 1-PR is linear in the state and measurement size and exhibits low cost variation with the number of outliers. Additionally, 1-PR operates over hypotheses after the integration of a data subset, which have corrected part of the predicted model error with respect to the real system.

Two contributions of this thesis have been: 1) the development and exhaustive validation of 1-PR (Section 3.2) reported in ([Civ+09a; Civ+10]); and 2) the application of the EKF + 1-PR + RLR combination in laparoscopy (Section 3.3) reported in [GGCM11] and [GG+14].

## 3.1 Relocalization

Active search is one of the system strengths, since it enables the system real-time operation, but it is also one of its weaknesses. The system works fine provided that the mapped features are found inside the elliptical search window. However, if the camera suffers from sudden motions, the image is blurred, there are large occlusions, or the scene is deformed, tracking will fail because no features will be matched within several frames.

In order to avoid this problem, the use of a relocalization system is a must. The ideal relocalization system should detect loss of tracking and stop EKF integration to avoid map corruption due to incorrect data associations, and then enable a recovery procedure. The tracking should deemed lost if all attempted matches in a frame have been unsuccessful, the camera pose uncertainty has grown too large, or if all the predicted mapped features are out of the predicted camera FoV. During the recovery procedure, the relocalization should find matches between the current image and the already estimated map in a data-driven manner without assuming priors about the camera localization with respect to the map.

Randomized List Relocalization (RLR), proposed in [WKR07] and based on Randomized Trees [LF06], is a cutting-edge feature-based relocalization in visual SLAM that complies with the requirements demanded to an ideal relocalization system.

RLR casts the image-to-map matching as a classification problem. When the system detects a tracking failure, a few thousand of the strongest FAST features [RD05], detected in the current image, are fed to the classifier to

find putative image-to-map matches.

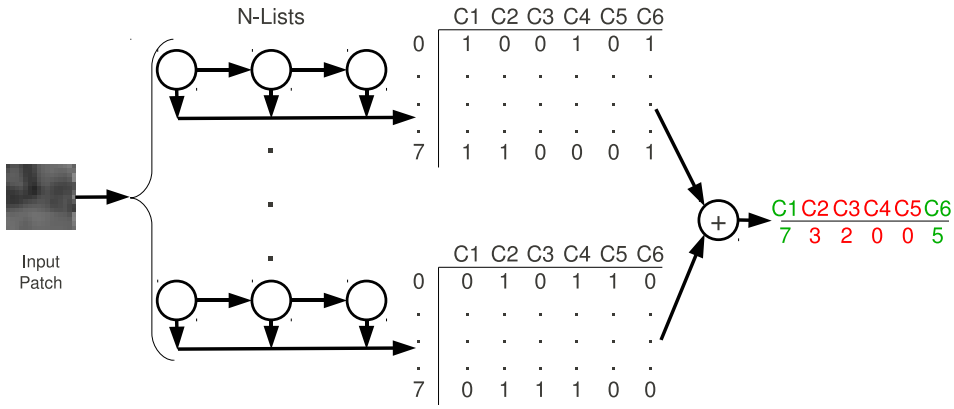



Figure 3.1: RLR classifier. Hypotheses selection for an input patch (putative matches) in RLR. The selected hypotheses are those whose value is greater than a threshold, in this example the threshold is 4.

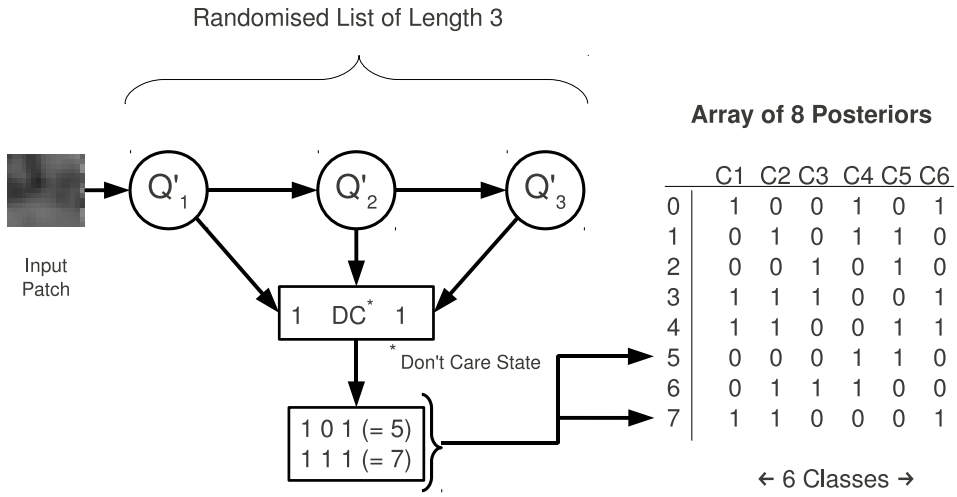



Figure 3.2: Randomised List Structure, example with a "don't care bit". The input patch is fed to the list. The test $Q'_2$ measures noise –it does not agree with (3.2)–, and then it is marked as "don't care" Finally two binary words are composed and used to recover their respective posteriors.

Internally, the RLR classifier is implemented as $N$ lists of $D$ sequential random binary tests and $C$ classes –one per map point– (Figure 3.1 shows an example of a small RLR classifier. Typical classifier values are: $N = 40$, $D = 18$, $C = [100 - 200]$. Figure 3.2 depicts a list with $D = 3$). For each list, the result of the tests forms a binary word which indexes into an array of $2^D$ posteriors. Each binary test $Q'_i$ compares the Gaussian-smoothed intensity values of the feature patch $I_\sigma(\cdot)$ at two different pixel locations $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\mathbf{Q'_i} = \begin{cases} 0, & if \quad I_\sigma(a_i) - I_\sigma(b_i) >= z_i \\ 1, & otherwise. \end{cases} \tag{3.1}$$

Both $\boldsymbol{a}$ and $\boldsymbol{b}$ are randomly determined when the binary test is created. Equation (3.1) has a $z_i$ term which is used for the purpose of not measuring noise and favoring repeatability in areas of uniform color. Each $z_i$ takes a random value in the range $[0 - 20]$ which is also fixed during the test creation. Besides, each test $Q'_i$ is explicitly checked to see if its result is close to a noise threshold according to:

$$|I_\sigma(a_i) - I_\sigma(b_i) - z_i| < th_{noise}. \tag{3.2}$$

If this is the case, the $i$-th bit is set to a "don't care" state (the test takes both "0" and "1" values). Thus, when the test word is formed, the scores for all possible values of the word are obtained from the array of posteriors achieving more noise tolerance. Figure 3.2 depicts a Randomised List structure with $D$=3 and 6 different classes. In this example, test $Q'_2$ agrees with (3.2) and hence two binary words are composed and their corresponding posteriors recovered.

The array of posteriors stores, for each entry, a binary score string of one bit per class (each map feature is a class). Then, when a class activates one array entry during training, its corresponding bit is set to 1. If the class has never activated an entry, its corresponding bit will be 0. The feature correspondence hypotheses are selected by counting the times that a class is present in all indexed posteriors of the $N$ lists. When a class appears in more than a *threshold* of posteriors, it is included in the set of potential feature correspondences (putative matches) that will be used during relocalization. As can be seen, map features may be similar to each other and then multiple feature correspondence hypotheses must be considered. Figure 3.1 depicts this procedure. In this figure, all recovered posteriors of the N lists are added and the classes *C1* and *C6* are selected as putative matches for the input patch because their sums are greater than the threshold established to 4.

Once the set of putative matches has been obtained, RANSAC is applied over this set in order to relocalize the camera with respect to the map. Camera location is hypothesized from three feature correspondences using the 3-Point-Pose (PnP) algorithm proposed in [FB81]. Each camera location hypothesis is rated according to how many other map features can be matched in the image. Once a good pose hypothesis is found, it is optimized in a "moving camera observing a fixed map" manner, and then the SLAM system is reinitialized. If the pose estimate is indeed close enough to the true estimate, then one or two fixed map EKF iterations are sufficient to refine the camera pose. It should be noticed that map integrity is fundamental not only for tracking, but also for relocalization.

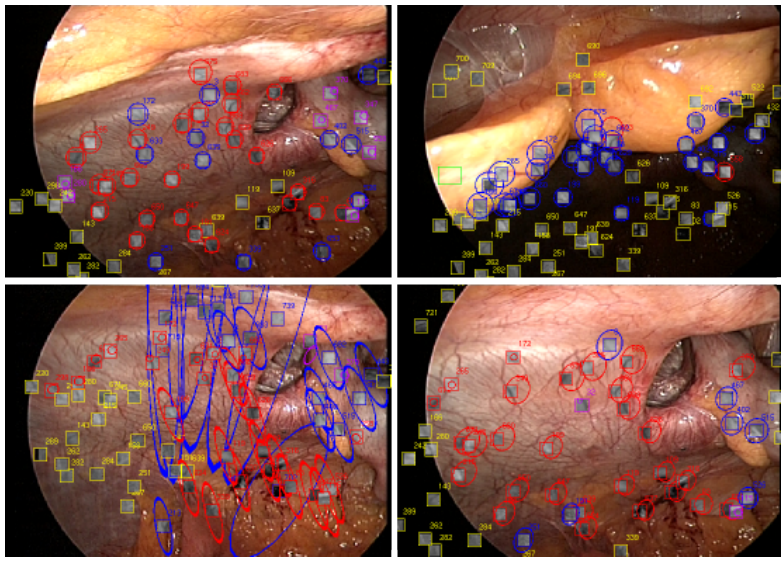
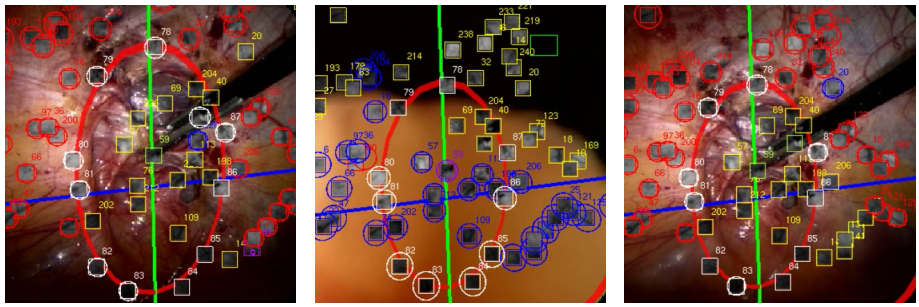


Figure 3.3: Example of relocalization after laparoscope extraction and reinsertion in a 874 frame laparoscopic sequence. Upper left, system just before tracking loss. Upper right, laparoscope partially out of the cavity. Lower left, unstable relocalization. Lower right, system after total tracking recovery.

Regarding the classifier training, a two-stage online procedure is applied for every map feature. First, at feature initialization, a new class is added in the classifier and 400 warped versions of the texture patch around the feature are GPU-synthesized from the image where the feature is first observed. The warped patches are used to train the classifier. The second stage harvests texture patches during EKF operation which are used for online training. RLR considers each class score independently, facilitating the continuous online training, since the classification rate of any class is not affected by the



(a) The system is on track. (b) A large occlusion causes the tracking loss. (c) Finally, the system relocalizes and tracking is recovered.

Figure 3.4: Example of relocalization after a large occlusion for the operation in Figure 4.9c.

addition of other classes. The classifier is also exploited for selecting the most distinctive features at initialization: only features scoring low in the classifier with respect to other features already in the map are eventually initialized. Doing so, the map features are trackable, locally salient and also distinctive for recognition and relocalization.

RLR has proven to be valid in the laparoscopic experiments performed in this thesis. Figure 3.3, corresponding to video [GGf] (0:38 - 0:49), shows a real laparoscopic example of a loss of tracking due to an extraction and reinsertion of the laparoscope and its posterior relocalization. Figure 3.4 shows another real laparoscopic example of a loss of tracking in this case caused by a large occlusion. This example can be found in video [GGh] (1:13 - 1:32).

## 3.2 1-Point RANSAC

A robust search for correspondences, or data association, generally operates by checking the consistency of the data against the global model assumed to be generating the data, and discarding as spurious any that does not fit into it. Among robust estimation methods, Random Sample Consensus (RANSAC) [FB81] stands out as one of the most successful and widely used, especially in the Computer Vision community. One contribution of this thesis is the integration of RANSAC into the EKF framework –1-Point RANSAC (1-PR).

As a motivation and in order to highlight the requirements and benefits

of the RANSAC, a simple 2D line estimation example with spurious data is used to explain the standard RANSAC algorithm (Figure 3.5) and, after that, its adaptation to the EKF framework with the proposed 1-PR algorithm (Figure 3.6). 1-PR is thoroughly detailed along this section and is shown as a practical matching algorithm.

Standard RANSAC starts from a set of data, 2D points in this simple example, and the underlying model that generates the data, a 2D line. In the first step, RANSAC constructs hypotheses for the model parameters and selects the one that gathers most support. Hypotheses are randomly generated from the minimum number of points necessary to compute the model parameters, which is two in the case of line estimation. Support for each hypothesis can be computed in its most simple form by counting the data points inside a threshold (related to the data noise), although more sophisticated methods have been used [TZ00].

Hypotheses involving one or more outliers are assumed to receive low support, as is the case in the third hypothesis in Figure 3.5. The number of hypotheses $n_{hyp}$ necessary to ensure that at least one spurious-free hypothesis has been tested with probability $p$ can be computed from this formula:

$$n_{hyp} = \frac{log\,(1-p)}{log\,(1-(1-\epsilon)^m)}\ , \tag{3.3}$$

where $\epsilon$ is the outlier ratio and $m$ the minimum number of data points necessary to instantiate the model. The usual approach is to adaptively compute this number of hypotheses at each iteration, assuming the inlier ratio is the support set divided by the total number of data points in this iteration [HZ04].

Data points that voted for the most supported hypothesis are considered clear inliers. In a second stage, clear inliers are used to estimate the model parameters. Individual compatibility is checked for each one of the rest of the points against the estimated model. If any of them is rescued as inlier, as happens in the example in Figure 3.5, the model parameters are re-estimated again in a third step.

Figure 3.6 illustrates the idea behind 1-PR in the same 2D line estimation problem. As the first key difference, the starting point is a data set and its underlying model, but also a prior probability distribution over the model parameters. RANSAC hypotheses are then generated based on this prior information and data points, differently from standard RANSAC hypothesis based solely on data points. The use of prior information can reduce the size of the data set that instantiates the model to the minimum size of one point, and it is here where the computational benefit of this method with

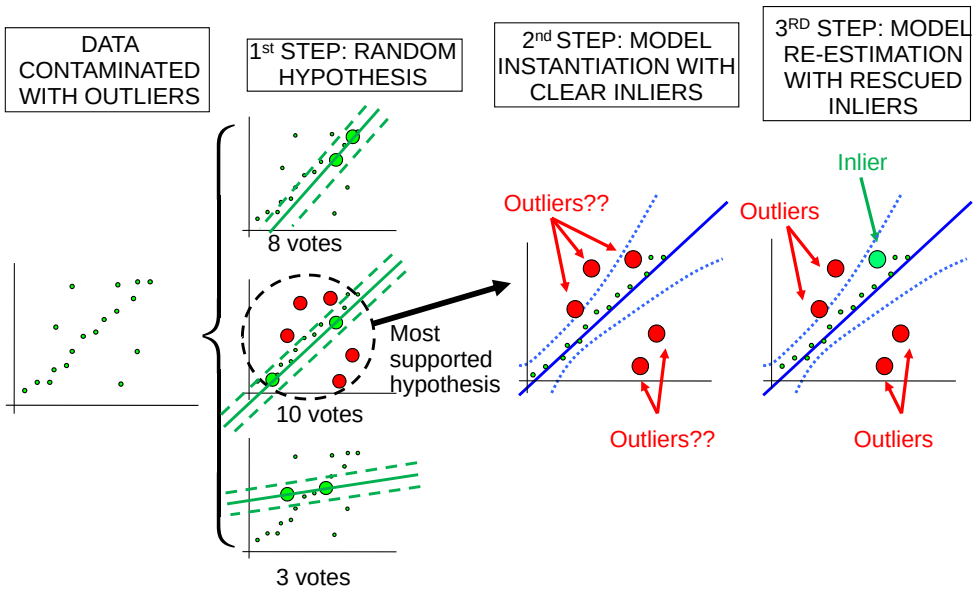



Figure 3.5: RANSAC steps for the simple 2D line estimation example: First, random hypotheses are generated from data samples of size two, the minimum to define a line. The most supported one is selected, and data voting for this hypothesis are considered inliers. Model parameters are estimated from those clear inliers in a second step. Finally, the remaining data points consistent with this latest model are rescued and the model is re-estimated again.

respect to RANSAC arises: according to Equation 3.3, reducing the sample size $m$ greatly reduces the number of RANSAC iterations and hence the computational cost.

The order of magnitude of this reduction can be better understood if instead of this simple estimation example, a real visual estimation application is used. According to [Nis04], at least five image points are necessary to estimate the 6 degrees of freedom (DoF) camera motion between two frames (so $m = 5$). Using Equation (3.3), assuming an inlier ratio of 0.5 and a probability $p$ of 0.99, the number of random hypotheses would be 146. Using the 1-PR scheme, assuming that probabilistic a priori information is available, the sample size $m$ can be reduced to one point and the number of hypotheses would be reduced to 7. Having an a priori probability distribution over the camera parameters is unusual in classical pairwise Structure from Motion (SfM) which assumes widely separated views [HZ04], and methods like standard RANSAC, which generate hypotheses from candidate feature matches,

are mandatory in this case. But in sequential SfM from video ([Dav03; KM08; Mou+09]), smooth interframe camera motion can be reasonably assumed and used to generate a prior distribution (prediction) for the image correspondences. For the specific case of the EKF implementation of sequential SfM, this prior probability is naturally propagated by the filter and is straightforwardly available.

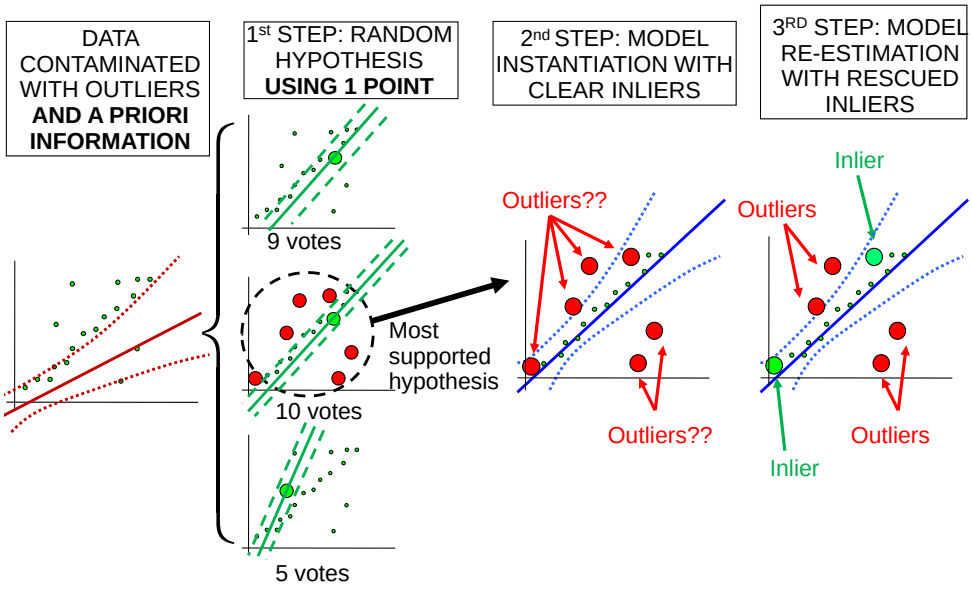



Figure 3.6: 1-PR steps for the simple 2D line estimation example: As a key difference from standard RANSAC, the algorithm assumes that an a priori probability distribution over the model parameters is known in advance. This prior knowledge allows to compute the random hypotheses using only 1 data point, hence reducing the number of hypotheses and the computational cost. The remaining steps do not vary with respect to standard RANSAC in Figure 3.5.

.

### 3.2.1 Related Work

#### Data Association

RANSAC [FB81] was introduced early in visual geometric estimation [TM93] and has been the preferred outlier rejection tool in the field. Recently, an important stream of research has focused on reducing the model verification

cost in standard RANSAC ([RFP08; CM08; Cap05; Nis05]) via the early detection and termination of bad hypotheses. The 1-PR algorithm proposed here is related to this stream in the sense that it also reduces the hypothesis generation and validation cost. Nevertheless, it does so in a different manner: instead of fast identification of good hypotheses among a large number of them, the number of hypotheses is greatly reduced from the start by considering the prior information given by a dynamic model.

Incorporating probabilistic information into RANSAC has rarely been discussed in the computer vision literature. Only Moreno *et al.* ([MNLF08]) have explored the case where weak a priori information is available in the form of probabilistic distribution functions.

More related to this method, the combination of RANSAC and Kalman filtering was proposed by Vedaldi *et al.* [Ved+05]. 1-PR might be considered a specific form of Vedaldi's quite general approach. They propose an iterative scheme in which several minimal hypotheses are tested; for each hypothesis, all the consistent matches are iteratively harvested; no statement about the cardinality of the hypotheses is made. Here, a definite and efficient method, in which the cardinality of the hypotheses generator size is 1, and the inlier harvesting is not iterative but in two stages, is proposed. The method is described in reproducible detail to deal efficiently with the EKF algorithm by splitting the expensive EKF covariance update in two stages in order to reach real time.

RANSAC using 1-point hypotheses has also been proposed in [SFS09] as the result of constraining the camera motion. While at least 5 points would be needed to compute monocular SfM for a calibrated camera undergoing general 6 DoF motion [Nis04], fewer are needed if the motion is known to be less general: as few as 2 points in [OM01] for planar motion and 1 point in [SFS09] for planar and nonholonomic motion. As a clear limitation of both approaches, any motion performed out of the model will result in estimation error. In fact, it is shown in real-image experiments in [SFS09] that although the most constrained model is enough for RANSAC hypotheses (reaching then 1-PR), a less restrictive model offers better results for motion estimation.

In the case of the 1-PR method, extra information for the predicted camera motion comes from the probability distribution function that the EKF propagates over time. The method presented is then, in principle, not restricted to any specific motion, being suitable for 6 DoF estimation. The only assumption is the existence of tight and highly correlated priors. This assumption is reasonable within the EKF framework since the filter itself only works in such circumstances.

Among non-RANSAC-based methods for data association, JCBB has

been the preferred technique within the EKF framework being successfully used both in visual [Cle+07; WKR07] and non-visual SLAM [FNL02]. As discussed in Section 2.3, JCBB extracts the maximum set of matches that is jointly compatible with the multivariate Gaussian prediction from all IC matches. Nevertheless, JCBB entails two limitations: its exponential computational cost in the number of measurements, and its lack of accuracy for operating on the linearized predicted state of the measurements, which are overcome by 1-PR. Regarding the former, the computational complexity of 1-PR is linear in the number of measurements with low cost variation in the number of spurious matches (outliers). Regarding the latter, JCBB operates with the prediction of the measurements before fusing them, in contrast, 1-PR, and RANSAC in general, operates after fusing a subset of them, which corrects part of the predicted model error with respect to the real system.

Two methods are also of interest for this work. First, Active Matching (AM) [CD08] which is a clear inspiration for 1-PR. In AM, feature measurements are integrated sequentially; the choice of a measurement, at each step, is driven by expected information gain; the results of each measurement in turn are used to narrow the search for subsequent correspondences. 1-PR can be seen as lying in the middle ground between RANSAC or JCBB, which obtain point correspondence candidates and then aim to resolve them, and AM with its fully sequential search for correspondences. The first step of 1-PR is very similar to AM confirming that integrating the first match highly constrains the possible image locations of other features but, afterwards, both algorithms diverge. A problem with AM is the unreasonably high computational cost when scaling to large numbers of feature correspondences per frame (1-PR has much better properties in this regard), though an improvement to AM has also addressed this issue in a different way [Han+10].

The second method is Randomized Joint Compatibility proposed by Paz *et al.* [PTN08]. This basically randomizes the jointly compatible set search by avoiding the complete Branch and Bound search. At the first step, an initial small set of jointly compatible inliers is obtained via Branch and Bound search in random sets. Then, the joint compatibility of each remaining match is checked against the initial set. Although this approach lowers the computational cost of the JCBB, it still faces the accuracy problems derived from the use of the predicted measurement function before data fusion.

### Benchmarking

Carefully designed benchmark datasets and methods have come into standard use in the vision community [SS02; Eve+10]. Robotic datasets have reached a high level of detail presenting either detailed benchmarking proce-

dures [Kü+09], or datasets with reliable ground truth and open resources for comparison [Smi+09; BMG09].

The RAWSEEDS dataset [RAW11], which includes monocular streams for large scale scenarios, has been used for the validation of 1-PR. While being suitable to benchmark very large real-image experiments, robotic datasets face two main inconveniences: First, the robot motion is planar in all the datasets, thus not allowing to evaluate full 6-DoF motion estimation. And second, GPS only provides translational data so angular estimation cannot be benchmarked. Simulation environments, like the one described in [FP09], can provide the translational and angular ground truth for any kind of camera motion. Nevertheless, these simulation environments usually cannot represent full real world complexity.

The benchmarking method proposed and used in this thesis overcomes all these limitations. It consists of comparing the estimation results against a Bundle Adjustment solution over high resolution images. Full 6 DoF motion can be evaluated with low user effort (only the generation of a Bundle Adjustment solution is required), requirements for hardware are low (a high resolution camera), and any kind of motion or scene can be evaluated, since the method operates over the real images themselves.

This approach is not entirely new: the use of a global Bundle Adjustment solution to benchmark sequential algorithms has already been used in [ED07; Mou+09]. The contribution here is the validation of the algorithm showing that the Bundle Adjustment uncertainty is much lower than the sequential methods to benchmark. As another novelty, global Bundle Adjustment is applied over high resolution images further improving accuracy. While it is true that a Bundle Adjustment solution may still suffer from scale drift, it will be much lower than that of the sequential algorithms. Also, scale drift can be driven close to zero by carefully choosing the images over which to apply Bundle Adjustment, in order to form a well-conditioned network [Tri+00], so the validity of the method is not compromised.

### 3.2.2 1-PR EKF Algorithm

Algorithm 1 outlines the proposed combination of 1-PR inside the EKF framework in its most general form in the belief that this method may be of application in a large number of estimation problems. Figures 3.7 and 3.8 illustrate the algorithm steps over a laparoscopic image.

**Algorithm 1** 1-Point RANSAC EKF

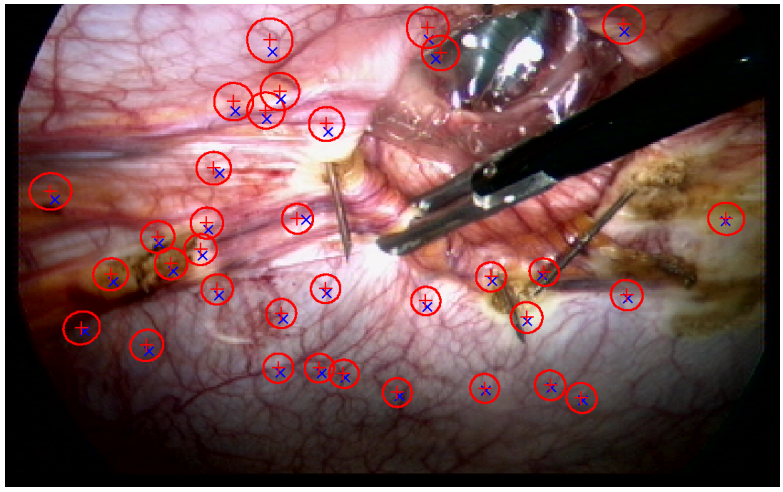
INPUT: $\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}$ {EKF estimate at step $k-1$}
  $th$ {Threshold for low-innovation points.}
OUTPUT: $\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}$ {EKF estimate at step $k$}

{A. EKF prediction and individually compatible matches}
$[\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}] = EKF\_prediction(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}, \mathbf{u})$
$[\hat{\mathbf{h}}_{k|k-1}, \mathbf{S}_{k|k-1}] = measurement\_prediction(\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1})$
$\mathbf{z}^{IC} = search\_IC\_matches(\hat{\mathbf{h}}_{k|k-1}, \mathbf{S}_{k|k-1})$

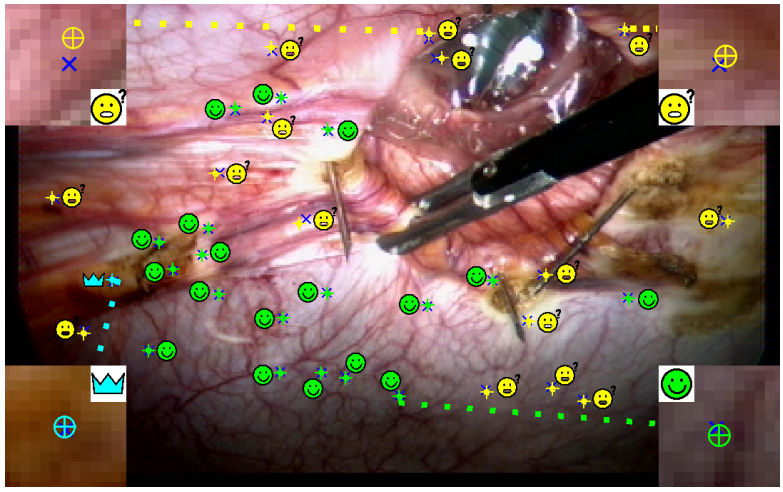
{B. 1-Point hypothesis generation and evaluation}
$\mathbf{z}^{li\_inliers} = [\,]$
$n_{hyp} = \infty$ {Initial value. Updated in the loop}
**for** $i = 0$ to $n_{hyp}$ **do**
  $\mathbf{z}_i = select\_random\_match(\mathbf{z}^{IC})$
  $\hat{\mathbf{x}}_i = EKF\_state\_update(\mathbf{z}_i, \hat{\mathbf{x}}_{k|k-1})$ {Only state; NO covariance}
  $\hat{\mathbf{h}}_i = predict\_all\_measurements(\hat{\mathbf{x}}_i)$
  $\mathbf{z}_i^{th} = find\_matches\_below\_a\_threshold(\mathbf{z}^{IC}, \hat{\mathbf{h}}_i, th)$
  **if** $size(\mathbf{z}_i^{th}) > size(\mathbf{z}^{li\_inliers})$ **then**
    $\mathbf{z}^{li\_inliers} = \mathbf{z}_i^{th}$
    $\epsilon = 1 - \frac{size(\mathbf{z}^{li\_inliers})}{size(\mathbf{z}^{IC})}$
    $n_{hyp} = \frac{log(1-p)}{log(1-(1-\epsilon))}$
  **end if**
**end for**

{C. Partial EKF update using low-innovation inliers}
$[\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}] = EKF\_update(\mathbf{z}^{li\_inliers}, \hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1})$

{D. Partial EKF update using high-innovation inliers}
$\mathbf{z}^{hi\_inliers} = [\,]$
**for** every match $\mathbf{z}^j$ above a threshold $th$ **do**
  $[\hat{\mathbf{h}}^j, \mathbf{S}^j] = point\_j\_prediction\_and\_covariance(\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}, j)$
  $\nu^j = \mathbf{z}^j - \hat{\mathbf{h}}^j$
  **if** ${\nu^j}^\top {\mathbf{S}^j}^{-1} \nu^j < \chi^2_{\alpha,d}$ {$\alpha$: Confidence level; $d$: DoF} **then**
    $\mathbf{z}^{hi\_inliers} = add\_match\_j\_to\_inliers(\mathbf{z}^{hi\_inliers}, \mathbf{z}^j)$
  **end if**
**end for**
$[\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}] = EKF\_update(\mathbf{z}^{hi\_inliers}, \hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k})$



(a) Individually compatible –IC– matches. State prediction (+) with their corresponding elliptical search regions.



(b) Consensus hypothesis and low-innovation matches. The match generating the hypothesis ( ). Low-innovation supporting matches ( ). Non-supporting matches ( ).

Figure 3.7: 1-PR stages corresponding to one frame for the operation in Figure 4.9a (I): (a) Individually compatible –IC– matches. (b) RANSAC winner hypothesis and consensus low-innovation matches. The estimated state is represented by its projection in the image, (+) stands for the estimate and the ellipse stands for the covariance. The measurements are displayed as (×). Different colors are used to code different matching categories. Zoom is made over 4 paradigmatic matches for each class of matches.

### EKF Prediction and Individually Compatible Matching (lines 5–7)

The algorithm begins with standard EKF prediction: the estimation for the state vector $\mathbf{x}_{k-1|k-1}$ at step $k-1$, modeled as a multidimensional Gaussian $\mathbf{x}_{k-1|k-1} \sim \mathcal{N}\left(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}\right)$, is propagated to step $k$ through the known dynamic model $\mathbf{f}_k$:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{f}_k\left(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_k\right) \tag{3.4}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^\top + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^\top \ . \tag{3.5}$$

In the above equation $\mathbf{u}_k$ stands for the control inputs to the system at step $k$; $\mathbf{F}_k$ is the Jacobian of $\mathbf{f}_k$ with respect to the state vector $\mathbf{x}_{k|k-1}$ at step $k$; $\mathbf{Q}_k$ is the covariance of the zero-mean Gaussian noise assumed for the dynamic model, and $\mathbf{G}_k$ is the Jacobian of $\mathbf{f}_k$ with respect to that noise at step $k$.

The predicted probability distribution for the state $\mathbf{x}_{k|k-1}$ can be used to ease the correspondence search (active search), as described in Section 2.2.4 for the visual SLAM case. Propagating this predicted state through the measurement model $\mathbf{h}_i$ offers a Gaussian prediction for each measurement:

$$\hat{\mathbf{h}}_i = \mathbf{h}_i\left(\hat{\mathbf{x}}_{k|k-1}\right) \tag{3.6}$$

$$\mathbf{S}_i = \mathbf{H}_i \mathbf{P}_{k|k-1} \mathbf{H}_i^\top + \mathbf{R}_i \ , \tag{3.7}$$

where $\mathbf{H}_i$ is the Jacobian of the measurement $\mathbf{h}_i$ with respect to the state vector $\mathbf{x}_{k|k-1}$, and $\mathbf{R}_i$ is the covariance of the Gaussian noise assumed for each individual measurement. The actual measurement $\mathbf{z}_i$ should be exhaustively searched for inside the 99% probability region defined by its predicted Gaussian, $\mathcal{N}\left(\hat{\mathbf{h}}_i, \mathbf{S}_i\right)$, by comparison to the chosen local feature descriptor. Figure 3.7a shows measurement predictions of the map points ($\hat{\mathbf{h}}$: +), their elliptical search region and their corresponding found measurements ($\mathbf{z}$: ×) obtained for a monocular laparoscopic example. All measurements ($\mathbf{z}$) compose a set of individually compatible matches ($\mathbf{z}^{IC} = (\mathbf{z}_1, \cdots, \mathbf{z}_i, \cdots, \mathbf{z}_n)^\top$).

Active search allows computational savings and also constraints the matches to be individually compatible with the predicted state $\mathbf{x}_{k|k-1}$. Nevertheless, ensuring geometric compatibility for each separated match $\mathbf{z}_i$ does not guarantee the global consensus of the whole set. Therefore, the joint compatibility of the data against a global model still has to be checked for the set of individually compatible matches $\mathbf{z}^{IC}$ previous to the EKF update.

### 1-Point Hypothesis Generation and Evaluation (lines 9–21)

Following the principles of RANSAC, random state hypotheses $\hat{\mathbf{x}}_i$ are generated and data support is computed by counting measurements below a

threshold. It is assumed here that the predicted measurements are highly correlated, such that every hypothesis computed from one match reduces most of the common uncertainty producing an inlier uncertainty close to the measurement noise $\mathbf{R}_i$.

As the key difference with respect to standard RANSAC, random hypotheses will be generated not only based on the data $\mathbf{z}^{IC}$ but also on the predicted state $\mathbf{x}_{k|k-1} \sim \mathcal{N}\left(\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}\right)$. Exploiting this prior knowledge allows to reduce the sample size necessary to instantiate the model parameters from the minimal size to define the DoF of the model to only one data point. Since the termination criterion of the RANSAC algorithm in (3.3) grows exponentially with the sample size, using only one point reduces drastically the number of hypotheses to try.

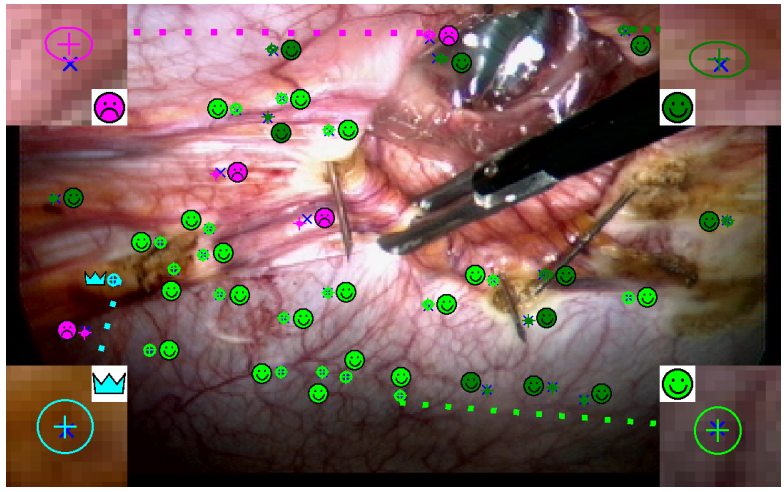
Another key aspect for the efficiency of the algorithm is that each hypothesis $\hat{\mathbf{x}}_i$ generation only needs an EKF state update using a single match $\mathbf{z}_i$. A covariance update, which is of quadratic complexity in the size of the state, is not computed and hence the cost per hypothesis is low. Hypothesis support is calculated by projecting the updated state into the camera, which can also be performed at very low cost compared with other stages in the EKF algorithm. All features whose Euclidean distance between their measurement and their new estimate is lower than an arbitrary threshold (originally this threshold was established as 2 times the measurement noise) are considered as supporters.

Figure 3.7b shows a match generating a hypothesis, its supporting match set (low-innovation inliers), and the non-supporting matches for the monocular laparoscopic example.

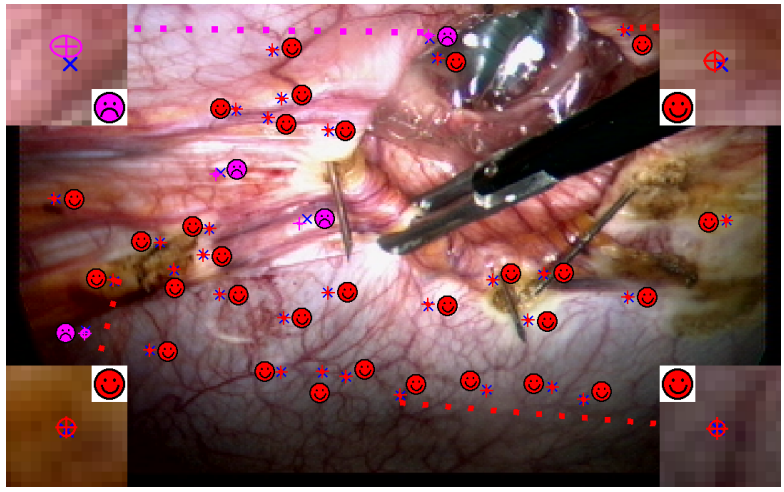
### Partial Update with Low-Innovation Inliers (line 23)

Data points voting for the most supported hypothesis, $\mathbf{z}^{li\_inliers}$, are designated as low-innovation inliers. They are assumed to be generated by the true model, since they are at a small distance from the most supported hypothesis. The rest of the points may be outliers but also inliers, even if they are far from the most supported hypothesis.

It is well known that distant points are useful for estimating camera rotation while close points are necessary to estimate translation. In the RANSAC hypothesis generation step, a distant feature would generate a highly accurate 1-point hypothesis for rotation, while translation would remain inaccurately estimated. Other distant points would in this case have low innovation and would vote for this hypothesis. But as translation is still inaccurately estimated, nearby points would presumably exhibit high innovation even if they are inliers.



(a) Low-innovation inliers update & high-innovation inliers rescue. Set of low-innovation inliers ( ). Rescued high-innovation inliers ( ) are now inside the search region and then accepted. Spurious matches ( ) remain out of the new search region.



(b) Final update. The updated state results from the integration of high and low-innovation inliers ( ). Outliers ( ) are not integrated.

Figure 3.8: 1-PR stages corresponding to one frame for the operation in Figure 4.9a (II): (a) Low-innovation partial update and the rescued high-innovation inliers. (b) Fully updated map. The estimated state is represented by its projection in the image, (+) stands for the estimate and the ellipse stands for the covariance. The measurements are displayed as (×). Different colors are used to code different matching categories. Zoom is made over 4 paradigmatic matches for each class matches.

So after having determined the most supported hypothesis and the other points that vote for it, some inliers still have to be "rescued" from the high-innovation set. Such inliers will be rescued after a partial state and covariance update using only the reliable set of low-innovation inliers:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}' \left( \mathbf{z}^{li\_inliers} - \mathbf{h}' \left( \hat{\mathbf{x}}_{k|k-1} \right) \right) \tag{3.8}$$

$$\mathbf{P}_{k|k} = \left( \mathbf{I} - \mathbf{K}'\mathbf{H}' \right) \mathbf{P}_{k|k-1} \tag{3.9}$$

$$\mathbf{K}' = \mathbf{P}_{k|k-1}\mathbf{H}'^{\top} \left( \mathbf{H}'\mathbf{P}_{k|k-1}\mathbf{H}'^{\top} + \mathbf{R}' \right)^{-1} \tag{3.10}$$

where $\mathbf{H}' = (\mathbf{H}'_1, \cdots, \mathbf{H}'_i, \cdots, \mathbf{H}'_n)^{\top}$ stands for the Jacobian of the measurement equation $\mathbf{h}' \left( \hat{\mathbf{x}}_{k|k-1} \right)$ that projects the low-innovation inliers into the sensor space. $\mathbf{R}'$ is the covariance assigned to the sensor noise.

### Partial Update with High-Innovation Inliers (lines 25–33)

After a partial update using low-innovation inliers, most of the correlated error in the EKF prediction is corrected and the covariance is greatly reduced. This high reduction will be exploited for the recovery of high-innovation inliers: as correlations have weakened, consensus for the set will not be necessary to compute and individual compatibility will suffice to distinguish inliers from outliers.

An individual Gaussian prediction $\mathbf{h}^j \sim \mathcal{N} \left( \hat{\mathbf{h}}^j, \mathbf{S}^j \right)$ will be computed for each high innovation measurement $\mathbf{z}^j$ by propagating the state after the first partial update $\mathbf{x}_{k|k}$ through the projection model. The match will be accepted as an inlier if it passes a $\chi^2$ test, based on the new and more accurate innovation covariance, with an $\alpha$ confidence level (typically $\alpha = 95\%$) test and $d$ DoF ($d = 2$ in the monocular visual SLAM case). Matches which do not pass the test are marked as spurious (outliers). The final number of spurious matches is rather low, but their rejection is a must for performance.

Figure 3.8a shows the EKF update with the final low-innovation inlier set. Some of the previous non-supporting matches are now supporting matches and are rescued (high-innovation inliers). The final non-supporting matches are the final spurious matches.

After testing all the high-innovation measurements, a second partial update will be performed with all the points classified as inliers, $\mathbf{z}^{hi\_inliers}$, following the usual EKF equations. Figure 3.8b shows the final EKF update after integrating the high-innovation inliers.

It is worth mentioning that splitting the EKF update does not have a noticeable effect on the computational cost. If $n$ is the state size and $m$

the measurement vector size, and in the usual SLAM case where the state is much bigger than the locally measured set $n >> m$, the main EKF cost is the covariance update which is $\mathcal{O}\left(mn^2\right)$. If the update is divided into two steps of measurement vector sizes $m_1$ and $m_2$ ($m = m_1 + m_2$), this covariance update cost stays almost the same. Some other minor costs grow, like the Jacobian computation which has to be done twice. But also some others are reduced, like the measurement covariance inversion which is $\mathcal{O}\left(m^3\right)$. Nevertheless, the effect of the last two is negligible and for most EKF estimation cases the cost is dominated by the covariance update and remains approximately the same.

### 3.2.3 1-PR EKF Exhaustive Algorithm

The previous section detailed the original and most general version of the 1-PR algorithm. However, in the case of monocular EKF, and laparoscopy in particular, an exhaustive version of 1-PR may be used (Algorithm 2). An explanation of this algorithm can be found in the video [GGh](1:33 - 2:02).

In monocular EKF SLAM the cardinality of the individual compatible match set is low, in the order of tens, and hypotheses generation is cheap since they are generated from just one measurement. As a result, the cardinality of the hypothesis set is the same as the IC match set. Therefore, it is possible to exhaustively test all hypotheses.

The exhaustive 1-PR mainly differs from the original (Algorithm 1) in the random hypothesis generation. Tasks of the original version such as selecting a random match from the set of IC matches, or recomputing the number of hypotheses, which required to ensure that at least one spurious-free hypothesis has been tested, are not needed in the exhaustive version where all IC matches are considered as hypotheses.

Besides, unlike the original *hypothesis generation and consensus* stage, where the supporting test consists on an arbitrary threshold (typically, it was 2 times the measurement noise), in the exhaustive 1-PR only a cheap $\chi^2$ test is applied to identify the support for the hypothesis. Since the predicted measurements are assumed to be highly correlated, every hypothesis computed from one match reduces most of the common uncertainty producing an inlier uncertainty close to the measurement noise $\mathbf{R}_k$, so that the innovation covariance may be approximated as the measurement noise covariance $\mathbf{S}_k \approx \mathbf{R}_k$. Therefore, this approximation enables the use of the cheap $\chi^2$ test.

Although in the exhaustive version the hypothesis generation is not random, the RANSAC name is still kept because, in any case, this method is quite akin to the popular algorithm.

**Algorithm 2** Exhaustive Hypotheses 1-PR EKF-Update

IN: $\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}$ {EKF prediction at step $k$}
  $\mathbf{z}^{IC}, \mathbf{R}_k$ {IC matches & Meas. Error Covariance}
OUT: $\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}$ {EKF estimate at step $k$}
{*A. 1-Point hypothesis generation and consensus*}
**for** every $\mathbf{z}_i$ match in $\mathbf{z}^{IC}$ **do**
  $\hat{\mathbf{x}}_i = EKF\_state\_update(\mathbf{z}_i, \hat{\mathbf{x}}_{k|k-1})$
  $\hat{\mathbf{h}}_i = predict\_all\_measurements(\hat{\mathbf{x}}_i)$
  $[\mathbf{z}_i^{su}, \mathbf{z}_i^{ns}] = find\_supporters(\mathbf{z}^{IC}, \hat{\mathbf{h}}_i, \chi^2_{2,0.95}, \mathbf{R}_k)$
  **if** $size(\mathbf{z}_i^{su}) > size(\mathbf{z}^{li\_inliers})$ **then**
    $\mathbf{z}^{li\_inliers} = \mathbf{z}_i^{su}$;  $\mathbf{z}^{nonsupport} = \mathbf{z}_i^{ns}$
  **end if**
**end for**
{*B. Partial EKF update using low-innovation inliers & rescue high-innovation inliers*}
$[\hat{\mathbf{x}}^{li}_{k|k}, \mathbf{P}^{li}_{k|k}] = EKF\_update(\mathbf{z}^{li\_inliers}, \hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1})$
**for** every $\mathbf{z}^j$ match in $\mathbf{z}^{nonsupport}$ **do**
  $[\hat{\mathbf{h}}^j, \mathbf{S}^j] = point\_j\_prediction\_and\_covariance(\hat{\mathbf{x}}^{li}_{k|k}, \mathbf{P}^{li}_{k|k}, j)$
  $\nu^j = \mathbf{z}^j - \hat{\mathbf{h}}^j$
  **if** ${\nu^j}^\top {\mathbf{S}^j}^{-1} \nu^j < \chi^2_{2,0.95}$ **then**
    $\mathbf{z}^{hi\_inliers} = add\_match\_j\_to\_inliers(\mathbf{z}^{hi\_inliers}, \mathbf{z}^j)$
  **end if**
**end for**
{*C. Partial EKF update using high-innovation inliers*}
$[\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}] = EKF\_update(\mathbf{z}^{hi\_inliers}, \hat{\mathbf{x}}^{li}_{k|k}, \mathbf{P}^{li}_{k|k})$

### 3.2.4 Experimental Validation: Benchmark Method for 6 DoF Camera Motion Estimation

The first step of the method takes an image sequence of the highest resolution in order to achieve the highest accuracy. In this Section, a $1224 \times 1026$ pixel sequence was taken at 22 frames per second. A sparse subset of $n$ camera locations $\mathbf{x}_{BA}^{C_1}$ –Equations (3.11, 3.12)– are estimated, by Levenberg-Marquardt Bundle Adjustment with a robust likelihood model [Tri+00], over the corresponding $n$ images in the sequence $\{I_1, \cdots, I_n\}$. Images are manually selected to ensure they form a strong network. The reference frame is attached to the camera $C_1$, corresponding to the first frame of the sequence $I_1$. For the next "*1-Point RANSAC vs 5-Point RANSAC*" and "*1-Point RANSAC vs JCBB*" experiments (below in this section), 62 overlapping camera locations

were reconstructed by manually matching 74 points spread over the images. $15-20$ points are visible in each image.

$$\mathbf{x}_{BA}^{C_1} = \begin{pmatrix} \mathbf{x}_{1,BA}^{C_1} \\ \vdots \\ \mathbf{x}_{n,BA}^{C_1} \end{pmatrix} , \tag{3.11}$$

$$\mathbf{x}_{i,BA}^{C_1} = \left(X_{i,BA}^{C_1}, Y_{i,BA}^{C_1}, Z_{i,BA}^{C_1}, \phi_{i,BA}^{C_1}\theta_{i,BA}^{C_1}, \psi_{i,BA}^{C_1}\right)^\top \tag{3.12}$$

where each camera location is represented by its position $\left(X_{i,BA}^{C_1}, Y_{i,BA}^{C_1}, Z_{i,BA}^{C_1}\right)^\top$ and its orientation encoded as Euler angles $\left(\phi_{i,BA}^{C_1}, \theta_{i,BA}^{C_1}, \psi_{i,BA}^{C_1}\right)^\top$. The covariance of the solution is computed by back-propagation of reprojection errors $\mathbf{P}_{BA}^{C_1} = \left(\mathbf{J}^\top \mathbf{R}^{-1} \mathbf{J}\right)^{-1}$, where $\mathbf{J}$ is the Jacobian of the projection model and $\mathbf{R}$ is the covariance of the Gaussian noise assumed in the model.

The input sequence is then reduced by dividing its width and height by four. The algorithm to benchmark is applied over the subsampled sequence. The reference frame is also attached to the first camera $C_1$, which is taken to be the same first one as in Bundle Adjustment. Each image for which a Bundle Adjustment estimate is available is selected and stored $\mathbf{x}_{i,MS}^{C_1}$ along with its individual covariance $\mathbf{P}_{i,MS}^{C_1}$ directly extracted from the EKF at each step.

Since the reference has been set to the same first image of the sequence, the Bundle Adjustment and sequential estimation solutions only differ in the scale of the reconstruction. Therefore, in order to compare them, the relative scale $s$ is estimated first by minimizing the error between the two trajectories. The Bundle Adjustment trajectory is then scaled $\mathbf{x}_{BA}^{C_1} = f_{scale}\left(\mathbf{x}_{BA}^{C_1}\right)$ and, together with its covariance $\mathbf{P}_{BA}^{C_1} = \mathbf{J}_{scale}\mathbf{P}_{BA}^{C_1}\mathbf{J}_{scale}^\top$.

Finally, the error is computed as the relative transformation between the two solutions:

$$\epsilon = \oplus \mathbf{x}_{BA}^{C_1} \ominus \mathbf{x}_{MS}^{C_1} \ ; \tag{3.13}$$

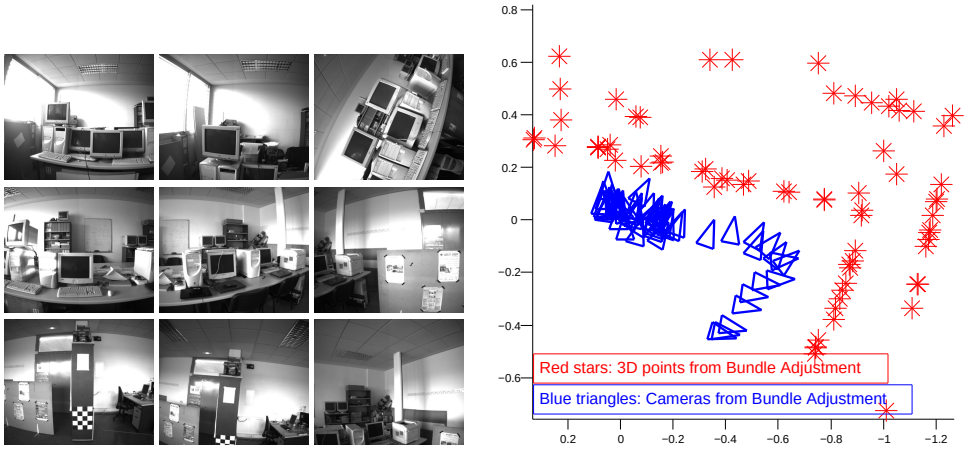
and the corresponding covariance of the error is computed by propagating the covariances of the global optimization and sequential estimate:

$$\mathbf{P}_\epsilon = \mathbf{J}_{\epsilon BA}\mathbf{P}_{BA}^{C_1}\mathbf{J}_{\epsilon BA}^\top + \mathbf{J}_{\epsilon MS}\mathbf{P}_{MS}^{C_1}\mathbf{J}_{\epsilon MS}^\top \ . \tag{3.14}$$

It was checked in the experiments that the covariance term from Bundle Adjustment, $\mathbf{J}_{\epsilon BA}\mathbf{P}_{BA}^{C_1}\mathbf{J}_{\epsilon BA}^\top$, was negligible with respect to the summed

covariance $\mathbf{P}_\epsilon$. Since this is the case, the Bundle Adjustment results can be considered as a reliable ground truth to evaluate sequential approaches. In the following figures, only uncertainty regions coming from filtering, $\mathbf{J}_{\epsilon MS}\mathbf{P}_{MS}^{C_1}\mathbf{J}_{\epsilon MS}^{\top}$ are shown.

The same subsampled sequence was used for all the experiments in the following "*1-Point RANSAC vs 5-Point RANSAC*" and "*1-Point RANSAC vs JCBB*" experiments (below in this section). The camera moves freely in 6 DoF in a computer lab, with the maximum distances between camera locations being around 5 meters. Filter tuning parameters were equal for all the experiments: motion dynamic and measurement model noise were kept the same, the number of measured features in the image was limited to 30 and all the thresholds (e.g. for feature deletion, cross-correlation, inverse depth to Euclidean conversion and initialization) were also kept the same. The reader should be aware that despite all the care taken, the experiments are not exactly the same: One of the reasons is that the outlier rate is different for each method; some methods need to initialize more features in order to keep measuring 30. Nevertheless, it is thought that this is the fairest comparison, since the algorithms try to measure always the same number of points and hence gather an equivalent amount of sensor data.




(a) Sample images from the sequence used for benchmarking.

(b) Reference solution computed by Bundle Adjustment.

Figure 3.9: Images extracted from the sequence used in the experiments and reference camera positions extracted.

Figure 3.9 shows example images from the sequence used in the following two sections for 1-PR and JCBB benchmarking. The 62 camera locations

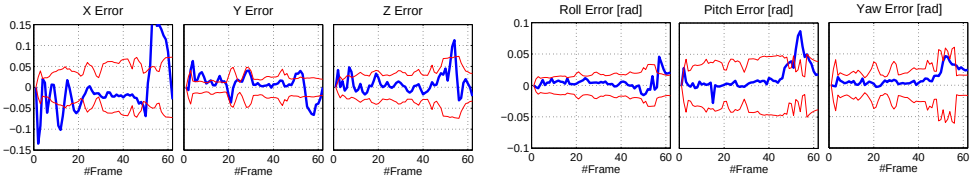
from the 2796 images long sequence are also displayed. Results for different experiments using this benchmarking method have been grouped for better visualization and comparison: Figures 3.10 and 3.12 show estimation errors for different tunings of 1-PR and JCBB; and Figure 3.13 details their computational cost. All the experiments were run on an Intel(R) Core(TM) i7 processor at 2.67GHz.

### 1-Point RANSAC vs 5-Point RANSAC

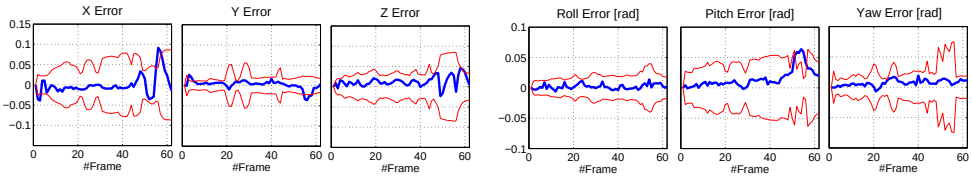
First, the performances of 5-point RANSAC (5-PR) and 1-PR are compared, to ensure that there is no degradation of performance when the sample size is reduced. Figures 3.10a and 3.10b show the errors of both algorithms with respect to the reference camera motion, along with their 99% uncertainty regions. It can be observed that reducing the sample size from 5 to 1 does not have a significant effect either on the accuracy or the consistency of the estimation. On the contrary, the figure even shows 1-PR outperforming 5-PR. This may be attributed to the fact that, unlike in classical SfM algorithms [RFP08], the theoretical number of hypotheses, given by Equation 3.3, was not inflated in the experiments. By increasing the number of iterations, 5-PR comes close to 1-PR; but it is remarkable that without this augmentation 1-PR already shows good behavior. The standard deviation of image noise was chosen to be 0.5 for the experiments since subpixel matching is used.

While the accuracy and consistency remains similar, the computational cost is much higher for the usual 5-PR than the proposed 1-PR. The details of the computational cost of both algorithms can be seen in Figures 3.13a and 3.13b. The cost of RANSAC is low compared with the rest of the EKF computations for the 1-PR case, but it is several orders of magnitude higher and is the main cost in the 5-PR case. This is caused by the increase in the number of random hypotheses in frames with a large number of spurious matches. Figures 3.11a and 3.11b show the number of hypotheses in both cases, revealing that in 5-PR this is two orders of magnitude.

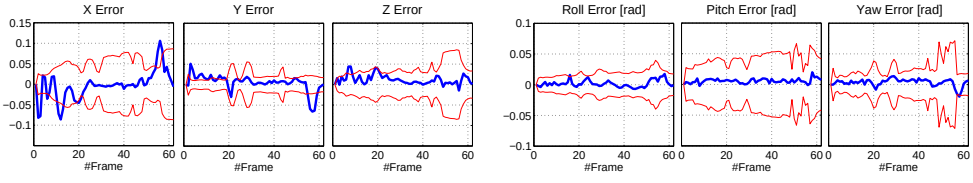
Hypothesis generation from a single point opens the possibility of an exhaustive approach rather than a random one: while an exhaustive generation of all the possible combinations of 5 points in the measurement subset would be impractical, an exhaustive generation of 1-point hypotheses implies only as many hypotheses as measurements. Figure 3.10c details the errors for the 1-point exhaustive hypothesis generation case. Compared with 1-point random hypothesis generation in Figure 3.10b, a similar accuracy and consistency is observed. Figure 3.11c shows the number of iterations needed for comparison with the random adaptive case (Figure 3.11b). The computational cost is increased but, as shown in Figure 3.13c, it is still dominated by the EKF



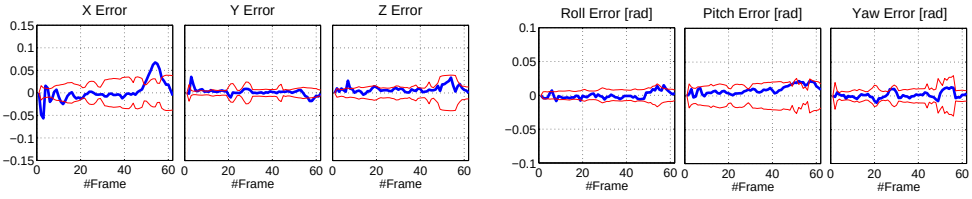

(a) 5-PR, $\sigma_z = 0.5$ pixels.




(b) 1-PR, $\sigma_z = 0.5$ pixels.

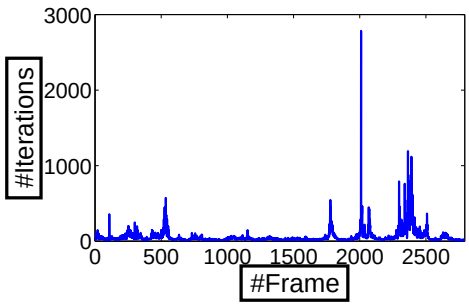



(c) 1-PR exhaustive hypothesis, $\sigma_z = 0.5$ pixels.

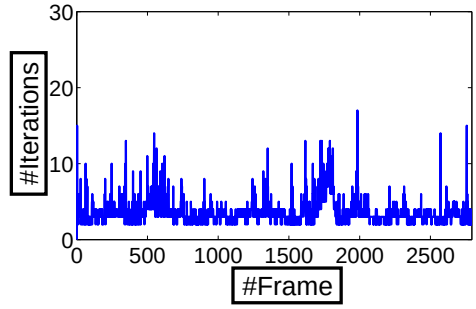



(d) 1-PR, $\sigma_z = 0.2$ pixels.

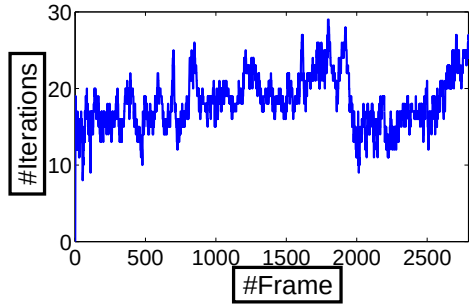
Figure 3.10: Camera location error (thick blue line) and uncertainty (thin red line) for different RANSAC configurations. Similar error and consistency are shown for 5-PR and 1-PR in Figures 3.10a and 3.10b, respectively. Figure 3.10c also reports similar results for exhaustive hypothesis testing. Figure 3.10d shows smaller errors as a result of making 1-PR stricter by reducing the standard deviation of measurement noise.




(a) Number of iterations along the sequence for 5-PR.




(b) Number of iterations along the sequence for 1-PR.




(c) Number of iterations along the sequence for exhaustive hypothesis generation.

Figure 3.11: Number of iterations for 5-PR and 1-PR. Notice that the several orders of magnitude for the 5-PR case cause a large cost overhead when compared with 1-PR (Figures 3.13a, 3.13b and 3.13c detail the computational cost for the three cases respectively).

update cost. Both options are thus suitable for real-time implementation.

Analyzing the computational cost in Figure 3.13b it can be concluded that the cost for 1-PR is always low compared with EKF computation even when the spurious match rate is high (the spurious match rate is shown in Figure 3.14b). As will be shown later, the latter becomes an important advantage over JCBB, whose cost grows exponentially with the rate of spurious matches. This efficiency opens the possibility of making the RANSAC algorithm stricter by reducing the measurement noise standard deviation and hence discarding high noise points in the EKF. Such analysis can be done by reducing the standard deviation from 0.5 to 0.2 pixels: high noise points were discarded as outliers, as can be seen in Figures 3.14b and 3.14d. The computational cost increases, as shown in Figure 3.13e with respect to 3.13b,

but still remains small enough to reach real-time performance at 22 Hz. The benefit of discarding high noise points can be observed in Figure 3.10d: errors and their uncertainty were reduced (but still kept mostly consistent) as a result of measuring more accurate points.

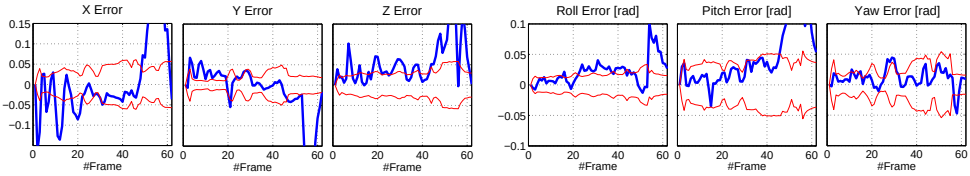
### 1-Point RANSAC vs Joint Compatibility Branch and Bound (JCBB)

RANSAC and JCBB tuning is a thorny issue when benchmarking both algorithms. Since both cases assume Gaussian distributions for the measurements and decide based on probability, choosing equal significance levels for the probabilistic tests of both algorithms is considered the fairest. The significance level was chosen to be 0.05 in the $\chi^2$ test that JCBB performs to ensure joint compatibility for the matches. Consistently, the probabilistic threshold for RANSAC was set to 95% for voting (line 15 in Algorithm 1 in Section 3.2.2) and for the rescue of high-innovation matches (line 29 in the algorithm).

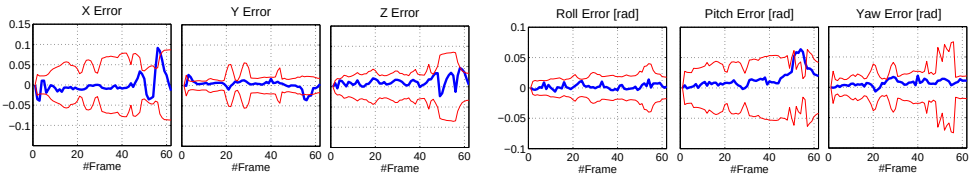
The results of benchmarking JCBB are shown in the following figures. First, Figure 3.12a details the errors and uncertainty regions for the EKF using JCBB. It can be observed that the estimation in Figure 3.12a shows larger errors and inconsistency than the 1-PR one in Figure 3.12b, repeated here for visualization purposes. The reason can be observed in Figure 3.14 where the outlier rates for 1-PR and JCBB are shown: the number of matches considered outliers by 1-PR is greater than by JCBB. The points accepted as inliers by JCBB are the ones that spoil the estimation.

A stricter version of JCBB has been benchmarked by reducing the standard deviation of uncorrelated measurement noise to 0.2 pixels, as was done with 1-PR. The spurious match rates of both algorithms, shown in Figures 3.14c and 3.14d, show that 1-PR remains more discriminative and hence produces a more accurate estimation than JCBB (Figure 3.12c). 1-PR errors for the same tuning are repeated in Figure 3.12d for comparison purposes. Also, as previously noted, the computational cost of JCBB grows exponentially when made stricter: Figure 3.13f shows peaks over one second in the worst cases.

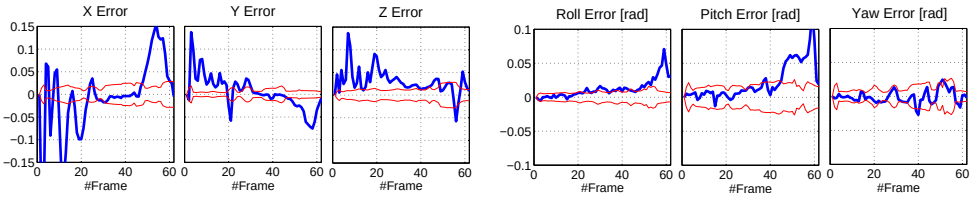
JCBB can also be made stricter by increasing the significance level $\alpha$ of the $\chi^2$ test it performs to check the joint compatibility of the data. Several experiments were run varying this parameter. The lowest estimation errors, shown in Figure 3.12e, were reached for $\alpha = 0.5$ instead of the usual $\alpha = 0.05$. Estimation errors for this best JCBB tuning are still larger than in any of the 1-PR experiments.



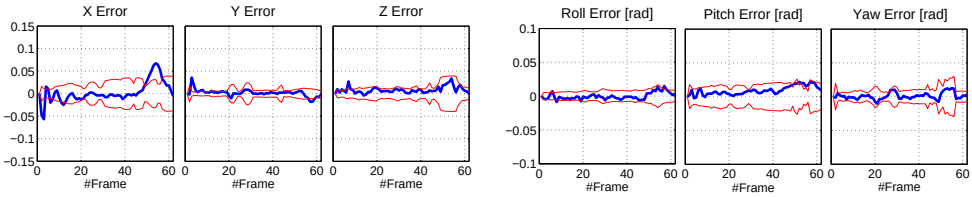

(a) JCBB, $\sigma_z = 0.5$ pixels



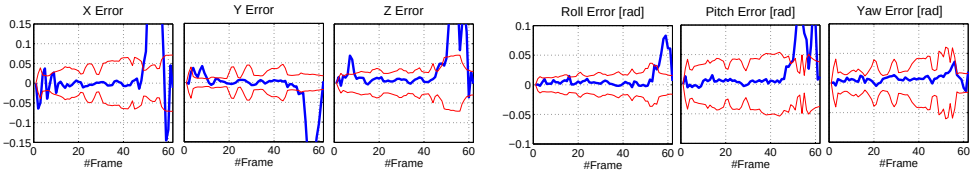

(b) 1-PR, $\sigma_z = 0.5$ pixels




(c) JCBB, $\sigma_z = 0.2$ pixels

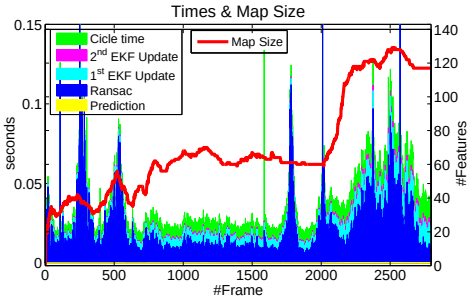



(d) 1-PR, $\sigma_z = 0.2$ pixels

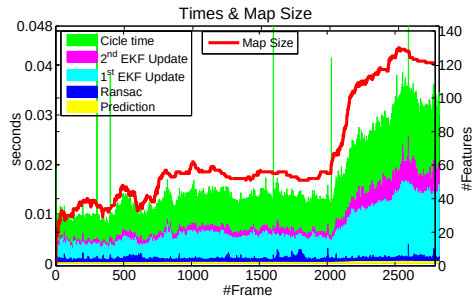



(e) JCBB, $\sigma_z = 0.2$ pixels, $\alpha = 0.5$

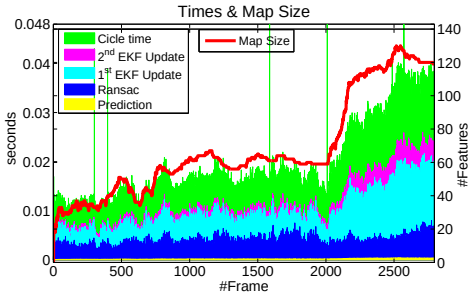
Figure 3.12: Camera location errors when using JCBB is shown in Figures 3.12a and 3.12c, for standard deviations of 0.5 and 0.2 pixels respectively. Figures 3.12b and 3.12d show 1-PR results for the same filter tuning, are repeated here for comparison. It can be seen that 1-PR outperforms JCBB in both cases. Figure 3.12e shows the best JCBB tuning found which still gives worse results than 1-PR.



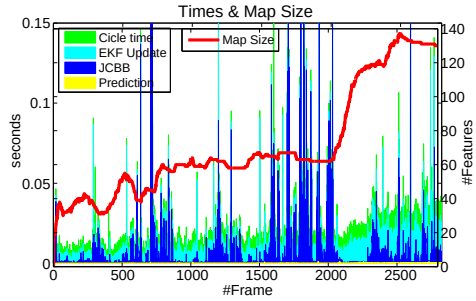

(a) 5-PR, $\sigma_z = 0.5$ pixels



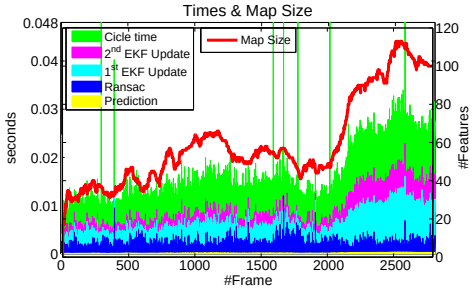

(b) 1-PR, $\sigma_z = 0.5$ pixels



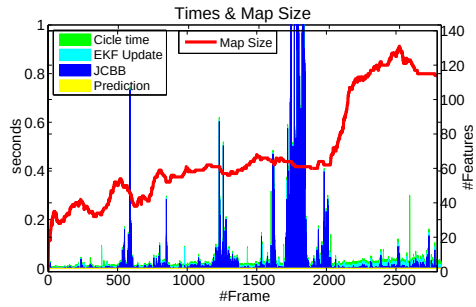

(c) 1-point exhaustive hypothesis, $\sigma_z = 0.5$ pixels




(d) JCBB, $\sigma_z = 0.5$ pixels

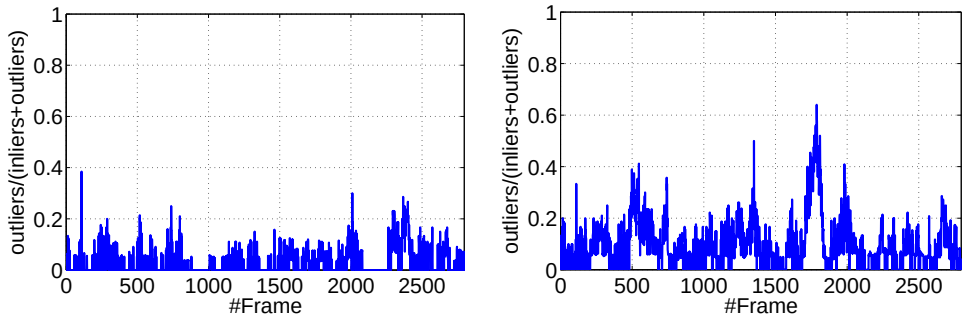



(e) 1-PR, $\sigma_z = 0.2$ pixels



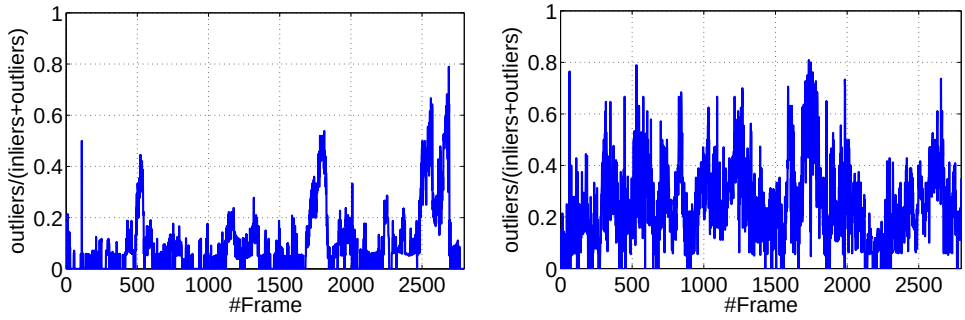

(f) JCBB, $\sigma_z = 0.2$ pixels

Figure 3.13: Detail of times and map sizes for different RANSAC and JCBB configurations in double y-axis figures: times are shown as areas and measured in seconds on the left y-axis; the map size is displayed as a a red line and is measured on the right y-axis. 1-PR exhibits much lower computational cost than 5-PR and JCBB. 1-PR also shows only a small increase when made exhaustive or stricter, making it suitable for real-time implementation at 22 Hz for the map size detailed in the figures.




(a) JCBB, $\sigma_z = 0.5$ pixels.

(b) 1-PR, $\sigma_z = 0.5$ pixels




(c) JCBB, $\sigma_z = 0.2$ pixels

(d) 1-PR, $\sigma_z = 0.2$ pixels.

Figure 3.14: Spurious match rate for JCBB and RANSAC when measurement noise standard deviation $\sigma_z$ is reduced to 0.2 pixels. It can be observed that reducing the measurement noise makes both techniques stricter, but 1-PR remains more discriminative.

### 3.2.5 Experimental Validation: Monocular EKF-Based Estimation for Long Outdoor Sequences

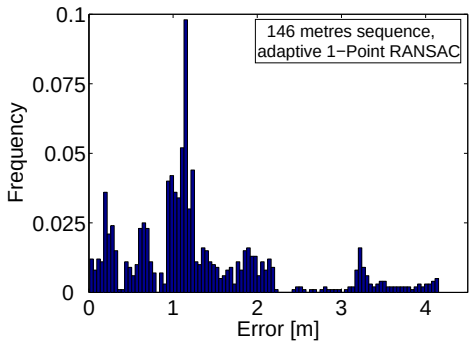
Three different sequences from the *RAWSEEDS* [RAW11] dataset have been used to test the validity of the 1-PR EKF for long-term camera motion estimation. All sequences were recorded by a $320 \times 240$ Unibrain camera with a wide-angle lens capturing at 30 fps. The estimated camera trajectories were validated against GPS data by means of an Euclidean distance:

$$\epsilon_k = \sqrt{\left(\mathbf{r}^W_{C_k} - \mathbf{r}^W_{GPS_k}\right)^\top \left(\mathbf{r}^W_{C_k} - \mathbf{r}^W_{GPS_k}\right)}. \tag{3.15}$$

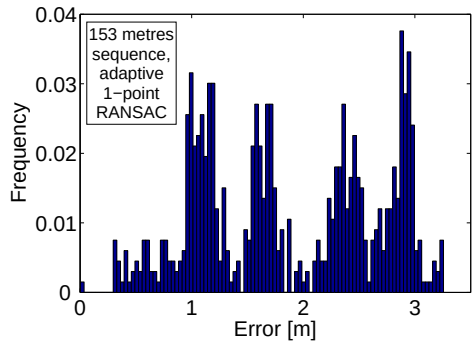
$\mathbf{r}^W_{C_k}$ corresponds with the estimated position (not the orientation) for camera $k$ and $\mathbf{r}^W_{GPS_k}$ corresponds with the GPS position for the same camera after aligning and scaling both trajectories by means of optimization.

Table 3.1: EKF-based visual estimation error for long camera trajectories.

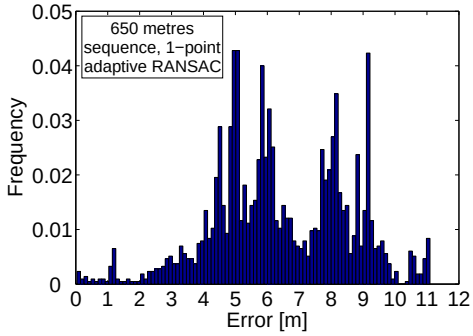
| Trajectory length [m] | Sensor used | Mean error [m] | Maximum error [m] | % mean error over the trajectory |
|---|---|---|---|---|
| 146 | monocular | 1.3 | 4.2 | 0.9% |
| 153 | monocular | 1.9 | 3.3 | 1.1% |
| 650 | monocular | 6.4 | 11.1 | 1.0% |



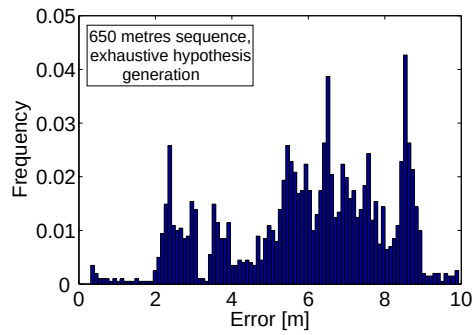

(a) 146 meters trajectory




(b) 156 meters trajectory




(c) 650 meters trajectory

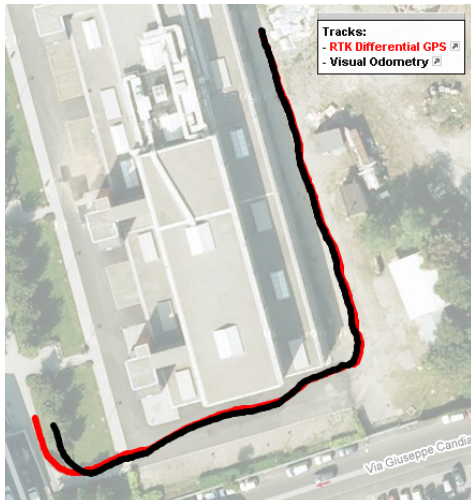



(d) 650 meters trajectory; 1-PR exhaustive

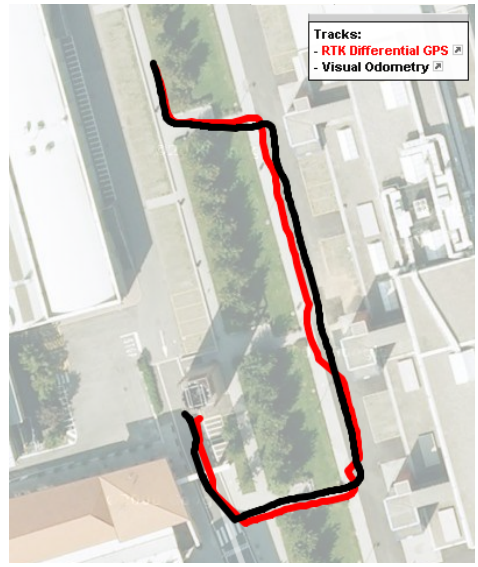
Figure 3.15: Histograms of the errors for the three experiments.

In the first sequence, consisting of 6000 images, the robot translates around 146 meters. The second sequence has 5400 images and the robot describes a similar trajectory length, around 153 meters. Finally, a very long and challenging sequence is evaluated that consists of 24180 frames (13.5 minutes of video) in which the robot describes a trajectory of 650 meters.

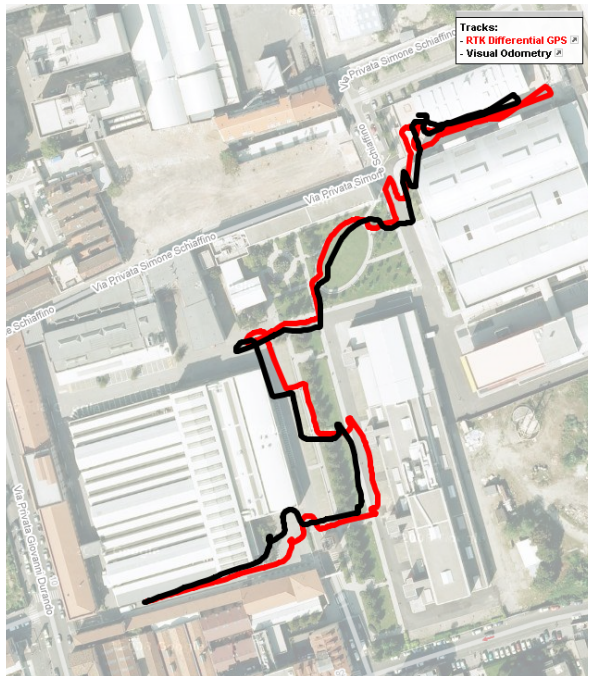
In order to reduce scale drift error, around two hundred features per frame




(a) 146 meters trajectory




(b) 156 meters trajectory




(c) 650 meters trajectory

Figure 3.16: Estimated trajectories from monocular data and GPS data.

had to be measured. This high number increased the computational cost of the EKF beyond real-time bounds. In the particular experiments presented, the algorithm ran at about 1 Hz.

Table 3.1 details the maximum and mean errors obtained in these experiments. It is worth noting that although for the three experiments the accumulated drift makes the error noticeable when plotted with the GPS trajectory, the relative error with respect to the trajectory keeps a low value (1% of the trajectory length).

Figure 3.15 shows histograms of the errors for the three sequences. Subfigures 3.15c and 3.15d show histograms of the errors for the 650 meters experiment in two different versions of the 1-PR algorithm: the first one using Algorithm 1 and the second one replacing the random hypothesis generation with exhaustive hypothesis generation (Algorithm 2) as evaluated in Figure 3.10c. The conclusion from Section 3.2.4 –"*1-Point RANSAC vs 5-Point RANSAC*"– is confirmed here: exhaustive hypothesis generation improves very slightly the estimation errors.

Figure 3.16 shows the estimated (in black) and the GPS (in red) trajectories over a top view extracted from Google Maps for each sequence. The accuracy of the estimated trajectories is clear from visual inspection.

## 3.3 Laparoscopic Experiments

In order to prove the EKF + 1-PR (in its exhaustive version –Algorithm 2) + RLR combination performance in laparoscopy, two series of laparoscopic sequences were captured. The performance relies on the configuration of three thresholds: 1) FAST feature initialization threshold, which corresponds with a Shi-Tomasi score, and indicates how distiguible is the point (the higher the score is, the more distinguible the point is); 2) matching normalized correlation threshold (when it tends to 100%, the correlation is better); and 3) reobservation rate threshold, which determines the life time of features (when it tends to 0%, it is more difficult to remove features from the map). In case of laparoscopic sequences, these thresholds are defined as 30, 40% and 40%, respectively. In contrast, they are stricter for traditional robotics sequences (scenes of man-made environments) whose typical values are defined as 300, 95% and 75%.

The first series consists of a 874 frame laparoscopic sequence at 360x288@25 Hz. The sequence, which corresponds to an abdominal exploration where a real human ventral hernia (hole) can be seen, is the same as the second experiment (Figure 2.9) of Section 2.5.2. However, in this section, only 186 frames could be processed with the EKF + JCBB combination. The

complete sequence contains some typical challenging issues of laparoscopic sequences: sudden motions, surgical tool clutter, temporary tissue deformation and laparoscope extraction and reinsertion into the abdominal cavity. These issues result in a high spurious rate that must be coped with in real time.

This series was processed in the year 2010 on an Intel Core i7 processor at 2.67 GHz, and reported in [GGCM11]. The number of features that the algorithm tried to measure in each frame was fixed to 45. A video is available in [GGf].

The second series is composed of fifteen in-vivo human laparoscopic ventral hernia repairs. These interventions were captured between April 2011 and July 2012 at 384x288@25 Hz and were processed and reported in the year 2013 in [GG+14].

All operations of this series were run on an Intel Core i7 processor at 2.93 GHz. The number of features that the algorithm tried to measure in each frame was fixed to 40 and, unlike the first series, the map size was limited to 100 points. Other differences with respect to the first series are related with the code, where some parts corresponding with the map management were optimized, and with the configuration of the relocalization, where the time to try hypotheses was reduced from 20ms to 10ms and the artificially assigned camera covariance when a good pose is found was halved.

For all operations in both series, laparoscope intrinsic parameters were calibrated using a standard planar pattern calibration method, based on [Zha00] (Figure 4.11), followed by bundle adjustment. A two parameter radial distortion model was applied (Section 2.2.3).

### 3.3.1 Results

#### First series - 2010: 874 frame laparoscopic sequence

Figure 3.17 shows for each frame of the only sequence of this series: the map size, the number of measured features, and the number of outliers. A feature is considered as an outlier if it is matched by image correlation inside the active search region but deemed as inconsistent by the 1-PR. Some frames present equal or even higher number of outliers than inliers. This demonstrates the effectiveness of 1-PR when facing with high outlier rates. Some of these high spurious rates correspond with temporary tissue deformation caused by tools interacting with the tissue, or by surgeons pushing the cavity from outside (Figure 3.18). When temporary deformations happen, some matches are not found simply because the corresponding points are imaged out of the search window due to the severe deformation. Other matches with smaller deformation are imaged inside the elliptical search window but even-

tually marked as spurious by 1-PR. The 1-PR performance over this sequence can be seen in video [GGf] (0:31 - 0:38 and 0:53 - 1:06).

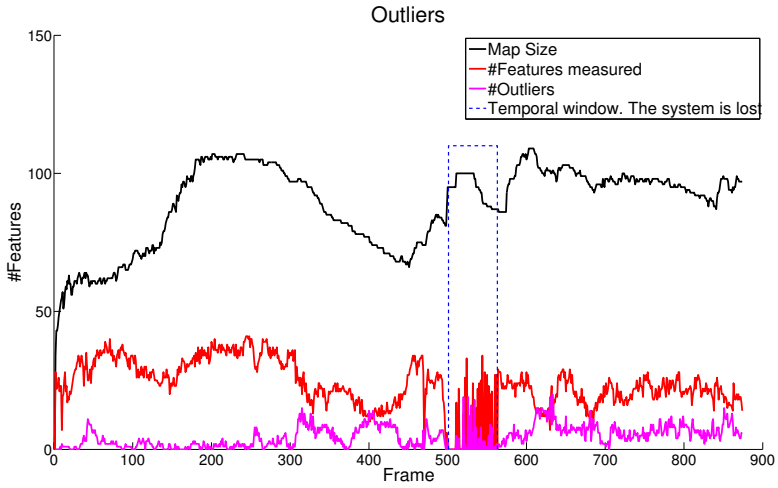



Figure 3.17: Total number of map features –black line. Measured features –red line. Spurious matches –magenta line, i.e. matches found inside the active search region marked as spurious by 1-PR. The blue dashed rectangle corresponds to frames where tracking is lost.

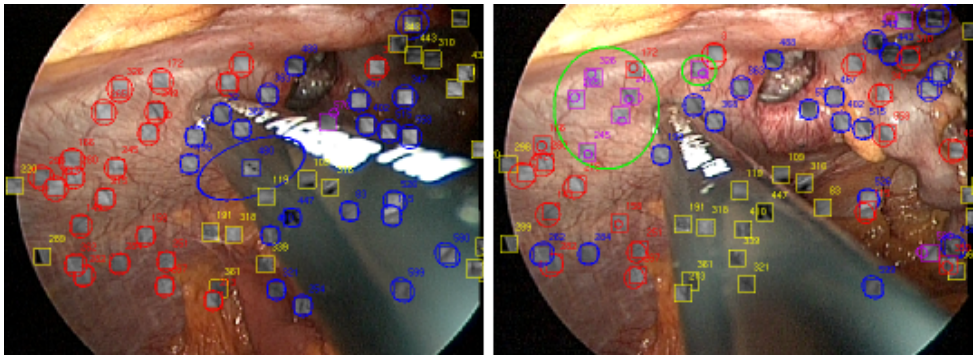


Figure 3.18: Left, frame #375 cavity undeformed, most of the map points are successfully reobserved. Right, frame #388 a significant number of map points around the tool-tissue contact point are marked as outliers and hence not measured avoiding map corruption; green circles enclose points marked as outliers; blue points close to the tool suffer such a big deformation that they are imaged out of the search window and hence not matched.

Figure 3.19 presents the total cycle time budget identifying EKF prediction, 1-PR, EKF update, EKF update for rescued matches, and initialization and map management. The EKF prediction represents an almost imperceptible share. It is also worth noting that the approximately constant time consumed by 1-PR, which is about 20% of the total budget. Notice also that the low CPU time consumed by the update for rescued matches signals that those rescued matches are just a few; however, they are very informative because they normally correspond with recently initialized points close to the camera which produce valuable translation information. Map management uses a significant fraction of the computation budget and needs a more careful optimization, above all with the integration of the RLR into the EKF + 1-PR system. This optimization will be performed in the second series experiments of 2013.

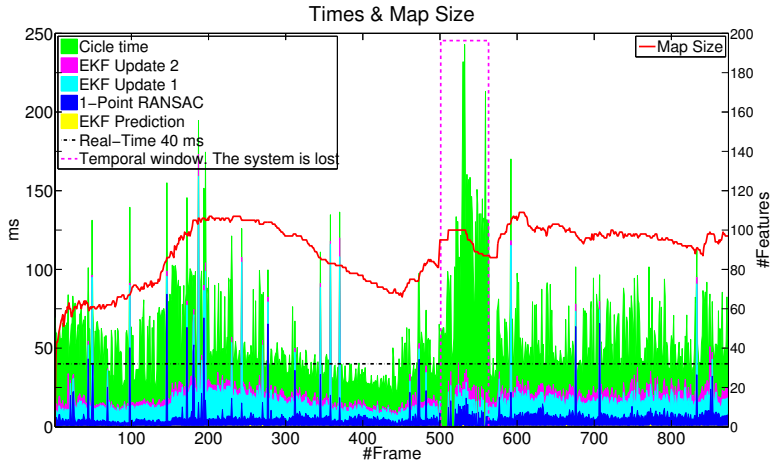



Figure 3.19: Computation time budget and map size in double y-axis figure; times are shown in milliseconds (ms) on the left axis. The map number of features is shown on the right axis. The magenta dashed rectangle signals frames where the tracking was lost.

Figure 3.20 shows the total computation time per frame as a histogram. It can be observed that the majority of frames take more than 40 ms but less than 100 ms. This data shows that the system is very close to real-time performance, which is easily achievable by optimization (and eventually achieved in the second series experiments of 2013).

During laparoscopic procedures is frequent extracting and reinserting the laparoscope into the body. This represents an extreme situation for relocalization. In Figure 3.3, four selected frames illustrate the tracking loss and

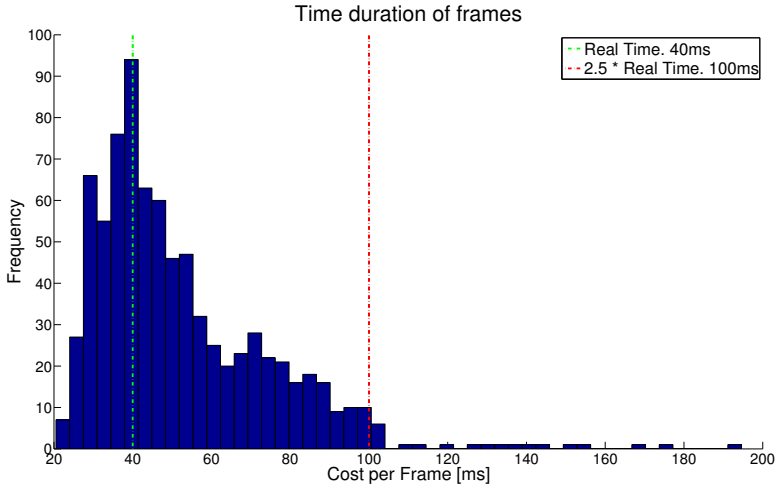



Figure 3.20: Histogram showing the computational cost.

recovery. Before total recovery, an unstable relocation stage is observed. This loss of tracking and its posterior relocalization can be found in video [GGf] (0:40 - 0:48).

An important quality of the produced map is the point persistence. The histogram in Figure 3.21 shows how long map features live. It is clear that an important fraction of the 396 initialized features dies early because they cannot be successfully reobserved. However, for a map of about 100 features, there are 54 features ($\approx 50\%$) that have survived more that 600 frames. These persistent features, selected in a survival-of-the-fittest way, are well spread over the observed cavity, locally salient, and suitable for camera relocalization (Figure 3.22).

### Second series - 2013: Fifteen laparoscopic ventral hernia repair

The proposed combination has been able to successfully compute the map and the camera trajectory for the fifteen laparoscopic ventral hernia repair sequences (Figure 4.9). It has been able to cope with a variety of illuminations, textures and input port geometries achieving real-time performance in all sequences.

To analyze the cycle time budget, the sequence corresponding to Figure 4.9c, and available in video [GGh] (1:13 - 1:32), has been selected because it is archetypical. It includes EKF routine operation and relocalization after tracking loss due to occlusion (Figure 3.4). Figure 3.23 displays the cycle time budget split in: EKF prediction, putative IC matching, 1-PR hypothe-

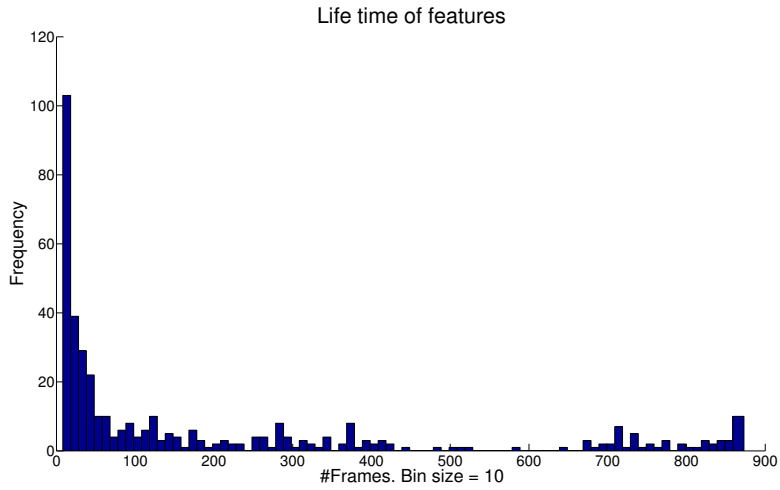



Figure 3.21: Histogram displaying feature persistence. 396 are initialized in the experiment. 54 of them survive for more than 600 frames. A new feature is tested for 10 frames, if trackable is kept otherwise is removed. For this reason, persistence lower that 10 correspond to non trackable features.

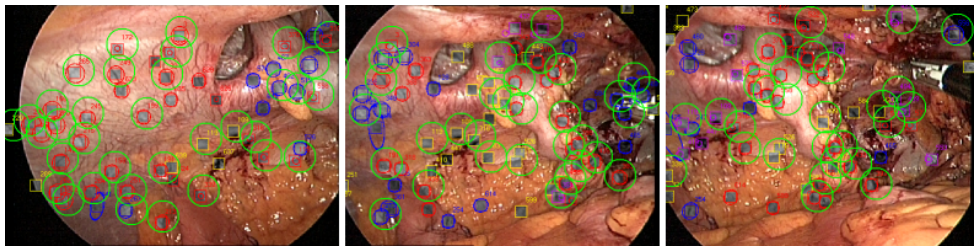


Figure 3.22: Features surviving for more than 600 frames are surrounded by a green circle.

ses generation and consensus, low innovation inliers update, high-innovation rescue and update, and map management (feature creation and removal). IC matching is time consuming due to image correlation and patch warping. As it can be seen in the image, all frames were processed in real time (25 Hz; <40 ms) even those corresponding with relocalization. In comparison with Figure 3.19 of the previous series, it is remarkable the drastic time reduction in the map management stage caused by code optimization.

Figure 3.24a displays the cost-per-frame histogram for all frames in all sequences (6473 frames). The cycle time mode is around 13 ms, the mean

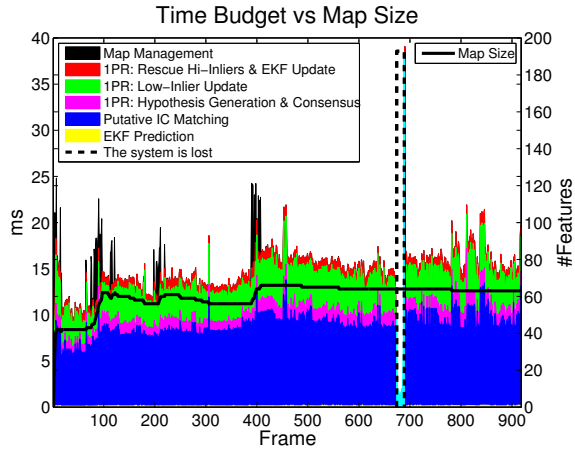
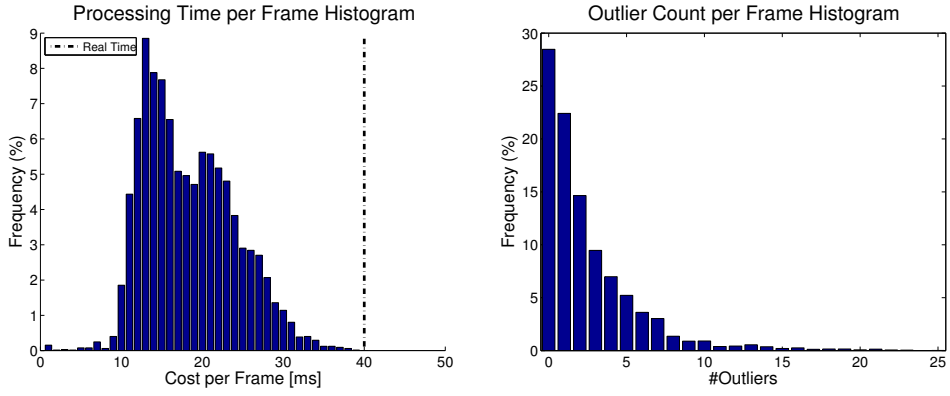



Figure 3.23: Cycle time split in the six main parts of the EKF cycle and map size for an archetypal execution corresponding to operation Figure 4.9c.

and the median being around 18 ms. Faster frames (<10 ms) correspond to relocalization when no features are detected (since in that case no relocalization hypotheses are generated), and with first sequence frames when the map is small. Times around 38 ms correspond to frames when the system has just relocalized and the camera location is still not refined. Thus, it can be concluded that robust real-time performance can be achieved.




(a) Cycle time histogram for all processed frames in the 15 sequences.

(b) Outlier histogram for all processed frames in the 15 sequences.

Figure 3.24: (a) Cycle times and (b) outlier histograms for all frames in all sequences.

Typical map sizes are between 50 and 100 points. Up to 40 map features are measured per frame. Figure 3.24b shows a histogram of the outlier count for all frames in all the sequences. Although nearly 30% of frames do not contain any spurious match, only one of the sequences can be successfully processed if 1-PR is disabled. Therefore, it can be concluded that algorithms robust to spurious data are a must for EKF SLAM even in the case of a low spurious-matches rate. 1-PR cost is linear in number of measurements and state size while the outliers have a low influence on the computational cost (<20% of the total budget corresponding to 1-PR hypotheses generation and consensus). Hence, the proposed system can achieve real time even when ∼25% of frames contain more than 3 outliers. In contrast, methods like exhaustive JCBB, with exponential complexity in the number of outliers, would not perform in real time.

## 3.4 Conclusions

### 3.4.1 1-Point RANSAC

This Chapter presents a novel RANSAC algorithm which, for the first time and differently from standard purely data-driven RANSAC, incorporates *a priori* probabilistic information into the hypothesis generation stage. As a consequence of using this prior information, the sample size for the hypothesis generation loop can be reduced to the minimum size of 1 point data. 1-PR has two main strengths that worth summing up here. First, as in standard RANSAC, model constraints are checked *after* hypothesis data has been fused with the a priori model. Second, using 1-point plus prior knowledge hypotheses greatly reduces the number of hypotheses to construct, and hence the computational cost compared with usual RANSAC based solely on data. In a practical sense, 1-PR presents a linear computational complexity in the number of outliers that means an overhead of less than 20% of the standard EKF cost, making it suitable for real-time implementation in visual SLAM.

Comparing with JCBB, where its relevance resides on their generality, the main advantage of 1-PR is its efficiency. The rich variety of correlation patterns that a covariance matrix can encode is manageable by JCBB. However, 1-PR exploits the very simple pattern where all the correlations are mainly explained by sensor motion, and hence small size data subsets are enough to constraint the rest of the measurements. Therefore, 1-PR is directed to the particular case of rigid scenes, thus for more complex models like non-rigid scenes, 1-PR may not offer such a satisfactory result.

Nevertheless, it is also true that estimation from a moving sensor data

stream in an almost rigid scene covers a great percentage of SLAM problems; and a specific method more efficient than general methods can be of importance. In this sense, 1-PR outperforms existing approaches by presenting lower cost and scaling well with the state vector and measurement size, and also with the outlier rate (1-PR presents a linear cost in the number of outliers versus the exponential complexity of JCBB).

Besides its efficiency, 1-PR has also some advantages in dealing with non-linearities as a result of checking rigidity after data fusion where some of the inaccuracies introduced by non-linearities have been compensated. On the contrary, JCBB checks rigidity before data fusion which is a serious drawback of the algorithm.

This chapter also presents a method for benchmarking 6-DoF camera-motion-estimation results. The method shows three clear advantages: Firstly, it is intended for real image sequences and includes effects difficult to reproduce by simulation (like non-Gaussian image noise, shaking handy motion, image blur or complex scenes). Secondly, it is easily reproducible as the only hardware required is a high resolution camera. And thirdly, the effort required by the user is low. The uncertainty of the estimated solution also comes as an output of the method and the appropriateness of BA estimation as reference can be validated. The method has been used to prove the claimed superiority of the 1-PR method.

The general EKF + 1-PR algorithm has been experimentally tested for the case of large camera trajectories in outdoor scenarios. Errors around 1% of the trajectory have been obtained for trajectories up to 650 meters from a publicly available dataset. The number of tracked features in the image has to be increased to two hundreds in order to avoid scale drift. This high number makes this case currently moves away from real-time performance, and the method runs at 1 frame per second.

Finally, it is also worth remarking that, although this thesis is focused on the particular case of monocular EKF-SLAM, the 1-PR method is independent of the type of sensor used. The only requirement is the availability of highly correlated prior information, which is typical of EKF-SLAM for any kind of sensor used. Also, as highly correlated priors are not exclusive to EKF-SLAM, the applicability of 1-PR could be even broader. As an example, the camera pose tracking in keyframe schemes [KM07; Mou+09] would benefit from 1-PR cost reduction provided that a dynamic model were added to predict camera motion between frames.

### 3.4.2 Laparoscopic Experiments

An improved robust version of the EKF-SLAM has been proposed in this chapter. The new version includes the integration of a relocalization system (RLR) into the EKF framework and the substitution of JCBB data association by 1-PR procedure.

This new combination has been tested over two laparoscopic sequence series. The first one consists of a 874 frame laparoscopic sequence. This sequence is the same as the second experiment (Figure 2.9) of the Section 2.5.2 and was processed in 2010 with an unoptimized version of the new combination. The second series is composed of fifteen in-vivo human laparoscopic ventral hernia repairs, which were processed in 2013 with an optimized code version achieving real-time performance.

The integration of RLR enables to recover the system after tracking losses. This is essential in laparoscopy where instrument occlusions or typical laparoscope maneuvers may cause the loss of tracking. RLR detects these losses and stops the EKF integration persevering the integrity of the map from a possible corruption. Then, RLR searches for possible putative matches between the current image and the fixed map. After a set of putative matches is found, RLR tries to recover the tracking by 3-point-pose PnP algorithm and RANSAC. When a good camera location is found, RLR reactivates the normal EKF working.

1-PR has also demonstrated to greatly outperform JCBB. 1-PR along with RLR have shown their performance over the 874 frame sequence of the first series. For this sequence, the EKF + JCBB combination (Chapter 2) only could process 186 out of 874 frames, but the new combination processed the complete sequence. This is mainly possible thanks 1-PR copes with a high number of outliers caused by a laparoscopic tool interaction (JCBB does not), and thanks RLR system which allows to support losses of tracking avoiding a complete system failure.

Then, the combination of EKF + 1-PR + RLR has shown to be appropriate to build a map from laparoscopic sequences. This combination has demonstrated to be able to cope with typical challenges in this kind of sequences: sudden motions, surgical tool cluttering, temporary tissue deformation, large occlusions and laparoscope extraction and reinsertion in the abdominal cavity.

Since the system is based on an EKF filter, the computational cost is quadratic in the map size and linear in the number of measured features in the image. In laparoscopic experiments, a significant number of map points need to be measured to achieve robust relocalization. In the first series of experiments, the map size was above 100 features and the number

of measured features was fixed to 45 resulting times lower than 3 times real-time (120 ms). However, in the second series, an optimized version of the code along with the reduction of measured features (40) and the limited map size to 100 features, achieve times lower than 40 ms per frame (25 Hz), thus it works in real time.

It must be noted that the proposed algorithm is able to compute a nice summary of the scene after processing the whole sequence. A survival-of-the-fittest process selects what scene features are included in the map. Only locally salient, trackable, and distinctive for relocation points are included in the final map. This rigid map is excellently exploited by relocalization procedure to recover from tracking losses and to relocate at reinsertions. The computed map might well be the starting point for learning priors to process sequences corresponding to similar procedures performed to different patients.

All results have been validated over real sequences, so it can be concluded that monocular SLAM in the abdominal cavity is a valid mapping method that does not need any additional sensor but just a standard monocular endoscope and commodity computers.

Despite an experimental validation has been provided for the method, it would be interesting to compare its solution with respect to a ground truth. Chapter 4 is devoted to an extensive validation of the accuracy of this system. Additionally to the accuracy validation, another validation from clinical point of view is carried out, showing the possible advantages that monocular visual SLAM entails with respect to the patient and inside the operating room.

# 4

# Exhaustive System Validation

Chapter 3 has shown the maturity of visual SLAM algorithms in the field of robotics both for relative small environments, like the first experiment compared with a bundle-adjustment ground truth, and for large environments, like those several-hundred-meter trajectories validated with GPS data. However, to date, the performance of these algorithms applied in medical imaging has not thoroughly been validated over real surgeries. Most of works in medical imaging make subjective or objective validations of the algorithms by using external trackers, phantoms, synthetic data, ex-vivo data, in-vivo animal data, or a combination of them (e.g. [MY10; Mir+12; Hu+12]). The most important contribution of this thesis is precisely the exhaustive validation of these algorithms with human real laparoscopic interventions. The validation has been carried out with simulations and real in-vivo surgeries.

Fifteen laparoscopic ventral hernia repair (LVHR) operations have been captured to validate monocular visual SLAM because: 1) the scene is almost rigid and textured; 2) the standard LVHR procedure includes accurate distance measurements that can be used as ground-truth; 3) the surgical procedure has not been modified at all, except for the addition of an exploratory endoscope maneuver; 4) SLAM exploits the images simplifying the surgical procedure without a disruptive modification of the workflow; and 5) image sequences exhibit significant inter-patient variability in texture, illumination, input port placement, and exploratory trajectory.

This chapter is devoted to detailing the visual SLAM validation over laparoscopic surgeries both from engineering and clinical points of view. Both validations have been reported in [GG+14] and [Ber+], respectively. Section 4.1 describes the ventral hernia repair procedure. Section 4.2 explains the new exploratory endoscope maneuver needed to use SLAM in LVHR proposed in

[Gil+11a; Gil+11b]. Section 4.3 details the simulations performed to validate the accuracy of the scene reconstruction and the trajectory recovered by SLAM at different camera configurations [GG+14]. Section 4.4 explains the protocol followed for data acquisition. Finally, Sections 4.5 and 4.6 show the engineering and clinical validation of these methods in laparoscopic images [GG+14; Ber+].

It is worth noting that the works [Gil+11a], [Gil+11b] and [Ber+] correspond with clinical publications, and, for that reason, the first authors are surgeons. However, these publications would not exist without the engineering contribution by the author of this thesis that has been essential in all of them.

## 4.1 Ventral Hernia Repair Procedure

A ventral hernia is a defect –hole– that appears in the internal abdominal wall due to muscular strain, weak abdominal muscles, or as a result of previous surgery (incisional hernias). Ventral hernias are dangerous because part of an organ –usually the bowel or intestine– might protrude through the hernia and cause an obstruction or strangulation of the organ accompanied by intense pain and necrosis. The reported overall prevalence of ventral hernias ranges from 2% to 13% [MH85; SR93; Ban+12] for the incisional case being a common pathology confronted by surgeons.

Repair of a ventral hernia ideally involves placement of a prosthetic mesh in the preperitoneal subaponeurotic plane, in a tension-free manner with the edges well beyond the borders of the hernia defect. Uniformly distributed intra-abdominal pressure contributes to fixation of the mesh (Pascal's principle), reducing the risk of recurrence. Both open surgery with retromuscular mesh placement and laparoscopic surgery with intraperitoneal mesh placement can benefit from uniformly distributed intra-abdominal pressure (Figure 4.1).

Historically, the most widely used surgical treatment for ventral hernias was the open retromuscular (Rives-Stoppa) repair procedure, which had the best outcomes for most incisional hernias and some primary ventral hernias. This procedure involves extensive parietal dissection and placement of a nonresorbable polypropylene or polyester mesh behind the posterior rectus fascia. Developments in biocompatible materials and endoscopic surgery [LB93] have enabled laparoscopic placement of bilaminar intraperitoneal prosthetic mesh, with minimal dissection. The mesh overlies the hernia defect and extends 3–5 cm beyond the borders of the defect [LeB+03; LeB07], and is fixed to the abdominal wall with tackers using the double-crown technique

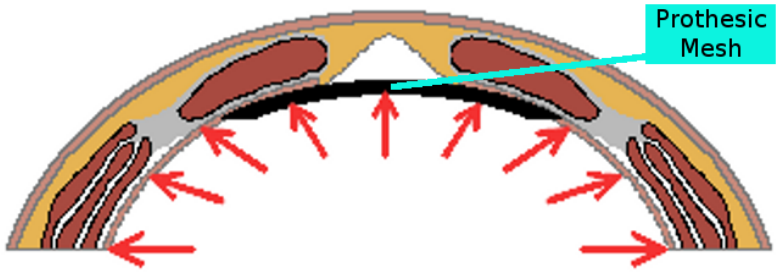



Figure 4.1: Intra-abdominal pressure helps to secure the prosthetic mesh in the intraperitoneal sublay position to the abdominal wall.

[MC+05], or transfascial sutures, or a combination of these methods; an evaluative review of the fixation methods can be found in [LeB07]. Primary closure of the hernia has good outcomes, but is technically complex [Ban+12; Ore+11]. Recently, biological adhesives such as fibrin have been used to fix the prosthetic mesh in place [Ste+10]. However, there are still uncertainties in the laparoscopic technique regarding the optimal mesh type, mesh fixation method, and measurement of hernia defect size; and the incidence of seromas. A video showing the LVHR procedure is available in [GGg].

The LVHR technique offers the advantages of the laparoscopic approach, i.e., a short hospital stay, less postoperative pain, and fast postoperative recovery. The procedure carries an acceptable risk of complications, a low risk of recurrence, and an excellent cosmetic result. LVHR is considered to be a good alternative to open surgery, at least in experienced hands [Ore+11; Sau+11].

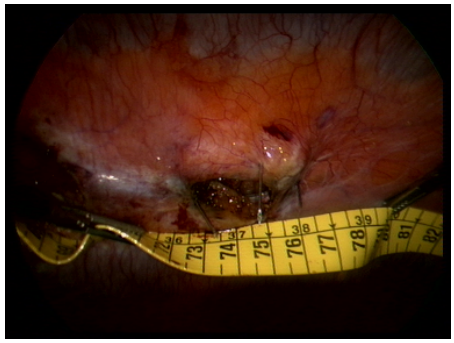
In the LVHR procedure, the hernia defect is measured in-vivo to cover the defect with a customized-in-size patch. The elliptical patch axes are those of the defect plus the predefined safety margin (3-5 cm). If possible, a piece of a sterilized tape measure is introduced inside the abdominal cavity and at least one of the two main hernia axes is measured (Figure 4.2c). If the tape measurement cannot be taken, other less accurate indirect methods as external measurement based on needle insertion (Figures 4.2a and 4.2b) are used. A video showing these two measuring techniques is avaliable in [GGi].

In this thesis, a cross-fertilization between LVHR and visual SLAM algorithms is established. On one hand, LVHR provides internal measurements that can be used as a ground-truth to validate the SLAM geometrical accuracy. A 0.5 cm tape measurement resolution determines the ground-truth accuracy. On the other hand, visual SLAM can be used as a computerized method of measuring the hernia dimensions by making use of only the image sequence gathered by the laparoscope, and a standard computer. The

visual SLAM method can be smoothly integrated into the LVHR procedure to provide measurements, that are as accurate as the classical methods (Figure 4.2), but take less time and do not require insertion of needles or a tape measure into the abdominal cavity.

(a) Internal view of the needles at the borders of the defect.

(b) External measurement between the needle insertion points.

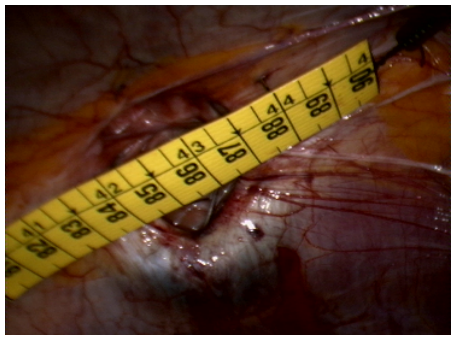



(c) Tape measure method, with direct internal measurement of the defect.

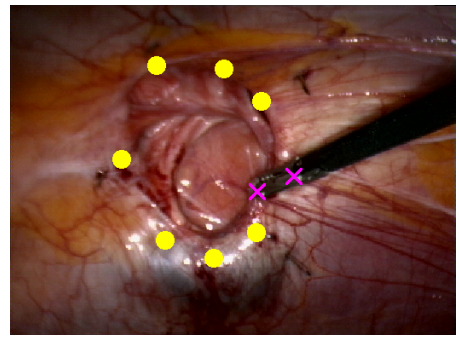
Figure 4.2: (a, b) Needle insertion method. (c) Tape measure method.

## 4.2 Hernia Repair SLAM Assisted Procedure

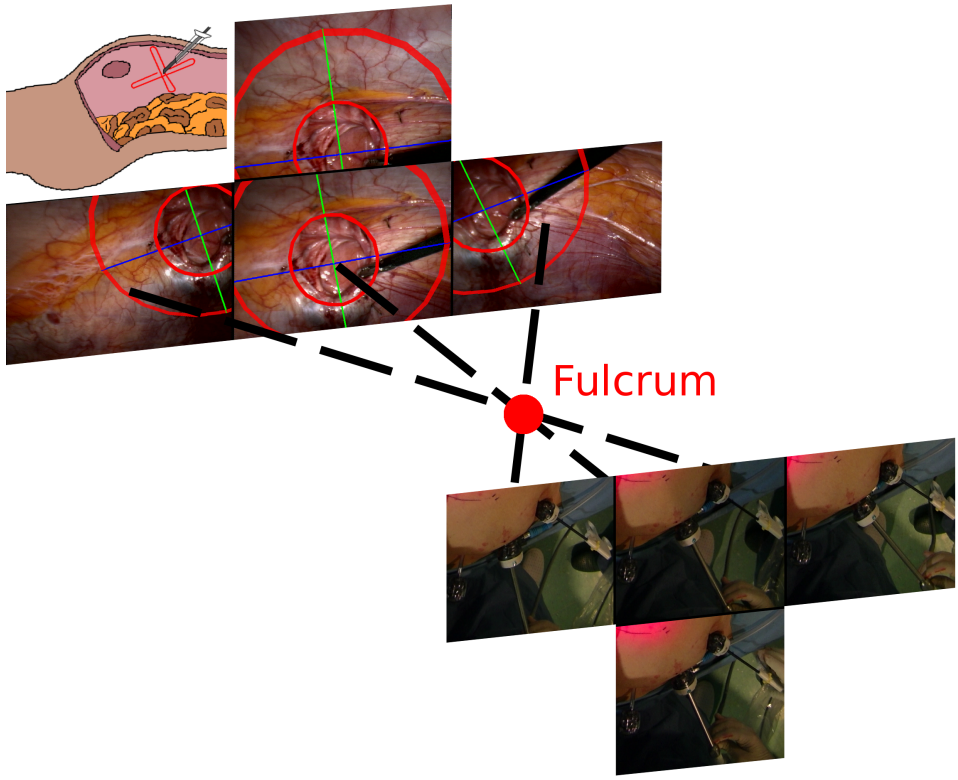
In order to enable the use of SLAM in LVHR, a new exploratory laparoscope maneuver, proposed in [Gil+11a; Gil+11b], extends the standard LVHR procedure at the measurement stage. This new exploratory laparoscope maneuver is performed aimed at translating the endoscope tip while the region of interest is kept in the field of view (FoV) (Figure 4.3c). Doing so, it is possible to gather a sequence with enough parallax for an accurate SLAM. This



(a) Tape measurement considered as ground-truth.

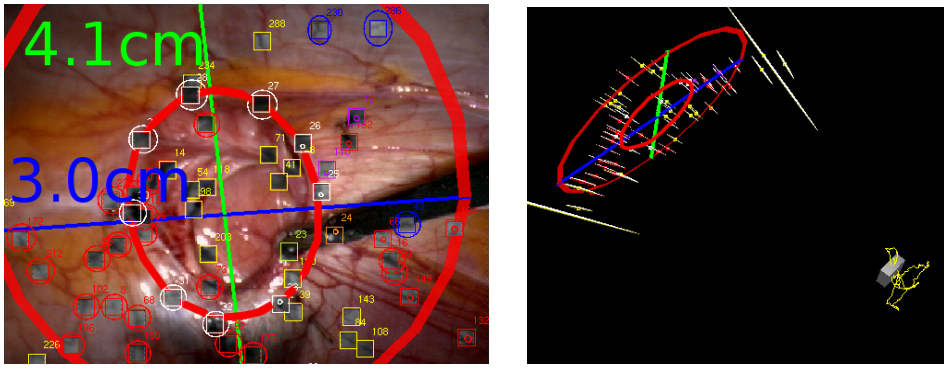


(b) Two points over a forceps define the scale (magenta crosses). Five or more points over the hernia defect boundary (yellow points).




(c) Internal and external hand-held exploratory laparoscope motion. It is worth noting that the hernia is always inside the FoV. Notice also the fulcrum effect between internal and external maneuver (when surgeon moves the laparoscope to left, the laparoscope tip move to right and vice versa).

Figure 4.3: Measurement process (I): internal measurement of the hernia, definition of the ellipse, and exploratory maneuver for the operation in Figure 4.9b.




(a) SLAM measurement, map and ellipses projected as augmented reality over a sequence frame.

(b) Camera trajectory, 3D map and ellipses in 3D. Top view.

Figure 4.4: Measurement process (II): estimated ellipses with the estimated hernia dimensions, 3D map and camera trajectory for the operation in Figure 4.9b.

sequence is processed to estimate a cavity map and the endoscope trajectory. Two videos explaining this maneuver can be found in [GGe] and in [GGh] (0:11 - 1:12).

Unlike visual SLAM using stereo images, monocular SLAM provides an up-to-scale 3D model of the cavity. To compute the actual measurements of the defect, the real scale of the 3D model is defined using a laparoscopic tool with a known tip size. Before the exploratory maneuver, additional key points are manually enforced to be in the map: two predefined points over a forceps to define the reconstruction scale, $s$, and several points (five or more) scattered over the defect boundary to estimate the hernia contour and size (Figure 4.3b).

The hernia defect is modeled as a virtual 3D ellipse in a three-stage way. In the first stage, an initial guess of the dominant plane defined by the five or more defect boundary points is computed by least squares. This guess is covariance-weighted in the second stage by an information filter extracting the needed covariances from the probabilistic map of the EKF monocular SLAM. After that, the points are projected on the weighted plane where the planar ellipse is fitted. Finally, the defect major and minor axis sizes are estimated from the ellipse (Figure 4.4a). Their dimensions are computed from the scale factor $s$ according with (2.31) where $d_m(i,j)$ and $d_m(r_1,r_2)$ correspond with the length of one of the axes and the relative distance between the two forceps points, respectively, both in the SLAM map.

Resulting from the exploration, the SLAM algorithm estimates the scene

3D map, the endoscope trajectory (Figure 4.4b) and the hernia dimensions. A second concentric ellipse defining the virtual border of the patch is visualized as an augmented reality annotation. Live augmented reality is possible due to the real-time 3D estimation of the endoscope position with respect to the 3D cavity (Figure 4.4a).

## 4.3 Simulation

The difficulty of obtaining an exact ground truth from in vivo data makes that the validation of the SLAM be an intricate issue. For that reason, a representative simulation has been designed in order to quantitatively evaluate the accuracy and robustness of the method.

The simulation mocks up the 3D geometry of the ventral hernia repair procedure where the human torso is modeled by means of an array of points on an ellipsoidal cap (Figure 4.5). Typical local non-rigid deformations of hernia repair emulating external forces, respiration or heartbeats have been applied over the cap. In the left flank of the cap, a virtual 30º DoV (Direction of View) (Figure 4.6) and 60º FoV (Field of View) endoscope, and a virtual tool tip defining the reconstruction scale have been inserted. From this setup, a synthetic image sequence is generated by moving the virtual endoscope around the fulcrum mimicking the real laparoscope movements. The 3D model points are projected according to the pinhole + two-radial-distortion parameter model (Section 2.2.3) and adding zero mean Gaussian noise with 0.5 pixels standard deviation. It has been simulated not only at the actual endoscope resolution $384 \times 288$ pixels but also double $768 \times 576$ and half $192 \times 144$.

The simulation focuses mainly on the $384 \times 288$ resolution because it corresponds to the endoscope used in the real surgeries. This simulation can be found in the video [GGh] (2:02 - 2:34). The simulation errors are aligned with respect to the last ground-truth camera. When the system is initialized, the translation and rotation uncertainties are large and hence their estimates are not very accurate. However, as the laparoscope moves and the cavity is seen with parallax, the errors decrease and the estimates improve. Presumably, the last estimated camera possesses the best estimates and the lower errors, and then the rest of the cameras must be aligned with it.

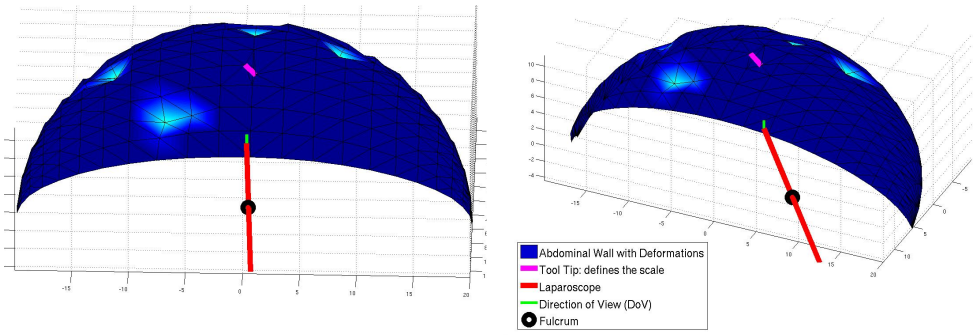



Figure 4.5: Simulation. Human torso modeled as an ellipsoidal cap. Navy blue corresponds to undeformed areas. Celeste corresponds to deforming areas.

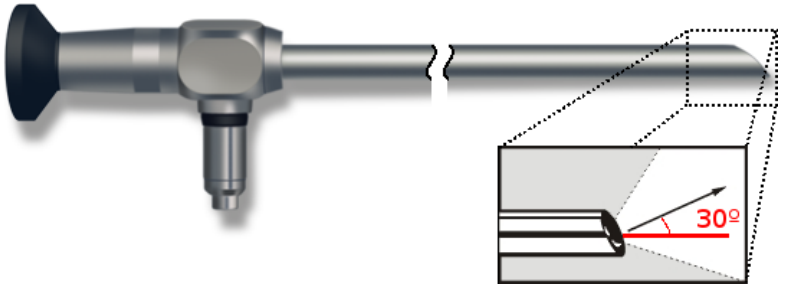


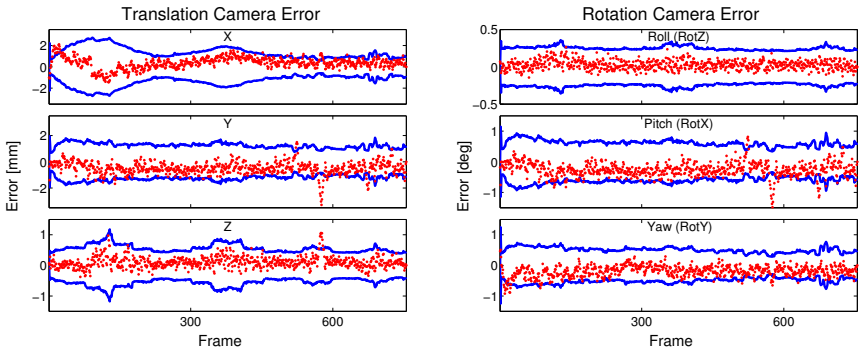

Figure 4.6: Laparoscopic 30º optics. Notice that the DoV (black arrow) does not correspond with the laparoscope main axis (red).

Let be two sets of camera locations $\mathbf{x}_G^{C_L}$ and $\mathbf{x}_S^{C_L}$:

$$\mathbf{x}_{(\star)}^{C_L} = \begin{pmatrix} \mathbf{x}_{1,(\star)}^{C_L} \\ \vdots \\ \mathbf{x}_{L,(\star)}^{C_L} \end{pmatrix}, \ (\star) = \{G|S\} \tag{4.1}$$

$$\mathbf{x}_{i,(\star)}^{C_L} = \left(X_{i,(\star)}^{C_L}, Y_{i,(\star)}^{C_L}, Z_{i,(\star)}^{C_L}, \phi_{i,(\star)}^{C_L} \theta_{i,(\star)}^{C_L}, \psi_{i,(\star)}^{C_L}\right)^\top \tag{4.2}$$

where each camera location is represented by its position $\left(X_i^{C_L} Y_i^{C_L} Z_i^{C_L}\right)^\top$ and its orientation in Roll-Pitch-Yaw angles $\left(\phi_i^{C_{Last}} \theta_i^{C_{Last}} \psi_i^{C_{Last}}\right)^\top$. These sets correspond with the ground-truth camera set and the SLAM set respectively and whose reference frames are attached to the last camera $C_L = \mathbf{x}_L^{C_L}$,




(a) 384x288 Camera translation error (red) and the corresponding $\pm 3\sigma$ acceptance region (blue).

(b) 384x288 Camera rotation error (red) and the corresponding $\pm 3\sigma$ acceptance region (blue).

Figure 4.7: Estimation camera error for the simulation results.

which is the same for both sets. The SLAM errors of the translation and rotation are computed as:

$$\epsilon = \oplus \mathbf{x}_G^{C_L} \ominus \mathbf{x}_S^{C_L} \ ; \tag{4.3}$$

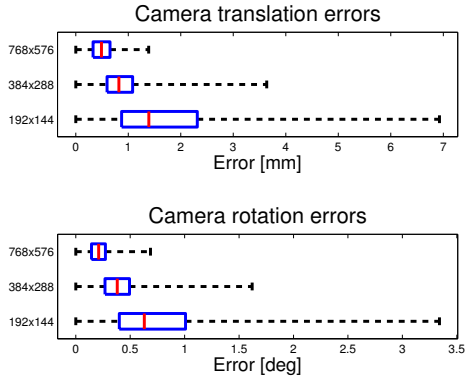
and the corresponding error covariance is computed by propagating the SLAM covariances:

$$\mathbf{P}_\epsilon = \mathbf{J}_{\epsilon S} \mathbf{P}_S^{C_L} \mathbf{J}_{\epsilon S}^\top \ , \qquad \mathbf{J}_{\epsilon S} = \frac{\partial \epsilon}{\partial \mathbf{x}_S^{C_L}} \tag{4.4}$$

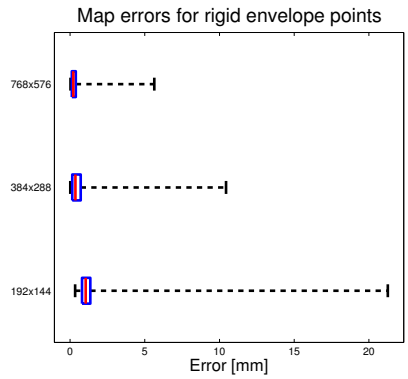
where $\mathbf{P}_S^{C_L}$ is the SLAM covariance aligned with respect to camera $C_L$. Notice that unlike (3.13), where the trajectory must be previously scaled, SLAM recovers a scaled trajectory because the virtual tool tip defines the real scale factor.

Figures 4.7a and 4.7b display the estimation error history for the camera translation and rotation respectively. Both the error and the $\pm 3\sigma$ acceptance region are represented. It can be concluded that the EKF provides a consistent estimation because the estimated value is mostly within the $3\sigma$ interval. Additionally, thanks to the covariance estimation, it can be evaluated how accurate the available estimation is at a given time step. The time evolution shows how initially the covariance grows due to the exploratory motion that departs from the initial camera location. As the estimation evolves, some features are reobserved and then the estimation error decreases.

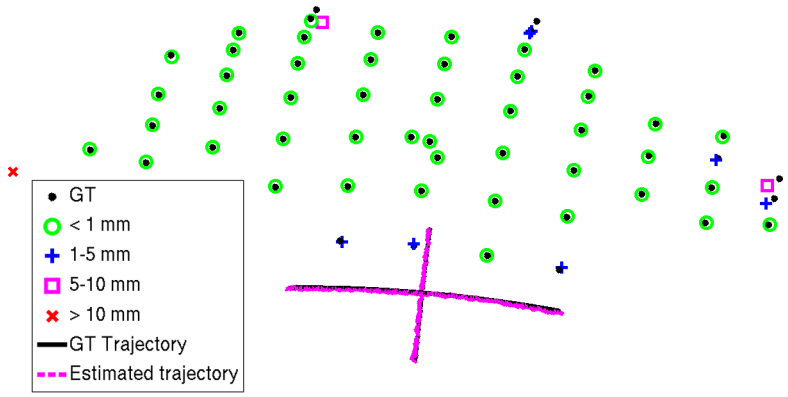
Figure 4.8a displays the estimation error distributions for the camera estimation history by means of box-and-whisker diagrams. The left and right



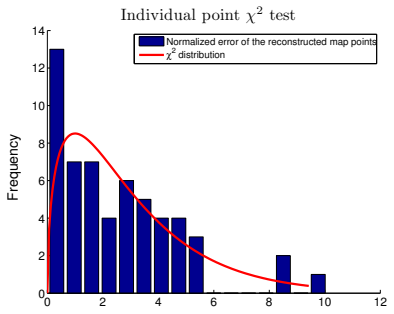

(a) Box-and-whisker diagrams for the camera location error in the three analyzed resolutions.



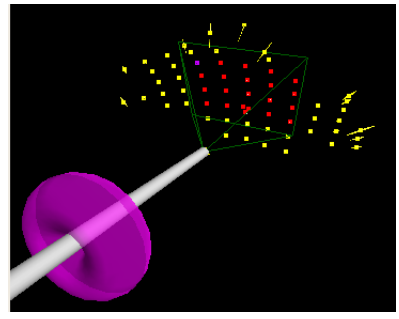

(b) Box-and-whisker diagrams for the error of the map points in the three analyzed resolutions.




(c) Estimated map point errors and camera trajectory after processing the whole $384 \times 288$ sequence. Notice that points whose location error is large, due to lack of parallax, are on the boundary.




(d) Map point error normalised with respect to the estimated covariance. It approximately distributes as the theoretical $\chi^2$ with 3 DoF



(e) Estimated map points and their covariances after processing the whole $384 \times 288$ sequence. The laparoscope (gray) is passing though the ground-truth fulcrum (magenta).

Figure 4.8: Estimation error for the simulation results.

of the box represent the first and third quartiles, the line inside the box is the median; the ends of the whiskers represent the minimum and maximum of all of the data. The errors are in the interval [0.6, 1.1] mm with 0.82 mm as the median for translation, and [0.27, 0.49] deg with 0.38 deg as the median for rotation.

The estimated map corresponds to the "rigid envelope" where none of the points in the cap are deformed. During the simulation, observations corresponding to non-rigid deformations are successfully marked as spurious by 1-PR and are not considered in the estimation (video [GGh] –2:12 - 2:15–). It is worth noting that if 1-PR is disabled, some spurious matches are marked as inliers and the estimation fails. For each time step and for each map point, the EKF provides both an estimate for the location and its covariance. As more images are processed, the covariance for a given point is reduced if the point is reobserved. Figure 4.8c displays the estimated map with absolute errors and Figure 4.8e displays the estimated map with the corresponding ellipsoidal $3\sigma$ acceptance regions after processing the whole sequence. Both figures show that most of the points have a small error except those at the map boundaries. Points on the boundary are only detected in a few images providing little parallax, hence, their location error is great. In any case, it has been verified that the estimation error normalized with the estimated covariance approximately distributes as a $\chi^2$ with 3 DoF (Figure 4.8d). It can be concluded that the map estimation is consistent, hence estimated covariances provide a per point accuracy measurement.

Figure 4.8b displays box-and-whisker diagrams for the estimation error for all the map points after processing the whole sequence. The errors are in the interval [0.15, 0.71] mm with 0.36 mm as the median, the maximum error is 10.44 mm corresponding to a point on the boundary.

From the $384 \times 288$ simulation, it can be concluded that the map estimation is accurate up to 1 mm for most of the points, in any case, the estimated covariance provides an assessment for each point accuracy. Regarding the effect of the camera resolution, the half and double resolution simulations show that EKF can make the most of the available resolution because error increases inversely with respect to the resolution (Figures 4.8a and 4.8b).

## 4.4 Experimental Validation Description

The goal of the experimental validation[1] is to prove the feasibility of using monocular visual SLAM in real surgical procedures. LHVR has been selected as a paradigmatic example because:

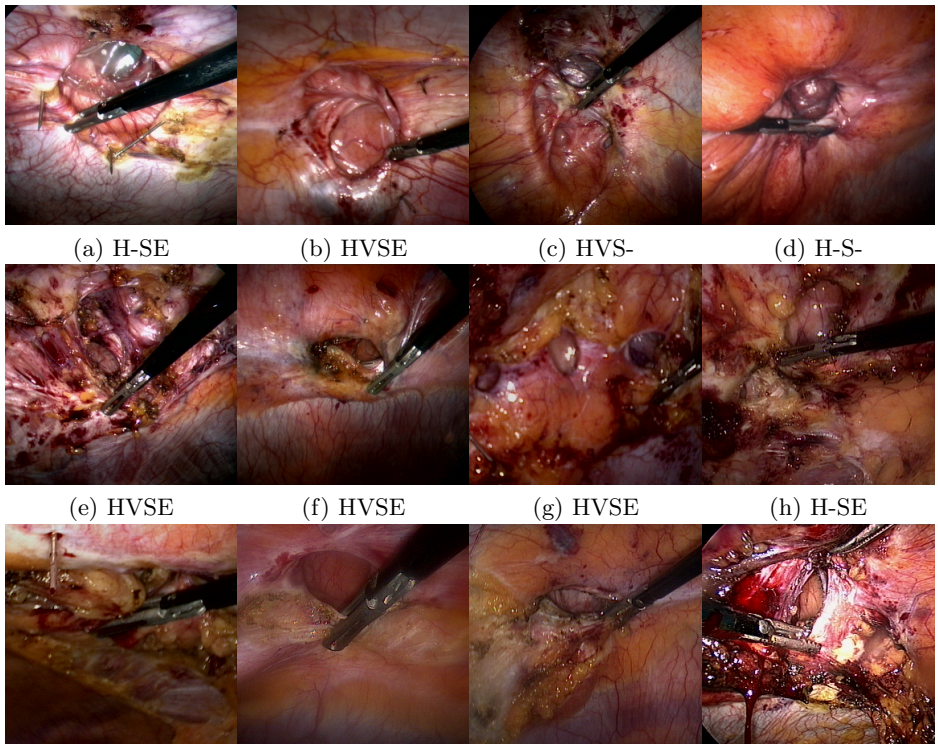
1. The scene is almost rigid and textured.
2. The standard procedure already includes accurate distance measurements that can be used as ground-truth to assess the visual SLAM geometrical accuracy.
3. The flexibility and robustness of visual SLAM methods are clearly tested because the surgical procedure has not been modified at all, except for the addition of an exploratory endoscope manoeuvre with a trajectory similar to other endoscope routine motions.
4. The SLAM version, just by making better use of the images, would simplify the surgical procedure without a disruptive modification of the workflow.
5. The image sequences exhibit significant inter-patient variability in texture, illumination, input port placement, and exploratory trajectory.

Fifteen in-vivo human LVHR interventions occurred between April 2011 and July 2012 were captured at 384x288@25 fps with an optics with 30º DoV (Figure 4.6) and 60º FoV angles (Figure 4.9 shows a thumbnail of each of them). The standard LVHR procedure has been extended with the additional exploratory endoscope maneuver (Section 4.2). For twelve of the operations, it was possible to take internal tape measurements for, at least, one main axis of the hernia (ground-truth) (Figures 4.13, 4.14, and 4.15). External measurements were taken for eleven of the operations (Figure 4.10 shows ten of them). The reasons for not taking some of the measurements were the difficulty in maneuverability or surgical time saving due to some patients' medical conditions.

### 4.4.1 Data Acquisition

The data acquisition equipment used during LVHR sequences consisted of a monocular endoscope (Image 1, Karl Storz, Germany) with a free PAL video

[1]The experiments developed in this work were approved by Comité Ético de Investigación Clínica de Aragón (CEICA) and governed according to the provisions of the Spanish Law *14/2007* regarding biomedical research.



(a) H-SE (b) HVSE (c) HVS- (d) H-S-

(e) HVSE (f) HVSE (g) HVSE (h) H-SE

(i) H-S- (j) H-S- (k) HVSE (l) –SE

(m) –SE (n) –SE (o) HV-E

Figure 4.9: The thumbnails –labeled from (a) to (o)– corresponding to the 15 ventral hernia repair surgeries used to validate the system. The "HVSE" code in the captions stands for the availability of (H) Horizontal tape measurement, (V) Vertical tape measurement, (S) SLAM measurement, and (E) External measurement. The SLAM map was successfully computed for all of them, while ellipse measurement was not possible in (o) due to the lack of texture around the defect.

output; a standard computer (Intel Core i7 CPU, 2.93 GHz, 4 GB RAM) equipped with a frame grabber; and a videocamera. In order to make the

most of each operation ensuring the data capture and minimizing technical problems during surgeries, the next action protocol for each operation was established:

1. *Computer connection with the endoscope*: The frame grabber of the computer was connected to a free PAL output of the endoscope. Frames were subsampled and captured at 384x288@25 fps in order to avoid the combing effect due to interlacing. The sequences were captured uncompress to avoid losses in the image quality.

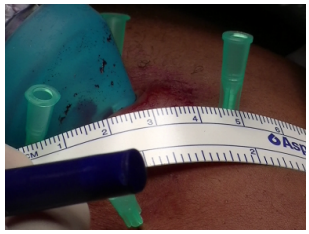
2. *Installation of the external camera*: All interventions except the corresponding to operation 4.9l were externally filmed with the external videocamera hanged on the roof lamps of the operating theatre. Video and audio of the external recordings were essential to: obtain external measurements; obtain measure times of the SLAM, internal, and external measuring methods; and capture the basic endoscope movements of the method. Figure 4.10 shows 10 out of 11 external measurements captured with the external camera.

3. *Endoscope configuration*: A correct endoscope illumination is mandatory for a correct visual SLAM performance. The endoscope illumination was configured to be greater than 60%.

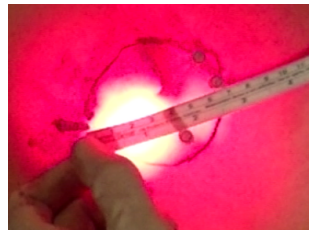
4. *Measurements*: At the moment of measuring, the three methods (SLAM, internal, and external) were carried out whenever was possible. In the case of SLAM method, the only one performed in the fifteen surgeries, a laparoscopic tool had to be fixed inside the laparoscope field of view. The other two are not available in all surgeries usually due to difficulty in maneuverability or patients' medical conditions.

5. *Laparoscope calibration*: When the surgery ended and previously to the optics removal from endoscope, a calibration planar pattern was imaged for laparoscope calibration according to Zhang's method [Zha00]. As the laparoscope has a 30º DoV, eight photos of the planar pattern were taken with a special laparoscopic positioning as shows Figure 4.11.

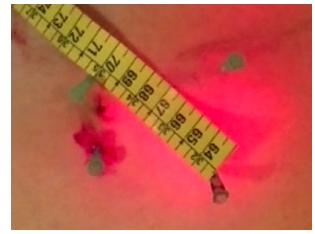
During surgeries all essential information was taken down. After the operation, this information was compared with the external and internal videos to ensure the correctness of data. The external recording for the intervention 4.9l is not available, therefore, the external measurements and times were taken trusting in operating room notes.



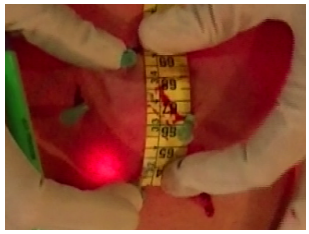
(a) Corresponds to Figure 4.9a.



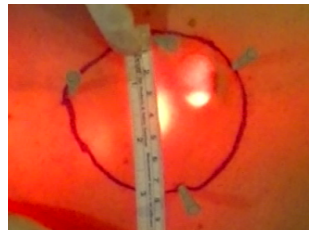
(b) Corresponds to Figure 4.9b.



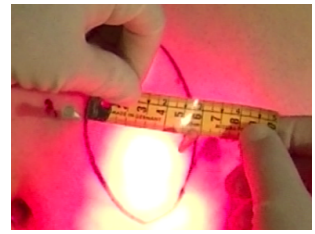
(c) Corresponds to Figure 4.9e.



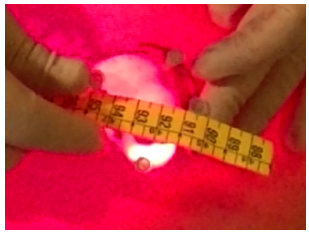
(d) Corresponds to Figure 4.9f.



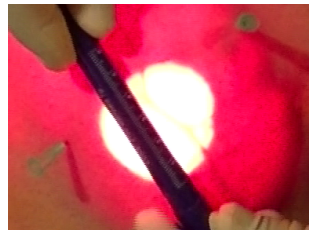
(e) Corresponds to Figure 4.9g.



(f) Corresponds to Figure 4.9h.

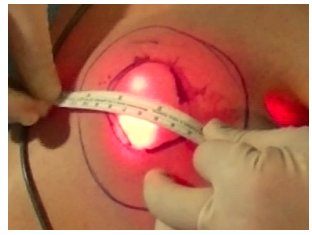


(g) Corresponds to Figure 4.9k.



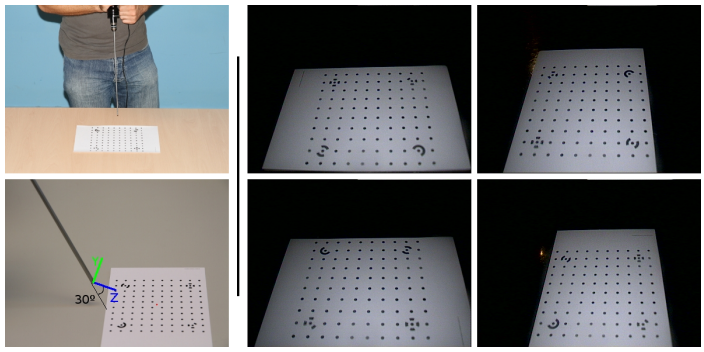
(h) Corresponds to Figure 4.9m.

(i) Corresponds to Figure 4.9n.

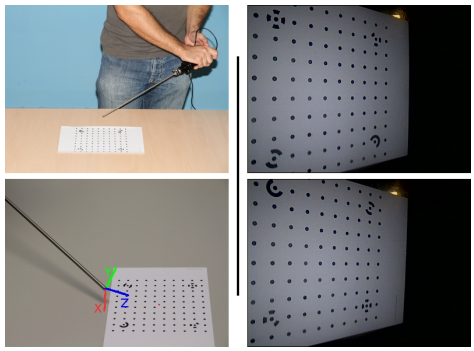


(j) Corresponds to Figure 4.9o.

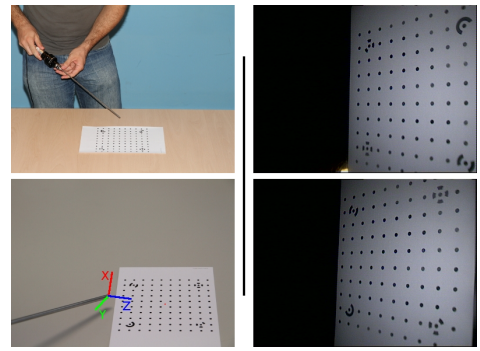
Figure 4.10: External measurements taken with the external videocamera corresponding to 10 out of 11 operations with external measurements.




(a) Two left images: laparoscope positioning, Z axis aims to the pattern center. Notice that the 30º optics DoV means that the Z axis is not the same that the laparoscope main axis. Four right images: planar pattern images, one per each side.



(b) Left images: laparoscope positioning, the laparoscope is 90º counter clock-wise rolled around the Z axis which aims to the pattern center. Right images: planar pattern images of two opposite sides.



(c) Left images: laparoscope positioning, the laparoscope is 90º clock-wise rolled around the Z axis which aims to the pattern center. Right images: planar pattern images of the another two opposite sides.

Figure 4.11: Procedure to take the calibration images. Laparoscope positioning and the eight images of the planar pattern needed to calibrate.

## 4.5 SLAM Engineering Validation

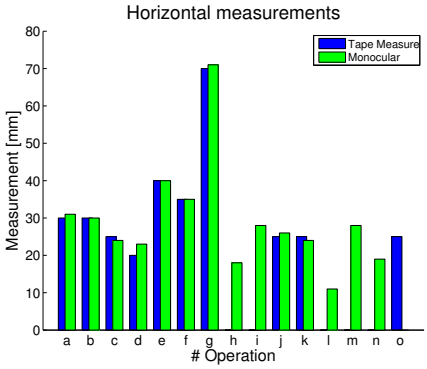
For the EKF SLAM validation, the same parameters, experimentally tuned, were applied for all of the experiments: image measurement error of 0.5 pixels standard deviation; 40% is the acceptance threshold for normalized correlation score to eventually accept a map point match in the new image; new features are assigned an initial 1 inverse depth, with an initial $\sigma_\rho = 1$, in order to have an initial direct depth acceptance region starting in 0.3 and extending to include infinite; regarding linear and angular accelerations,

standard deviations are $2.5\frac{1}{s^2}$ and $3\frac{rad}{s^2}$ respectively, as monocular cannot observe the scale, both depth and linear acceleration have no length units; finally, map management initializes features in order to have 40 map points observable in the image.

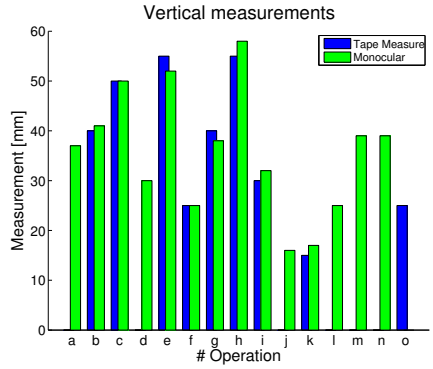
The proposed EKF SLAM has been able to successfully compute the map and the camera trajectory for the fifteen sequences (Figure 4.9). It has been able to cope with a variety of illuminations, textures and input port geometries. If a weakness has to be mentioned, it is the inability to perform the measurement in one of the sequences (Figure 4.9o) because of the lack of texture around the defect. In the rest of sequences, the EKF SLAM was always able to measure both ellipse axes because the defect visibility is required during the surgery and SLAM profits from that. In the failing case, the EKF SLAM was able to build the map; however, the clicked points signaling the defect were not trackable due to the lack of stable texture in the defect boundary area and the particular point detection method. A more dedicated work in image processing (e.g. using contours) is quite likely to overcome this limitation. In contrast, classical tape measuring procedure sometimes fails to produce the measurement because of the limited maneuverability resulting from the port placement.

The surgical time consumed by SLAM is mainly due to the exploration, which takes less than 1 minute irrespective of the sequence. Since the algorithm runs live (Figure 4.12c), no additional time is needed for the processing, except for selecting the points over defect boundary and over the forceps to define the scale. Both are easy to automate with the corresponding surgical time saving. In contrast, the internal tape measurement procedure, used as a ground truth in this validation, is rather uncertain (the time length ranges from 2 to 5 minutes). It has to be noted that in three cases where longer times were anticipated, the surgeons did not even try to measure. In any case, SLAM recovers not only two measurements but a full 3D model and the support for augmented reality.

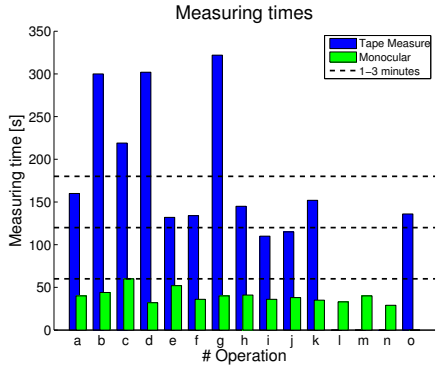
To validate the SLAM geometrical accuracy, the dimensions of the hernia defect's main axes have been estimated from the 3D recovered model and compared with those of internal tape measurement (the ground-truth), accurate up to 0.5cm. Figures 4.13, 4.14, and 4.15 show captures comparing the internal measurements with the SLAM measurements. No significant differences can be observed so it can be concluded that SLAM is as accurate as the internal tape measurement. Figures 4.12a - 4.12b depict measurements in the two axes.




(a) Horizontal axis measurement comparison.




(b) Vertical axis measurement comparison.

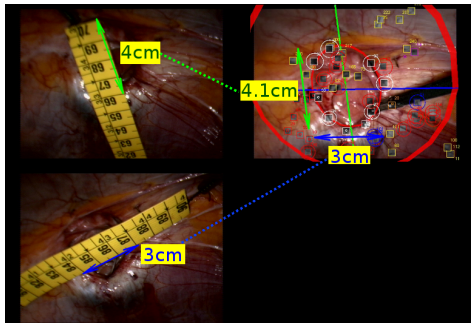



(c) Measuring time comparison.

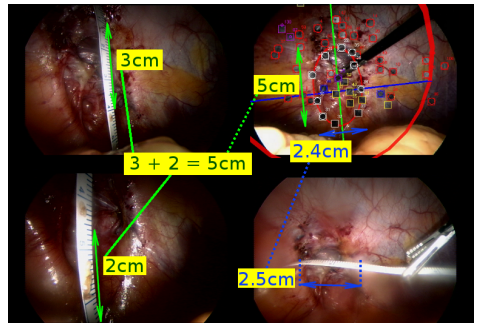
Figure 4.12: Measurement procedure comparison. Both accuracy (a), (b) and surgical time (c) are exhaustively plotted, one bar per operation per method. Missing data are represented as a missing bar. The labels correspond with those on Figure 4.9.

## 4.6 Clinical Validation

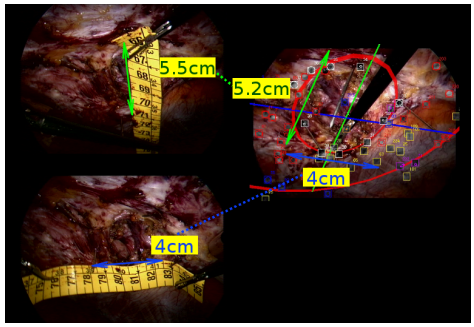
In addition to the engineering validation, a clinical validation has also been made. The clinical validation consists in a descriptive and comparative prospective study analyzing data from the fifteen LVHR. All LVHR were performed with a bilaminar intraperitoneal tissue-separating mesh. The mesh was fixed in place using transfascial non-absorbable sutures at the four cardinal points (four vertices of the hernia), and the edges of the mesh were fixed using absorbable tackers according to the double-crown technique [MC+05].



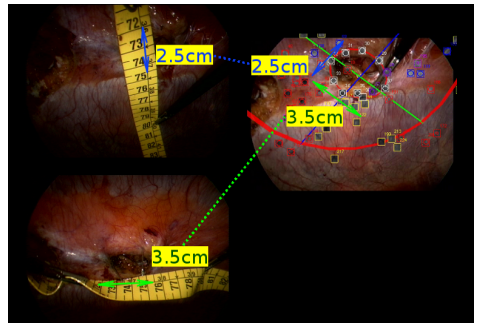

(a) Corresponds to Figure 4.9b.



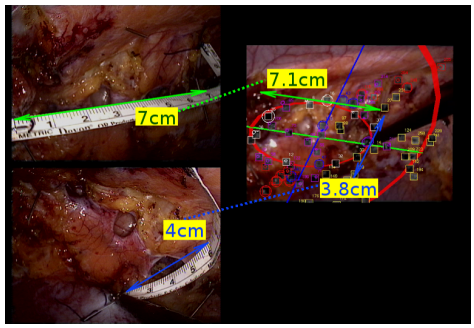

(b) Corresponds to Figure 4.9c.



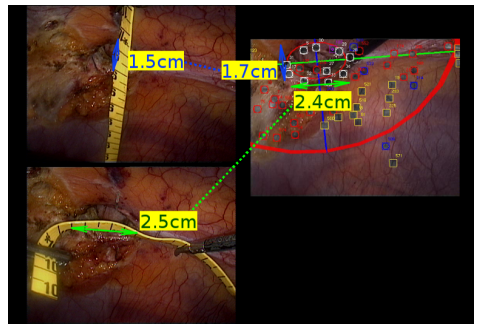

(c) Corresponds to Figure 4.9e.




(d) Corresponds to Figure 4.9f.




(e) Corresponds to Figure 4.9g.

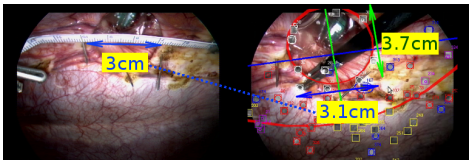



(f) Corresponds to Figure 4.9k.

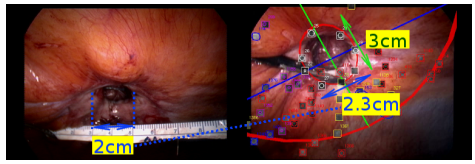
Figure 4.13: Comparison between ground-truth internal measurements and SLAM measurements (I). Internal measurements in both axes.

For each LVHR procedure, measurements were performed using three methods: the two classical methods (needle and tape. Figure 4.2), and the Visual SLAM Measurement (VSM) method.

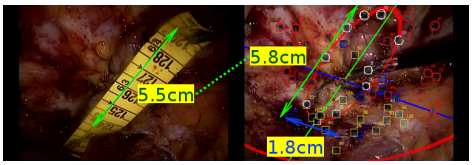
The study protocol, including the documents for obtaining informed con-



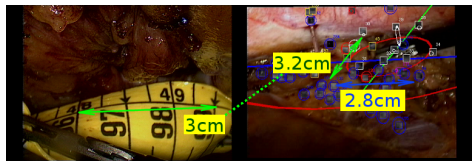

(a) Corresponds to Figure 4.9a.



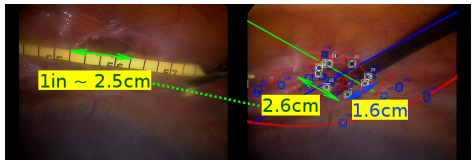

(b) Corresponds to Figure 4.9d.




(c) Corresponds to Figure 4.9h.

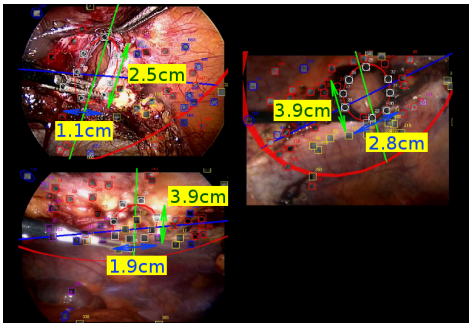



(d) Corresponds to Figure 4.9i.

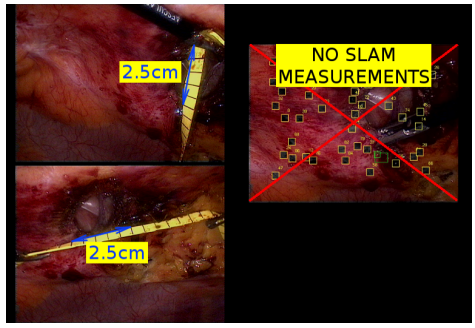



(e) Corresponds to Figure 4.9j.

Figure 4.14: Comparison between ground-truth internal measurements and SLAM measurements (II). Internal measurements only in one axis.




(a) Corresponds to Figures 4.9l, 4.9m, 4.9n




(b) Corresponds to Figure 4.9o.

Figure 4.15: Comparison between ground-truth internal measurements and SLAM measurements (III). (a) No internal but SLAM measurements. (b) No SLAM but internal measurements.

sent from patients, were approved by Comité Ético de Investigación Clínica de Aragón (CEICA) and were in accordance with the Spanish law 14/2007 regarding biomedical research.

### 4.6.1 Surgical procedure

Antibiotic and antithrombogenic prophylaxis were administered to all patients. Abdominal access was established. The edges of the parietal defect were drawn on the skin, guided by tactile localization. A pneumoperitoneum was created by inserting a Veress needle into the left upper quadrant. A pressure of 12 mmHg was used to safely separate the viscera from the abdominal wall. Three trocars were placed along a line as far as possible from the hernia defect: two 5-mm diameter working trocars, and a central 10-mm diameter trocar for the camera and for inserting the prosthetic mesh (Figure 4.16). A camera with 30$^{o}$ DoV (Figure 4.6) was used to examine the anterior abdominal wall, particularly the areas around the trocars. The abdominal cavity was explored to locate the viscera, identify adhesions, and locate and evaluate all hernia defects. 13 patients had a defect in the central abdominal wall or the right flank, and in these patients the ports were placed in the left flank. 2 patients had a defect in the left flank, and in these patients the ports were placed in the right flank.

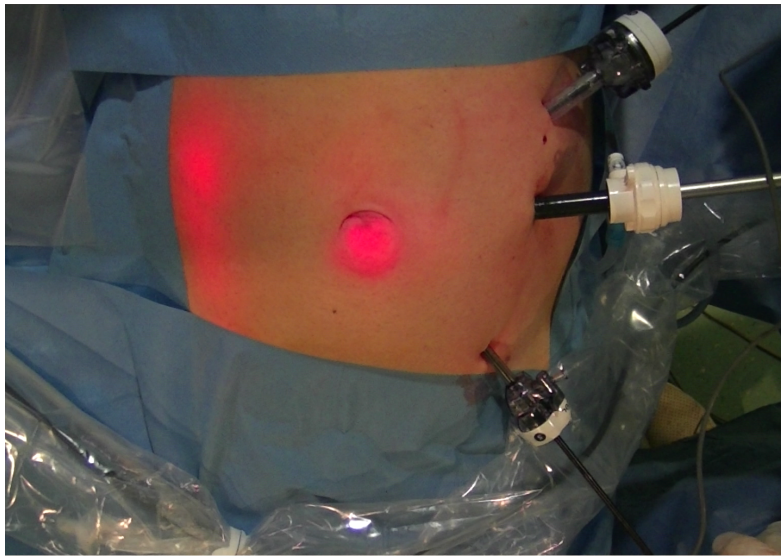


Figure 4.16: Trocar locations in the left flank.

After creating the pneumoperitoneum, the fat and visceral adhesions were dissected from the hernia sac. Adhesiolysis was performed at the borders of the hernia defect to locate the edges of the intact abdominal wall. For adhesions close to the intestines, monopolar coagulation was avoided to avoid inadvertent perforation.

To assess the size of the defect without enlarging the hernia, the pneumoperitoneum pressure was reduced to 8 mmHg. The two diameters of the defect were measured to determine the required size of the prosthetic mesh using the three measurement methods. First, four needles were placed through the abdominal wall under endoscopic guidance to determine the sizes of the two main axes of the defect, which was considered to be elliptical in shape; after correct insertion of the needles, an external tape measure was used to measure the distances between them (Figures 4.2a and 4.2b). Second, a sterilized tape measure was introduced into the abdominal cavity to measure the two axes of the defect (Figure 4.2c). Third, the defect was measured using the VSM method; the surgeon fixed a forceps inside the abdominal cavity and moved the tip of the laparoscope in a cross-shape, keeping the tip of the forceps and the defect in the field of view (Figure 4.3c); after the surgery was finished, the endoscopic sequence was processed to estimate the size of the defect (Figure 4.4a); in the first image of the sequence, several points were marked: two predefined points on the forceps, whose relative distance was known, to define the scale, and five or more points at the borders of the defect to estimate the hernia contour and size (Figure 4.3a).

The mesh was rolled along its major axis and grasped with forceps to insert it through the 10 mm trocar. Inside the abdominal cavity, the mesh was unrolled and oriented to cover the borders of the hernia defect. The mesh was fixed at the four cardinal points with non-absorbable monofilament sutures, and then fixed along the edges with tackers according to the double-crown technique, with 1 cm between tackers. An abdominal compression bandage was applied postoperatively. Oral ingestion was started after 8 hours and ambulation was started after 12 hours. Patients returned for a follow-up visit after 30 days.

The main steps of the whole procedure for the LVHR are shown in this video [GGg]. Besides, the video [GGi] shows the needle and the internal tape measurement methods.

### 4.6.2 Results

Fifteen ventral hernias were repaired (Figure 4.9), 9 females (60%) and 6 males (40%). The mean patient age was 42 years (range, 27–69 years). Ten patients (67%) had recurrent hernias and five (33%) had primary hernias. The mean operation time was 80 min (range, 40–120 min). Patient comorbidities included obesity (n = 9), hypertension (n = 7), smoking and alcoholism (n = 3), diabetes mellitus (n = 2), chronic obstructive pulmonary disease (n = 2), ischemic heart disease (n = 2), chronic renal failure (n = 1), and human immunodeficiency virus infection (n = 1). Six of the patients

did not have any comorbidities. The size of the defect ranged from $1 \times 2$ cm to $4 \times 7$ cm. There were no cases of seroma, hematoma, relapse, infection, or other complications related to the prosthetic material.

Figures 4.17a and 4.17b show the measurements obtained by the three methods: needles, tape, and VSM. The VSM method failed in one patient because of particularly poor image quality (Figure 4.9o), but both axes could be measured using the VSM method in all the other patients (93%). Regarding needle and tape methods, 7 patients had extensive intra-abdominal adhesions whose laparoscopic adhesiolysis was very time-consuming. As these patients were classified as ASA III patients and suffered from intraoperative hemodynamic instability during anesthesia, only one of the two measurement methods was used in order to minimize the operation time; the tape method was applied in 4 patients (Figures 4.9c, 4.9d, 4.9i, 4.9j) and the needle method was applied in 3 patients (Figures 4.9l, 4.9m, 4.9n). The tape method was finally applied in 12 patients obtaining 19 out of 24 measurements (79%; 1 measurement per hernia defect axis), or 19 out of 30 measurements (63%) if all 15 patients are considered. The main reason for inability to perform all measurements was the difficulty of the procedure because of limited range of motion of the laparoscopic tools. Finally, the needle method was only used in 11 out of 15 patients (73%).

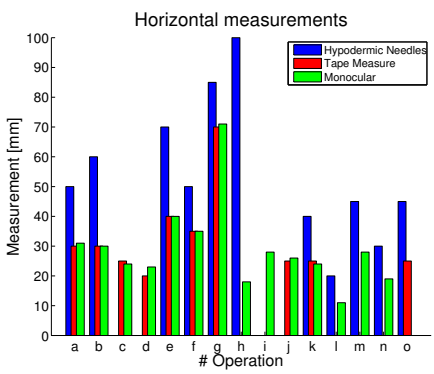
The accuracy of measurement methods was compared. The tape method was the most accurate (accuracy up to 0.5 cm because of tape resolution). The needle method was rather inaccurate, always resulting in an excessively large value (average excess of 3 cm). There were no significant differences between the VSM and tape methods, indicating that both methods are equally accurate.

Figure 4.17c shows the time taken to perform measurements. VSM was the fastest method with a mean time of 40 s (range, 29–60 s). The needle method had a mean time of 169 s (range, 66–300 s). The tape method had a mean time of 186 s (range, 110–322 s); note that this mean time would be greater if all 24 measurements had been obtained.

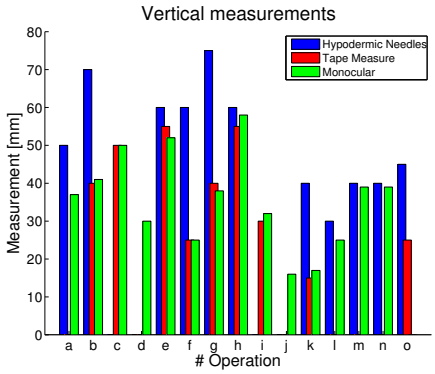
## 4.7 Conclusions and Future Work

Traditional endoscopic surgery displays and disposes of the image sequence. However, monocular SLAM, with the addition of an exploratory maneuver, is able to exploit the sequence allowing to use it to estimate measurements.

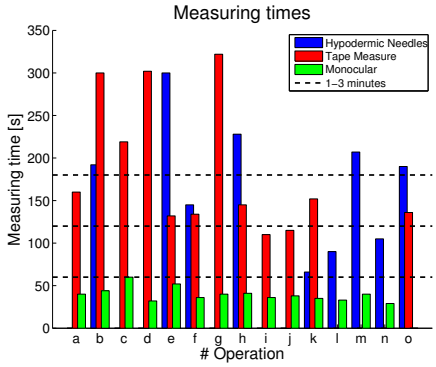
This thesis provides the first human in-vivo experimental validation for the feasibility of using EKF monocular SLAM as a proper method to deal with medical endoscope sequences. The 15 studied patients showed vari-




(a) Horizontal measurements of hernia defect size using the three different methods.




(b) Vertical measurements of hernia defect size using the three different methods.




(c) Times taken by the three different methods.

Figure 4.17: (a, b) Measurements of hernia defect size using the three different methods. Use of the tape obtained 63% of the total measurements, use of the needle obtained 73%, and use of VSM obtained 93%. Measurement using the tape is considered to be the most accurate. Measurement using VSM was found to be as accurate as measurement using the tape, but measurement using the needle was significantly less accurate. (c) Time taken to perform measurements using the different methods. The VSM method was the fastest.

ability in terms of textures, illumination, port placement, and exploratory trajectories. In spite of this variability, all the image sequences could be processed using the monocular SLAM method (VSM), indicating that this method is useful in a variety of situations.

The method has proved to be fast, non-invasive, and easily incorporated

to the existing LVHR surgical workflow by using solely images gathered from a hand-held standard monocular endoscope, standard laparoscopic tools and a simple cross-shaped motion performed by the surgeon.

### 4.7.1 Engineering Validation

A scene rigid model is assumed; however, thanks to 1-PR, the method has proven robust with respect to scene local non-rigid deformations such as respiration or external forces. The validation is based on synthetic data and on sequences coming from a real surgical environment over fifteen human in-vivo laparoscopic ventral hernia repair surgeries.

Unlike other experimental validations based on phantoms or animal imagery, the method has been tested over fifteen human surgeries that displayed the typical inter-patient variability (different textures, illumination, input port placement and exploratory trajectories) (Figure 4.9). Despite the variability, all the sequences have been processed with the same tuning, therefore they provide experimental evidence of the method usability.

The accuracy of the EKF monocular SLAM + 1-PR has been proved. In any case, any real-time visual SLAM method, either monocular or stereo, would perform equally well on condition that it has a robust-to-spurious policy.

The method cannot deal with non-rigid nor with textureless scenes. Besides, offline camera calibration is required.

Regarding the non-rigidity, Agudo *et al.* [ACM12b; ACM12a] have proved that the combination EKF-FEM (Finite Elements Method) can deal with deformations in real time. This approach is relevant for medical images because it can exploit the biomechanical availability. One of the immediate goals is to adapt these methods to the system. Concerning calibration, it would be desirable to solve the complete problem (3D structure recovery, camera location and camera calibration) during the exploratory movement. Finally, the lack of texture could be tackled using a monocular SLAM based on points and edges and researching the combination with photometric methods.

In the particular case of the hernia measurement, another minor limitation is that currently the scale and the hernia defect have to be selected by clicking on the images; an automatic detection of both would ease the use of the system.

Since the system is based on an EKF implementation, it only can handle a few hundred points. However, methods based on keyframe + bundle adjustment such as [KM07] can render a map composed of a few thousand points. This signals a clear way for increasing the map density.

Finally, a more ambitious goal is to exploit the camera location as a backbone for augmented reality providing additional visual information (multimodal registration images –CT or MRI– or another kind of annotations) in real time.

### 4.7.2 Clinical Validation

Compared with the traditional surgical methods of treating ventral hernias, LVHR in general and the VSM method in particular have irrefutable benefits, including accurate confirmation of the diagnosis and objective measurement of the hernia defect. These techniques can be used to identify and measure both the main defect and secondary defects, to ensure that the implanted mesh will cover all defects.

Use of VSM to perform measurements during LVHR minimizes the risk of infection, because VSM prevents exposure of the abdominal cavity to additional external instruments. Unlike the needle method, the VSM method does not cause injury to areas with scar tissue, and therefore does not expose the patient to the risk of contamination from areas of inflammation or microabscesses resulting from previous laparotomy.

This study focused on perfecting a method of measuring the size of the hernia defect. To date, no significant assessment of measurement method has been reported in the literature. The needle method provides approximate measurements, but the measurements tend to be too large because the needles are not inserted perfectly perpendicularly. This method is also invasive and carries a risk of hemorrhage if a blood vessel is injured. The tape method is accurate but is difficult to perform and can be time consuming making the measurement unfeasible. The VSM method is non-invasive, fast, and accurate.

On the contrary to the other methods assessed, VSM does not need additional tape measure or needles, which simplifies the workflow. The system was able to obtain accurate measurements. This system can be extended to support augmented reality insertions, to guide the surgeon during alignment of the prosthesis with respect to the border of the hernia defect and increase the ease of the procedure. Augmented reality insertions may also be used to display preoperative information to provide assistance during the procedure.

It would be interesting to extend this method to other surgical procedures such as flexible endoscopy and thoracoscopy. This method may also be useful for intra-abdominal measurements of organs such as the spleen or adrenal glands, to determine the required size of the extraction ports.

# 5 Conclusions and Future Work

## 5.1 Conclusions

From a robotics and computer vision point of view, laparoscopy can be posed as a monocular SLAM problem. In traditional laparoscopy, images gathered by laparoscope are shown in a screen and then disposed of. However, if laparoscopy were treated as a monocular SLAM problem, these images would be exploited, recovering, in real time, a 3D reconstruction of the abdominal cavity and the laparoscope localization with respect to that reconstruction.

SLAM algorithms have been thoroughly studied and validated in mobile robotics environments (indoors, outdoors, man-made environments, ...). However, not any previous work to this thesis, and devoted to applying SLAM in some endoscopic technique (endoscopy, laparoscopy, colonoscopy, ...), has extensively validated this type of algorithms. All of them make subjective validations, analyzing the appearance of the reconstruction, or objective validations by means of phantoms, ex-vivo data, in-vivo data from animal, or using additional devices only applicable in lab environments. This lack of validation makes unfeasible an immediate SLAM use in clinical applications.

This thesis has shown the feasibility of using these algorithms inside a clinical environment. For that, 15 real human laparoscopic sequences corresponding to ventral hernia repairs have been used to perform an exhaustive validation. In this type of operations, the surgeon needs to measure the dimensions of the hernia defect. These dimensions have been used as a ground truth to validate the monocular SLAM reconstructions. Additionally, several simulations with different system configurations have also been performed. Both real sequences and simulations have shown that it is possible to obtain 3D reconstructions in real time (25 fps) with millimetric errors. On the

other hand, the validation with real sequences has shown the robustness of this type of algorithms in the presence of inter-patient variability (different textures, illumination, input port placement and exploratory trajectories) since all sequences were processed with the same tuning.

In order to show the feasibility of these algorithms in laparoscopy, an EKF-based SLAM has been chosen. It has been selected the EKF technique because it is mature, thoroughly known, and works well in real time ($25 \sim 30$ fps) in small environments.

This thesis demonstrates that a monocular SLAM system will work in laparoscopy provided that it has implemented a robust and efficient spurious rejector system, and a good relocalization system. Due to the intrinsic nature of laparoscopic images, they are prone to contain a great number of spurious matches and to suffer from losses of tracking.

The main sources of spurious generation and losses of tracking are: temporary deformations caused by respiration, heartbeats, or external forces like forces exerted by laparoscopic tools; occlusions caused by tools or even by tissues or organs; blurred images; sudden laparoscope motions; or the extraction and reinsertion of the laparoscope into the abdominal cavity. These issues have been efficiently solved with the proposed 1-PR spurious detector and with the RLR system [WKR07].

The SLAM algorithm assumes that the scene is completely rigid. This assumption along with the 1-PR treatment of spurious have allowed that small deformations being considered as mismatches. In this way, it has been prevented that deformable features being integrated in the scene estimation. In the case they were integrated, the estimation would degrade, and even wreck, leading to a complete SLAM system failure. Finally, this way of treating small deformations has given rise to small maps with a tens of rigid points that are easily identifiable, reobservable, and well distributed along the scene. This kind of map allows that the RLR algorithm works efficiently in the presence of tracking losses, enabling monocular SLAM to process relatively long sequences of intracavitary explorations.

Recovered reconstructions have been demonstrated to be useful to synthetically recover the lost FoV by means of photorealistic reconstructions; to recover the lost depth caused by working with 2D images; to allow to make 3D distance measurements inside of body, or even they could be used to make surface measurements. In addition, they are used as a backbone for augmented reality annotations.

From a clinical point of view, these methods have shown to reduce the time taken by operations (that means less anesthesia required) and to provide security (both avoiding to introduce external devices inside the body and

enabling the possibility of making augmented reality annotations). Additionally, these methods have shown to be non-invasive and easily incorporated in routine surgical procedures without disturbing either surgeons or classical procedures. Therefore, in the future, these methods will become an essential surgical tool in the surgeon's armory.

## 5.2 Future Work

Despite the promising results shown in this thesis, monocular SLAM in laparoscopy, and in endoscopy in general, still presents several issues that must be solved before a real application inside the surgical room.

In the first place, the camera calibration problem should be faced with. Currently, camera calibration is performed after surgical intervention in order to avoid a possible contamination of the laparoscope, which is sterilized previous to the operation. This is one of the main issues because these algorithms are not usable in a surgical room yet. The ideal SLAM system should solve the complete problem (estimate the 3D structure of the scene, the camera location with respect to this structure, and the camera calibration parameters) during exploratory movements inside the abdominal cavity.

In the second place, these algorithms work with point features extracted from images, thus they cannot deal with textureless scenes. An interesting research would be to tackle this problem by means of SLAM systems that handle point features, edges and regions of interest, or even combining them with photometric methods.

In the third place, current SLAM methods assume that the scene is completely rigid. This assumption is extremely strong for internal scenes of the body. Recently, there is a great research in deformable SLAM field. Works such as those of Agudo *et al.* [ACM12b; ACM12a] have proved that the combination EKF-FEM can deal with deformations in real time. This approach is relevant for medical images because it can exploit the biomechanical characteristics of the tissues in order to support possible deformations instead of treating them as spurious.

In the fourth place, for the particular SLAM case presented in this thesis, EKF has a quadratic computational cost in the state size (directly related with map size), therefore, it only handles a few hundred points in real time. Adapting any method based on keyframes + BA such as the proposed by Klein & Murray [KM07] would be an interesting work. These methods enable to work with a few thousand points rendering dense scene maps that help a better understanding of the scene.

In the fifth place, this thesis has proposed to use the 3D reconstruction

as a backbone for augmented reality. The use of augmented reality has been shown with simple annotations over laparoscopic images and it has been mentioned its possible use along with multimodal registration images –CT or MRI– in real time. A research along this line would be very interesting and relevant since it would allow to show patient's preoperative data in real time during intervention. This would help tremendously surgeon's work during surgical procedures that entail a high level of risk.

Finally, this thesis proposes to apply SLAM techniques in laparoscopy and demonstrates their feasibility over 15 ventral hernia repairs. It would be interesting, from a clinical point of view, to search for other surgical procedures that could benefit from SLAM results. Some examples are thoracic surgery (thoracoscopy), joint surgery (arthroscopic surgery or arthroscopy), or gastrointestinal tract surgery (endoscopy, colonoscopy).

## 5.3 Conclusiones

Desde el punto de vista de la robótica y la visión por computador, la laparcosopia se puede interpretar como un problema de SLAM monocular. En la laparoscopia tradicional las imágenes capturadas por el laparoscopio únicamente son mostradas en un monitor para posteriormente ser desechadas. Sin embargo, si se tratase la laparoscopia como un problema de SLAM monocular, esas imágenes serían explotadas recuperando en tiempo real una reconstrucción 3D de la cavidad abdominal al mismo tiempo que se localizaría el laparoscopio con respecto a esa reconstrucción.

Los algoritmos de SLAM han sido profundamente estudiados y validados en entornos de robótica móvil (exteriores, interiores, construcciones humanas, ...), sin embargo, ningún trabajo anterior a esta tesis, y dedicado a aplicar estos algoritmos sobre técnicas endoscopicas (endoscopia, laparoscopia, colonoscopia, ...), ha validado de una forma extensiva este tipo de algoritmos. Estos trabajos hacen validaciones subjetivas, analizando la apariencia de la reconstrucción, o bien con maniquíes, datos ex-vivo, datos in-vivo de animales, o usando dispositivos adicionales lo que hace que queden bastante lejos de una posible inmediata aplicación clínica.

En esta tesis se ha demostrado la viabilidad de estos algoritmos dentro de un entorno clínico mediante la realización de una validación experimental exhaustiva con 15 operaciones reales de hernia ventral. En este tipo de operaciones el cirujano necesita medir las dimensiones del defecto herniario. Estas dimensiones han sido usadas como referencia para comprobar las reconstrucciones obtenidas por el SLAM monocular. Además de la validación con secuencias reales, también se han realizado simulaciones con diferentes config-

uraciones del sistema. Tanto las secuencias reales como las simulaciones han mostrado que se puede obtener reconstrucciones en tiempo real (25 fps) con errores milimétricos. Por otra parte, la validación sobre las 15 operaciones ha demostrado la robustez de estos algoritmos ante la variabilidad interpaciente (diferentes texturas, iluminaciones, disposiciones de los trocares y trayectorias exploratorias) ya que todas las secuencias han sido procesadas con los mismos parámetros de configuración.

Para mostrar la viabilidad de estos algoritmos en laparoscopia, se ha elegido un algoritmo de SLAM basado en EKF. Se ha seleccionado esta técnica de SLAM por ser una técnica madura, profundamente conocida, y que funciona bastante bien y en tiempo real ($25 \sim 30$ fps) en entornos reducidos.

En esta tesis se ha demostrado que un sistema de SLAM monocular funcionará correctamente en laparoscopia siempre y cuando tenga implementado un sistema robusto y eficaz de detección y rechazo de espurios, y un sistema de detección de pérdida del *tracking* con su posterior relocalización. Debido a la naturaleza intrínseca de las imágenes laparoscópicas, estas son propensas a contener grandes cantidades de espurios además de sufrir pérdidas de *tracking*.

Las fuentes principales de generación de espurios y de pérdidas de *tracking* son: la presencia de deformaciones temporales causadas por la respiración, los latidos del corazón o por fuerzas externas como las ejercidas por las herramientas; las oclusiones causadas por las herramientas o incluso por tejidos u órganos; imágenes borrosas; movimientos repentinos del laparoscopio; o la extración y reinserción del laparoscopio dentro de la cavidad abdominal. Todos estos problemas son eficazmente resueltos con el algoritmo 1-PR propuesto para el tratamiento de espurios y con el sistema de relocalización RLR [WKR07].

El algoritmo de SLAM utilizado asume que la escena es completamente rígida. Esta asunción de rigidez junto con la asociación de datos robusta del 1-PR han permitido que pequeñas deformaciones hayan sido consideradas como espurios. De esta forma, se ha impedido una posible integración de características deformables dentro de la estimación de la escena, lo que habría causado una degradación de esta e incluso un fallo completo del sistema de SLAM. Finalmente, este tratamiento de las deformaciones ha dado lugar a mapas de unas decenas de características rígidas fácilmente identificables, reobservables y bien distribuidas a lo largo de la escena. Este tipo de mapa permite que el algoritmo RLR se relocalice de una forma bastante eficiente ante posibles pérdidas de *tracking*, habilitando el procesamiento de secuencias de exploraciones intracavitarias relativamente largas.

En cuanto a las reconstrucciones obtenidas, estas han demostrado ser

útiles para ampliar sintéticamente el FoV perdido mediante reconstrucciones fotorrealistas; recuperar la profundidad perdida por trabajar con imágenes 2D; permitir realizar mediciones de distancias 3D en el interior del cuerpo, e incluso se podrían realizar mediciones de superficies; y soportar anotaciones en realidad aumentada.

Desde el punto de vista clínico, estos métodos han demostrado reducir el tiempo de las operaciones (menos anestesia para el paciente) y aportar seguridad (tanto por evitar la posible introducción de elementos extraños dentro del cuerpo, como por la posibilidad de realizar anotaciones en realidad aumentada). Adicionalmente, estos métodos son no invasivos y fáciles de incorporar en las rutinas quirúrgicas, sin llegar a ser una molestia para el cirujano ni interferir con los procedimientos habituales. Por lo tanto, en el futuro, estos métodos se pueden convertir en una nueva herramienta imprescindible dentro del arsenal quirúrgico del cirujano.

## 5.4 Trabajo Futuro

A pesar de los resultados prometedores mostrados en esta tesis, el SLAM en laparoscopia, y endoscopia en general, todavía presenta ciertos problemas que deben de ser solventados antes de tener un sistema para uso en quirófano.

En primer lugar está el problema de calibración de la cámara. Actualmente la calibración se realiza tras la intervención quirúrgica para evitar una posible contaminación del laparoscopio, el cual está esterelizado, antes de la operación. Este es uno de los principales motivos por el que estos algoritmos aún no se pueden utilizar dentro de quirófano. El sistema de SLAM ideal sería aquel que permitiese resolver el problema completo (estimación de la estructura 3D de la escena, de la localización de la cámara y de su calibración) durante los movimientos exploratorios dentro de la cavidad abdominal.

En segundo lugar, estos sistemas, al funcionar sobre características puntuales extraídas de las imágenes, no soportan escenas sin textura. Sería interesante abordar este problema mediante la utilización de sistemas de SLAM que soporten, además de características puntuales, segmentos y regiones de interés o incluso combinarlos con métodos fotométricos.

En tercer lugar, los métodos actuales de SLAM asumen que la escena es completamente rígida. Esta asunción es muy fuerte para el interior de cavidades corporeas. Actualmente hay una gran investigación en el campo del SLAM en escenas deformables. Trabajos como los presentados por Agudo et al. [ACM12b; ACM12a] han demostrado que la combinación de SLAM con elementos finitos pueden tratar las deformaciones en tiempo real. Este acercamiento es bastante relevante para el caso de las imágenes médicas ya

que permitiría explotar las caracteríticas biomecánicas de los tejidos para soportar las posibles deformaciones sin tener que tratarlas como espurios.

En cuarto lugar, para el caso particular presentado en esta tesis, el SLAM empleado está basado en una implementación en EKF, la cual tiene un coste computacional cuadrático en el tamaño del estado (tamaño del mapa). Por lo tanto, solo se pueden manejar mapas de unos pocos cientos de puntos. Una investigación interesante sería el adaptar algún método basado en *keyframes* + BA como el propuesto por Klein y Murray [KM07]. Estos métodos permiten trabajar con miles de puntos obteniendo reconstrucciones densas de la escena lo que ayudaría a una mejor comprensión de esta.

En quinto lugar, en esta tesis se ha propuesto utilizar la reconstrucción 3D de la escena como soporte para realidad aumentada. El uso de la realidad aumentada se ha mostrado con simples anotaciones sobre las imágenes de laparoscopia, y se ha nombrado su posible utilización junto con registro multimodal de imágenes de TAC o MRI en tiempo real. Una investigación en esta línea sería muy interesante y relevante ya que permitiría mostrar datos preoperativos del paciente en tiempo real durante la operación, ayudando enormemente al trabajo del cirujano en los procedimientos quirúrgicos que conlleven un elevado nivel de riesgo.

Finalmente, en esta tesis se ha propuesto emplear las técnicas de SLAM en laparoscopia y se ha demostrado su aplicación sobre 15 eventroplastias. Sería interesante, desde el punto de vista clínico, buscar otros procedimientos, no solo laparoscópicos, que se pudieran beneficiar de los resultados obtenidos por el SLAM. Algunos ejemplos podrían ser la cirugía torácica (toracoscopia), la cirugía en articulaciones (artroscopia), o la relacionada con el tubo digestivo (endoscopia, colonoscopia).

# Bibliography


[ACM12a] Antonio Agudo, Begoña Calvo, and J. M. M. Montiel. "3D Reconstruction of Non-Rigid Surfaces in Real-Time Using Wedge Elements". In: *5th Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (ECCV)*. Vol. 7583. 2012, pp. 113–122. DOI: 10.1007/978-3-642-33863-2_12.

[ACM12b] Antonio Agudo, Begoña Calvo, and J. M. M. Montiel. "Finite Element based Sequential Bayesian Non-Rigid Structure from Motion". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 1418–1425. DOI: 10.1109/CVPR.2012.6247829.

[Ban+12] Ambar Banerjee et al. "Laparoscopic ventral hernia repair: Does primary repair in addition to placement of mesh decrease recurrence?" In: *Surgical Endoscopy* 26.5 (2012), pp. 1264–1268. DOI: 10.1007/s00464-011-2024-3.

[Ber+] Ernesto Bernal et al. "Computer vision distance measurement from endoscopic sequences. Prospective evaluation in laparoscopic ventral hernia repairs." In: *Surgical Endoscopy* . Under revision ().

[BMG09] Jose-Luis Blanco, Francisco-Angel Moreno, and Javier Gonzalez. "A collection of outdoor robotic datasets with centimeter-accuracy ground truth". In: *Autonomous Robots* 27.4 (2009), pp. 327–351. DOI: 10.1007/s10514-009-9138-7.

[Bur+05] Darius Burschka et al. “Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery”. In: *Medical Image Analysis* 9.5 (2005), pp. 413–426. DOI: 10.1016/j.media.2005.05.005.

[Can86] John Canny. “A Computational Approach to Edge Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[Cap05] David Capel. “An Effective Bail-out Test for RANSAC Consensus Scoring”. In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2005, pp. 78.1–78.10. DOI: 10.5244/C.19.78.

[Cas+99] JA Castellanos et al. “The SPmap: a probabilistic framework for simultaneous localizationand map building”. In: *IEEE Transactions on Robotics and Automation* 15.5 (1999), pp. 948–952. DOI: 10.1109/70.795798.

[CB12a] Toby Collins and Adrien Bartoli. “3D Reconstruction in Laparoscopy with Close-Range Photometric Stereo”. In: *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 7511. 2012, pp. 634–642. DOI: 10.1007/978-3-642-33418-4_78.

[CB12b] Toby Collins and Adrien Bartoli. “Towards Live Monocular 3D Laparoscopy Using Shading and Specularity Information”. In: *Int. Conf. on Information Processing in Computer-Assisted Interventions(IPCAI)*. Vol. 7330. 2012, pp. 11–21. DOI: 10.1007/978-3-642-30618-1_2.

[CD08] Margarita Chli and Andrew J. Davison. “Active Matching”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2008, pp. 72–85. DOI: 10.1007/978-3-540-88682-2_7.

[CDM08] Javier Civera, Andrew J. Davison, and J. M. M. Montiel. “Inverse Depth Parametrization for Monocular SLAM”. In: *IEEE Transactions on Robotics (T-RO)* 24.5 (2008), pp. 932–945. DOI: 10.1109/TRO.2008.2003276.

[Civ+09a] Javier Civera et al. “1-Point RANSAC for EKF-Based Structure from Motion”. In: *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. 2009, pp. 3498–3504. DOI: 10.1109/IROS.2009.5354410.

[Civ+09b] Javier Civera et al. “Drift-Free Real-Time Sequential Mosaicing”. In: *Int. Journal of Computer Vision (IJCV)* 81.2 (2009), pp. 128–137. DOI: 10.1007/s11263-008-0129-5.

[Civ+10] Javier Civera et al. “1-Point RANSAC for Extended Kalman Filtering: Application to Real-Time Structure from Motion and Visual Odometry”. In: *Journal of Field Robotics* 27.5 (Sept. 2010), pp. 609–631. DOI: 10.1002/rob.20345.

[Cle+07] Laura A. Clemente et al. “Mapping Large Loops with a Single Hand-Held Camera”. In: *Robotics Science and Systems*. 2007.

[CM08] Ondrej Chum and Jiri Matas. “Optimal randomized RANSAC”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 30.8 (2008), pp. 1472–1482. DOI: 10.1109/TPAMI.2007.70787.

[CN08] Mark Cummins and Paul Newman. “FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance”. In: *The International Journal of Robotics Research* 27.6 (2008), pp. 647–665. DOI: 10.1177/0278364908090961.

[Dan+07] Koppel Dan et al. “Toward Automated Model Building from Video in Computer-Assisted Diagnoses in Colonoscopy”. In: *Proc. of the SPIE Medical Imaging Conf.* 2007. DOI: 10.1117/12.709595.

[Dav+07] Andrew J. Davison et al. “MonoSLAM: Real-Time Single Camera SLAM”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 29.6 (2007), pp. 1052–1067. DOI: 10.1109/TPAMI.2007.1049.

[Dav03] Andrew J. Davison. “Real-Time Simultaneous Localisation and Mapping with a Single Camera”. In: *Int. Conf. on Computer Vision (ICCV)*. 2003, 1403–1410 vol.2. DOI: 10.1109/ICCV.2003.1238654.

[Dis+01] M. Dissanayake et al. “A solution to the simultaneous localization and map building (SLAM) problem”. In: *IEEE Transactions on Robotics and Automation* 17.3 (2001), pp. 229–241. DOI: 10.1109/70.938381.

[DW+03] H. Durrant-Whyte et al. “A Bayesian Algorithm for Simultaneous Localisation and Map Building”. In: *Robotics Research: The Tenth International Symposium* (2003).

[ED06] E. Eade and T. Drummond. "Scalable Monocular SLAM". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2006, pp. 469–476. DOI: 10.1109/CVPR.2006.263.

[ED07] Ethan Eade and Tom Drummond. "Monocular SLAM as a Graph of Coalesced Observations". In: *Int. Conf. on Computer Vision (ICCV)*. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409098.

[ED08] Ethan D. Eade and Tom W. Drummond. "Unified Loop Closing and Recovery for Real Time Monocular SLAM". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2008, pp. 6.1–6.10. DOI: 10.5244/C.22.6.

[ENT05] C. Estrada, J. Neira, and JD Tardos. "Hierarchical SLAM: Real-time accurate mapping of large environments". In: *IEEE Transactions on Robotics (T-RO)* 21.4 (2005), pp. 588–596. DOI: 10.1109/TRO.2005.844673.

[ESL05] RM Eustice, H. Singh, and JJ Leonard. "Exactly Sparse Delayed-State Filters". In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2005, pp. 2417–2424. DOI: 10.1109/ROBOT.2005.1570475.

[Eve+10] Mark Everingham et al. "The Pascal Visual Object Classes (VOC) Challenge". In: *Int. Journal of Computer Vision (IJCV)* 88.2 (2010), pp. 303–338. DOI: 10.1007/s11263-009-0275-4.

[FB81] Martin A. Fischler and Robert C. Bolles. "RANdom SAmple Consensus: A Paradigm for Model Fitting with Applicatlons to Image Analysis and Automated Cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395. DOI: 10.1145/358669.358692.

[FNL02] John W. Fenwick, Paul M. Newman, and John J. Leonard. "Cooperative concurrent mapping and localization". In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Vol. 2. 2002, pp. 1810–1817. DOI: 10.1109/ROBOT.2002.1014804.

[FP09] Jan Funke and Tobias Pietzsch. "A Framework For Evaluating Visual SLAM". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2009, pp. 69.1–69.11. DOI: 10.5244/C.23.69.

[GG+09a] Óscar G. Grasa et al. "EKF Monocular SLAM 3D Modeling, Measuring and Augmented Reality from Endoscope Image Sequences". In: *5th Workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery. (MICCAI)*. Sept. 2009.

[GG+09b] Óscar G. Grasa et al. "Real-Time 3D Modeling from Endoscope Image Sequences". In: *Workshop on Advanced Sensing and Sensor Integration in Medical Robotics (ICRA)*. May 2009.

[GG+14] Óscar G. Grasa et al. "Visual SLAM for Hand-Held Monocular Endoscope". In: *Transactions on Medical Imaging* 33.1 (2014), pp. 135–146. DOI: 10.1109/TMI.2013.2282997.

[GGa] Óscar G. Grasa. *Augmented Reality Video*. http://webdiis.unizar.es/~oscgg/videos/garcia_etal_miccai09_2.avi.

[GGb] Óscar G. Grasa. *Laparoscopic Measurement Video*. http://webdiis.unizar.es/~oscgg/videos/garcia_etal_miccai09_3.avi.

[GGc] Óscar G. Grasa. *Pattern Measurement Video*. http://webdiis.unizar.es/~oscgg/videos/patternMeasurements.mp4.

[GGd] Óscar G. Grasa. *Photorealistic Reconstruction Video*. http://webdiis.unizar.es/~oscgg/videos/garcia_etal_miccai09_1.avi.

[GGe] Óscar G. Grasa. *Video of Additional SLAM Maneuver in Laparoscopy*. http://webdiis.unizar.es/~oscgg/videos/essr11.mp4.

[GGf] Óscar G. Grasa. *Video of EKF + RLR + 1PR in Laparoscopy*. http://webdiis.unizar.es/~oscgg/videos/garcia_etal_icra11.mp4.

[GGg] Óscar G. Grasa. *Video of Laparoscopic Ventral Hernia Repair Procedure (LVHR)*. http://webdiis.unizar.es/~oscgg/videos/LVHR.mp4.

[GGh] Óscar G. Grasa. *Video of SLAM Validation in Laparoscopy*. http://webdiis.unizar.es/~oscgg/videos/garcia_etal_TMI13.mp4.

[GGi] Óscar G. Grasa. *Video of the Classical Measurement Methods of the Hernia Defect*. http://webdiis.unizar.es/~oscgg/videos/ClassicalMethods.mp4.

[GGCM11] Óscar G. Grasa, Javier Civera, and J. M. M. Montiel. "EKF Monocular SLAM with Relocalization for Laparoscopic Sequences". In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2011, pp. 4816–4821. DOI: 10.1109/ICRA.2011.5980059.

[Gil+11a] I. Gil et al. "Augmented Reality and 3D Measurement for Monocular Laparoscopic Abdominal Wall Hernia Repair". In: *33rd Congress of the European Hernia Society (EHS2011)*. 2011.

[Gil+11b] Ismael Gil et al. "Augmented Reality and 3D Measurement for Monocular Laparoscopic Abdominal Wall Hernia Repair". In: *46th Congress of the European Society for Surgical Research (ESSR11)*. 2011. DOI: 10.1002/bjs.7577.

[Haa+13] Sven Haase et al. "ToF/RGB Sensor Fusion for 3-D Endoscopy". In: *Current Medical Imaging Reviews* 9.2 (2013), pp. 113–119. DOI: 10.2174/1573405611309020006.

[Han+10] Ankur Handa et al. "Scalable Active Matching". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2010, pp. 1546–1553. DOI: 10.1109/CVPR.2010.5539788.

[Hat+06] Martin Hatzinger et al. "Hans Christian Jacobaeus: Inventor of Human Laparoscopy and Thoracoscopy". In: *Journal of Endourology* 20 (11 2006), pp. 848–850. DOI: 10.1089/end.2006.20.848.

[HKM09] S.A. Holmes, G. Klein, and D.W. Murray. "An O($N^2$) Square Root Unscented Kalman Filter for Visual Simultaneous Localization and Mapping". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.7 (2009), pp. 1251–1263. DOI: 10.1109/TPAMI.2008.189.

[HS88] Chris Harris and Mike Stephens. "A Combined Corner and Edge Detector". In: *Proceedings of the 4th Alvey Vision Conference*. 1988, pp. 23.1–23.6. DOI: 10.5244/C.2.23.

[Hu+12] Mingxing Hu et al. "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. DOI: 10.1016/j.media.2010.11.002.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[JU97] S.J. Julier and J.K. Uhlmann. "A new extension of the Kalman filter to nonlinear systems". In: *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*. Vol. 3. 1997. DOI: 10.1117/12.280797.

[KM07] Georg Klein and David Murray. "Parallel Tracking and Mapping for Small AR Workspaces". In: *Int. Symp. on Mixed and Augmented Reality (ISMAR)*. 2007, pp. 225–234. DOI: 10.1109/ISMAR.2007.4538852.

[KM08] Georg Klein and David Murray. "Improving the Agility of Keyframe-Based SLAM". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 5303. Springer. 2008, pp. 802–815. DOI: 10.1007/978-3-540-88688-4_59.

[Kü+09] Rainer Kümmerle et al. "On measuring the accuracy of SLAM algorithms". In: *Autonomous Robots* 27.4 (2009), pp. 387–407. DOI: 10.1007/s10514-009-9155-6.

[LB93] K. A. LeBlanc and W. V. Booth. "Laparoscopic repair of incisional abdominal hernias using expanded polytetrafluoroethylene: preliminary findings". In: *Surgical Laparoscopy & Endoscopy* 3.1 (1993), 39–41.

[LeB+03] K. A. LeBlanc et al. "Laparoscopic incisional and ventral hernioplasty: lessons learned from 200 patients". In: *Hernia* 7.3 (2003), pp. 118–124. ISSN: 1265-4906. DOI: 10.1007/s10029-003-0117-1.

[LF06] Vicent Lepetit and Pascal Fua. "Keypoint Recognition using Randomized Trees". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 28.9 (2006), pp. 1465–1479. DOI: 10.1109/TPAMI.2006.188.

[Low04] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *Int. Journal of Computer Vision (IJCV)* 60.2 (2004), pp. 91–110. DOI: 10.1023/B:VISI.0000029664.99615.94.

[Mau+12] Xavier Maurice et al. “A structured light-based laparoscope with real-time organs’ surface reconstruction for minimally invasive surgery”. In: *IEEE Int. Conf. of the Engineering in Medicine and Biology Society (EMBC)*. 2012, pp. 5769–5772. DOI: 10.1109/EMBC.2012.6347305.

[MBC11] Abed Malti, Adrien Bartoli, and Toby Collins. “Template-Based Conformal Shape-from-Motion from Registered Laparoscopic Images”. In: *Conference in Medical Image Understanding and Analysis (MIUA)*. 2011.

[MBC12] Abed Malti, Adrien Bartoli, and Toby Collins. “Template-Based Conformal Shape-from-Motion-and-Shading for Laparoscopy”. In: *Int. Conf. on Information Processing in Computer-Assisted Interventions(IPCAI)*. Vol. 7330. 2012, pp. 1–10. DOI: 10.1007/978-3-642-30618-1_1.

[MBM01] Edward M. Mikhail, James S. Bethel, and J. Chris McGlone. *Introduction to Modern Photogrammetry*. John Wiley & Sons, 2001.

[MC+05] S. Morales-Conde et al. “Laparoscopic ventral hernia repair without sutures-double crown technique: Our experience after 140 cases with a mean follow-up of 40 months”. In: *INTERNATIONAL SURGERY* 90.3, S (2005), S56–S62.

[MDCM01] Fabien Mourgues, Frédéric Devernay, and Ève Coste-Manière. “3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery”. In: *IEEE/ACM Symp. on Augmented Reality*. 2001, pp. 191–192. DOI: 10.1109/ISAR.2001.970537.

[MH+13] L. Maier-Hein et al. “Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery”. In: *Medical Image Analysis* 17.8 (2013), pp. 974–996. DOI: 10.1016/j.media.2013.04.003.

[MH85] Marjorie Mudge and L. E. Hughes. “Incisional hernia: A 10 year prospective study of incidence and attitudes”. In: *British Journal of Surgery* 72.1 (1985), pp. 70–71. DOI: 10.1002/bjs.1800720127.

[Mir+12] Daniel Mirota et al. “A System for Video-Based Navigation for Endoscopic Endonasal Skull Base Surgery”. In: *IEEE Transactions on Medical Imaging (TMI)* 31.4 (2012), pp. 963–976. DOI: 10.1109/TMI.2011.2176500.

[MNLF08] Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua. "Pose Priors for Simultaneously Solving Alignment and Correspondence". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2008, pp. 405–418. DOI: 10.1007/978-3-540-88688-4_30.

[Mon+02] Michael Montemerlo et al. "FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem". In: *Proceedings of the AAAI National Conference on Artificial Intelligence*. 2002, pp. 593–598.

[Mou+06] Peter Mountney et al. "Simultaneous Stereoscope Localization and Soft-Tissue Mapping for Minimal Invasive Surgery". In: *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 4190. 2006, pp. 347–354. DOI: 10.1007/11866565_43.

[Mou+09] E. Mouragnon et al. "Generic and real-time structure from motion using local bundle adjustment". In: *Image and Vision Computing* 27.8 (2009), pp. 1178–1193. DOI: 10.1016/j.imavis.2008.11.006.

[MY10] Peter Mountney and Guang-Zhong Yang. "Motion Compensated SLAM for Image Guided Surgery". In: *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 6362. 2010, pp. 496–504. DOI: 10.1007/978-3-642-15745-5_61.

[Nic+11] Stéphane Nicolau et al. "Augmented reality in laparoscopic surgical oncology". In: *Surgical Oncology* 20.3 (2011), pp. 189–201. DOI: 10.1016/j.suronc.2011.07.002.

[Nis04] David Nistér. "An efficient solution to the five-point relative pose problem". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 26.6 (2004), pp. 756–770. DOI: 10.1109/TPAMI.2004.17.

[Nis05] David Nistér. "Preemptive RANSAC for live structure and motion estimation". In: *Machine Vision and Applications* 16.5 (2005), pp. 321–329. DOI: 10.1007/s00138-005-0006-y.

[NT01] José Neira and Juan D. Tardós. "Data Association in Stochastic Mapping Using the Joint Compatibility Test". In: *IEEE Transactions on Robotics and Automation* 17.6 (2001), pp. 890–897. DOI: 10.1109/70.976019.

[Oku+11] Asli Okur et al. “MR in OR: First analysis of AR/VR visualization in 100 intra-operative Freehand SPECT acquisitions”. In: *Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Oct. 2011, pp. 211–218. DOI: 10.1109/ISMAR.2011.6092388.

[OM01] D. Ortín and J. M. M. Montiel. “Indoor robot motion based on monocular images”. In: *Robotica* 19.03 (2001), pp. 331–342. DOI: 10.1017/S0263574700003143.

[Ore+11] Sean B. Orenstein et al. “Outcomes of laparoscopic ventral hernia repair with routine defect closure using “shoelacing” technique”. In: *Surgical Endoscopy* 25.5 (2011), pp. 1452–1457. DOI: 10.1007/s00464-010-1413-3.

[Paz+08] Lina M. Paz et al. “Large-Scale 6-DOF SLAM With Stereo-in-Hand”. In: *IEEE Transactions on Robotics (T-RO)* 24.5 (2008), pp. 946–957. DOI: 10.1109/TRO.2008.2004637.

[PT08] P. Piniés and J.D. Tardós. “Large Scale SLAM Building Conditionally Independent Local Maps: Application to Monocular Vision”. In: *IEEE Transactions on Robotics (T-RO)* 24.5 (2008), pp. 1094–1106. DOI: 10.1109/TRO.2008.2004636.

[PTN08] Lina M. Paz, Juan D. Tardós, and José Neira. “Divide and Conquer: EKF SLAM in *O(n)*”. In: *IEEE Transactions on Robotics* 24.5 (2008), pp. 1107–1120. DOI: 10.1109/TRO.2008.2004639.

[RAW11] RAWSEEDS. RAWSEEDS *public datasets web page*. http://www.rawseeds.org/. 2011.

[RD05] Edward Rosten and Tom Drummond. “Fusing Points and Lines for High Performance Tracking”. In: *Int. Conf. on Computer Vision (ICCV)*. Vol. 2. Oct. 2005, pp. 1508–1515. DOI: 10.1109/ICCV.2005.104.

[RFP08] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. “A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 5303. Lecture Notes in Computer Science. 2008, pp. 500–513. DOI: 10.1007/978-3-540-88688-4_37.

[Sau+11] Stefan Sauerland et al. “Laparoscopic versus open surgical techniques for ventral or incisional hernia repair”. In: *Cochrane Database of Systematic Reviews*. 3. John Wiley & Sons, Ltd, 2011. DOI: 10.1002/14651858.CD007781.pub2.

[Sch+12] Christoph Schmalz et al. “An endoscopic 3D scanner based on structured light”. In: *Medical Image Analysis* 16.5 (2012), pp. 1063–1072. DOI: 10.1016/j.media.2012.04.001.

[SDY05] Danail Stoyanov, Ara Darzi, and Guang-Zhong Yang. “A Practical Approach Towards Accurate Dense 3D Depth Recovery for Robotic Laparoscopic Surgery”. In: *Computer Aided Surgery* 10.4 (2005), pp. 199–208. DOI: 10.3109/10929080500230379.

[SFS09] Davide Scaramuzza, Friedrich Fraundorfer, and Roland Siegwart. “Real-Time Monocular Visual Odometry for On-Road Vehicles with 1-Point RANSAC”. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2009, pp. 4293–4299. DOI: 10.1109/ROBOT.2009.5152255.

[Smi+09] Mike Smith et al. “The New College Vision and Laser Data Set”. In: *The International Journal of Robotics Research* 28.5 (2009), pp. 595–599. DOI: 10.1177/0278364909103911.

[SR93] T. A. Santora and J. J. Roslyn. “Incisional Hernia”. In: *Surgical Clinics Of North America* 73.3 (1993), 557–570.

[SS02] Daniel Scharstein and Richard Szeliski. “A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms”. In: *Int. Journal of Computer Vision (IJCV)* 47.1-3 (2002), pp. 7–42. DOI: 10.1023/A:1014573219977.

[SSC87] R. Smith, M. Self, and P. Cheeseman. “A stochastic map for uncertain spatial relationships”. In: *4th International Symposium on Robotics Research*. 1987.

[ST94] J. Shi and C. Tomasi. “Good Features to Track”. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 1994, pp. 593–600.

[Ste+10] Olmi Stefano et al. “Laparoscopic Incisional Hernia Repair With Fibrin Glue in Select Patients”. In: *Journal of the Society of Laparoendoscopic Surgeons (JSLS)* 14.2 (2010), pp. 240–245. DOI: 10.4293/108680810X12785289144359.

[Tar81] J. C. Tarasconi. “Endoscopic Salpingectomy”. In: *Journal of Reproductive Medicine* 26.10 (1981), 541–545.

[TBF05] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[Thr+04] Sebastian Thrun et al. "Simultaneous localization and mapping with sparse extended information filters". In: *The International Journal of Robotics Research* 23.7-8 (2004), pp. 693–716. DOI: 10.1177/0278364904045479.

[TM93] Philip H. S. Torr and David W. Murray. "Outlier detection and motion segmentation". In: *Proc. SPIE, Sensor Fusion VI* 2059 (1993), pp. 432–443. DOI: 10.1117/12.150246.

[Tot+11] Johannes Totz et al. "Dense Surface Reconstruction for Enhanced Navigation in MIS". In: *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. 6891. 2011, pp. 89–96. DOI: 10.1007/978-3-642-23623-5_12.

[Tri+00] Bill Triggs et al. "Bundle Adjustment — A Modern Synthesis". In: *Vision Algorithms: Theory and Practice*. Vol. 1883. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2000, pp. 298–372. DOI: 10.1007/3-540-44480-7_21.

[TZ00] P.H.S. Torr and A. Zisserman. "MLESAC: A new robust estimator with application to estimating image geometry". In: *Computer Vision and Image Understanding* 78.1 (2000), pp. 138–156. DOI: 10.1006/cviu.1999.0832.

[Ved+05] Andrea Vedaldi et al. "KALMANSAC: Robust filtering by consensus". In: *Int. Conf. on Computer Vision (ICCV)*. Vol. 1. 2005, pp. 633–640. DOI: 10.1109/ICCV.2005.130.

[Wan+08] Hanzi Wang et al. "Robust Motion Estimation and Structure Recovery from Endoscopic Image Sequences with an Adaptive Scale Kernel Consensus Estimator". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2008, pp. 1–7. DOI: 10.1109/CVPR.2008.4587687.

[WKR07] Brian Williams, Georg Klein, and Ian Reid. "Real-Time SLAM Relocalisation". In: *Int. Conf. on Computer Vision (ICCV)*. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409115.

[WSC07] Chia-Hsiang Wu, Yung-Nien Sun, and Chien-Chen Chang. "Three-Dimensional Modeling From Endoscopic Video Using Geometric Constraints Via Feature Positioning". In: *IEEE Trans. on Biomedical Engineering* 54.7 (2007), pp. 1199–1211. DOI: 10.1109/TBME.2006.889767.

[Zha00] Zhengyou Zhang. “A Flexible New Technique for Camera Calibration”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)* 22.11 (2000), pp. 1330–1334. DOI: 10.1109/34.888718.

[LeB07] LeBlanc, K. A. “Laparoscopic incisional hernia repair: are transfascial sutures necessary? A review of the literature”. In: *Surgical Endoscopy* 21.4 (2007), pp. 508–513. DOI: 10.1007/s00464-006-9032-8.