SIMULTANEOUS LOCALISATION AND MAPPING FOR MINIMALLY INVASIVE SURGERY

Peter Edward Mountney

Royal Society/Wolfson Foundation Medical Image Computing Laboratory
Department of Computing
Imperial College of Science, Technology and Medicine
University of London

2010

This thesis is submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy. Except for where indicated, it presents entirely my own work and describes the results of my own research.

Abstract

In recent years, Minimally Invasive Surgery (MIS) has transformed the general practice of surgery. The benefits are well documented and include reduced trauma, hospitalisation and comorbidity leading to faster recovery. Despite the benefits, current instrument design and visualisation make MIS challenging. The clinical benefits of Image Guided Intervention (IGI) are well established for procedures such as neurosurgery, where tissue motion is manageable. IGI provides visualisation below the tissue surface allowing the surgeon to avoid critical structures and identify target anatomy. In minimally invasive cardiac, gastrointestinal, or abdominal surgery, significant tissue deformation prohibits accurate registration of pre- and intra-operative data. In this thesis, computer vision and machine learning techniques are explored to estimate 3D tissue deformation for improved intra-operative navigation and visualisation.

The main focus of this thesis is concerned with modelling 3D tissue deformation from a mobile intra-operative device. Two methods are proposed to improve region tracking using machine learning techniques. The first is based on tracking-by-detection. A set of region descriptors is systematically selected, which are robust to deformation. This set is fused in a probabilistic framework to combine multiple cues and boost tracking. In the second method, a context specific technique is developed. It is capable of learning the information that best distinguishes a region from its surroundings. The information is adaptively updated online to learn a representation that is robust to deformation.

3D tissue models are built sequentially from a moving imaging device using Simultaneous Localisation And Mapping (SLAM). To this end, an optical biopsy mapping system based on SLAM is proposed. The system registers multi-modal, intra-operative images to a common coordinate space. The resulting Augmented Reality (AR) visualisation aids biopsy site retargeting and navigation. A second SLAM based system is proposed for dynamic view expansion. By using the localised camera position, a photorealistic tissue model is augmented onto the laparoscopic video. This expands the camera's field-of-view to aid navigation and reduce disorientation. Significantly, in this thesis, a re-formulation of the static SLAM problem is proposed. This is called Motion Compensated SLAM (MC-SLAM) which is capable of accurate localisation and dynamic mapping in periodically deforming environments. The work is validated using simulated, phantom, *ex vivo* and *in vivo* data. Finally, the future research directions and potential improvements to the techniques presented in this thesis are outlined.

Acknowledgements

I would like to thank my advisor, Professor Guang-Zhong Yang, for the opportunity to pursue a PhD in Medical Imaging. He has pushed me to achieve more than I ever thought possible and given me confidence in my research abilities. His vision, clinical and technical knowledge have shaped my work and his enthusiasm for research will remain with me.

I would also like to thank Andrew Davison. His work has been a constant influence throughout my PhD, and I consider myself fortunate to have had the opportunity to collaborate with him. I am extremely grateful for his practical advice, help, insight and patience, especially at the outset of my studies when I was first finding my way.

Throughout my PhD research, I have had the pleasure of working with very talented people who have offered me their help and advice on countless occasions. I would like to thank Dan S, Matina, Selen, Mirna, Dave, Marco, George, Dan E, Fani and Adrian. I have been lucky enough to work with some excellent clinicians including, Jim C, Dan L and Mike.

I have also been lucky enough to share my PhD experience with the friends I have made along the way. They made dealing with the ups and downs of research easier, and for that, I am incredibly grateful to: Andy D, Andy H, Doug, Salman, James, Rachel, Valentina, Julien, Alex, Johannes, Chris, Toby, Ka Wai, Vincent, Jim P, Neil and to everyone who has attended the Thursday meeting. I want to thank all my friends from London, Bristol and Reigate for always asking when I will finish and get a real job. I owe a special thanks to Cath for her patients and support.

I would not have pursued a PhD had it not been for the support of my family. My brother Andy has always been there to support me and given me the self belief to complete my PhD. I want to thank my mum Ricia for encouraging me to pursue what makes me happy and Andrew for his constant support. I would also like to thank my Dad and Christine for their support and advice.

To my family

Contents

List of Figures

List of Tables

List of Acronyms

Area Under Curve ..... (AUC)
Augmented Reality ..... (AR)
Bayesian Framework for Feature Selection ..... (BFFS)
Bidirectional Reflectance Distribution Function ..... (BRDF)
Blur Robust ..... (BR)
Blur Robust Colour Based Object Recognition ..... (BR-CBOR)
Blur Robust Colour Constant Colour Indexing ..... (BR-CCCI)
Charge-Coupled Device ..... (CCD)
Colour Based Object Recognition ..... (CBOR)
Colour Constant Colour Indexing ..... (CCCI)
Colour Cross Correlation ..... (CCC)
Colour Differential Invariants ..... (CDI)
Colour Gradient Location-Orientation Histogram ..... (CGLOH)
Colour Image Moments ..... (CMOM)
Colour Scale Invariant Feature Transform ..... (CSIFT)
Colour Speeded Up Robust Features ..... (CSURF)
Colour Spin Images ..... (CSpin)
Colour Steerable Filter ..... (CSF)
Computer Assisted Surgery ..... (CAS)
Computer-Integrated Surgery ..... (CIS)
Computed Tomography ..... (CT)
Coronary Artery Bypass Graft ..... (CABG)
Cross Correlation ..... (CC)
Degrees-of-Freedom ..... (DOF)
Difference Of Gaussian ..... (DOG)
Differential Invariants ..... (DI)
Directed Acyclic Graph ..... (DAG)
Ear, Nose and Throat ..... (ENT)
Efficient Second-order Minimisation ..... (ESM)
Extend Kalman Filter ..... (EKF)

Field-of-View ..... (FOV)
Finite Element Method ..... (FEM)
General-Purpose Computing on Graphics Processing Units ..... (GPGPU)
Geodesic Intensity Histograms ..... (GIH)
Global Positioning System ..... (GPS)
Graded Index ..... (GRIN)
Gradient Location-Orientation Histogram ..... (GLOH)
Graphics Processing Unit ..... (GPU)
Image Guided Intervention ..... (IGI)
Image Guided Surgery ..... (IGS)
Image Moments ..... (MOM)
Independent Component Analysis ..... (ICA)
Inertia Measurement Unit ..... (IMU)
Infrared Light Emitting Diodes ..... (IRED)
Intra-operative Computed Tomography ..... (iCT)
Intra-operative Magnetic Resonance Imaging ..... (iMRI)
Iterative Closest Point ..... (ICP)
Left Internal Mammary Artery ..... (LIMA)
Left Internal Thoracic Artery ..... (LITA)
Levenberg-Marquardt ..... (LM)
Light-Emitting Diode ..... (LED)
Lucas Kanade ..... (LK)
Magnetic Resonance Imaging ..... (MRI)
Maximally Stable Extremal Regions ..... (MSER)
Minimally Invasive Direct Coronary Artery Bypass ..... (MIDCAB)
Minimally Invasive Surgery ..... (MIS)
Motion Compensated Simultaneous Localisation And Mapping ..... (MC-SLAM)
Mutual Information ..... (MI)
Natural Orifice Transluminal Endoscopic Surgery ..... (NOTES)
Naive Bayesian Network ..... (NBN)
Normalised Cross Correlation ..... (NCC)
One-dimensional ..... (1D)
Optical Coherence Tomography ..... (OCT)
Positron Emission Tomography ..... (PET)
Principal Component Analysis ..... (PCA)

Radio Frequency ..... (RF)
Radio Frequency Ablation ..... (RFA)
Random Sample Consensus ..... (RANSAC)
Receiver Operating Characteristic ..... (ROC)
Red, Green and Blue ..... (RGB)
Scale Invariant Feature Transform ..... (SIFT)
Shape-From-Shading ..... (SFS)
Simultaneous Localisation And Mapping ..... (SLAM)
Single Photon Emission Tomography ..... (SPET)
Singular Value Decomposition ..... (SVD)
Speeded Up Robust Features ..... (SURF)
Steerable Filter ..... (SF)
Structure-From-Motion ..... (SFM)
Sum of Absolute Differences ..... (SAD)
Sum of Squared Differences ..... (SSD)
Three-dimensional ..... (3D)
Totally Endoscopic Coronary Artery Bypass Graft ..... (TECAB)
Two-dimensional ..... (2D)

Chapter 1

Introduction

Over the past two decades, Minimally Invasive Surgery (MIS) has played a major role in reshaping the general practice of surgery. It greatly reduces patient trauma leading to faster recovery time and reduced hospitalisation and risk of comorbidity. Although the benefits of MIS are well documented, the current tools make procedures challenging for surgeons and limit what can be achieved. The elongated tools lack tactile feedback, have limited degrees of freedom of motion, and suffer from the fulcrum effect. Furthermore, the internal organs are visualised using a laparoscopic camera displayed onto a 2D monitor. This results in a loss of direct 3D vision, off-axis visualisation, and a limited view of the surgical site. Current technologies have reached a glass ceiling in the functionality they can provide. The future of MIS depends on ergonomically improved instruments combined with effective visualisation. Increasing the current functionality of MIS will require combining both pre- and intra-operative imaging and sensing data - potentially with the assistance of a surgical robot for enhanced manual dexterity and access.

In recent years, Image Guided Intervention (IGI) has demonstrated its clinical potential of enhanced visualisation by using pre-operative data to guide intra-operative manipulation. Important information such as target anatomies, access routes, and critical structures can be defined using pre-operative data. This information is presented intra-operatively to guide the surgeon, enabling navigation and visualisation beyond the exposed tissue surface during surgery. IGI is currently limited to procedures such as

neurosurgery and orthopaedics, where tissue deformation is manageable. However, the theoretical benefits of IGI are even greater in MIS procedures such as cardiac, abdominal, and gastrointestinal surgeries. In these cases, accurate registration of pre- and intra-operative data is a significant challenge due to the amount of tissue deformation involved. The accurate estimation of the deforming 3D geometry *in situ* is, therefore, a fundamental pre-requisite to enable accurate, robust registration.

During MIS, intra-operative imaging techniques such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasound can be used to estimate the anatomical structure and temporal deformation. However, their practical use is restricted by what is already a very complex operating room setting. Laparoscopes and endoscopes are the standard intra-operative imaging devices used in MIS. It is desirable to use these modalities to estimate tissue deformation, but these devices only capture 2D information. Consequently, it is necessary to develop 3D reconstruction techniques based on computer vision algorithms. This, in itself, is challenging in the presence of tissue deformation and camera motion.

During MIS, the laparoscope or endoscope is controlled by the surgeon and used to navigate internal cavities toward the target anatomy and around the surgical site. Therefore, if these images are to be used to estimate deformation and perform registration, the motion of the camera must first be estimated and removed. Simultaneous estimation of the motion of the camera and the surrounding 3D structure is a well-studied topic in computer vision. Common approaches include Structure-from-Motion and Simultaneous Localisation And Mapping (SLAM). These approaches are mainly concerned with natural scenery with rigid structures. Deforming environments is a challenging research topic, which is further complicated by the requirement of online estimation and motion prediction during MIS.

The purpose of this thesis is to investigate image based techniques for estimating the spatial structure and temporal deformation of soft-tissue, as well as the pose of the intra-operative imaging devices used during MIS. Thus, the central aims and objectives of this thesis are:

assess the current state-of-the-art vision techniques for tracking tissue deformation during MIS;

investigate the use of machine learning and context specific information to enhance tissue tracking performance;

examine the feasibility of using SLAM based on laparoscopic images to simultaneously estimate laparoscope camera motion and 3D tissue morphology;

develop and evaluate a novel vision-based framework for image guided surgery with application to multi-modality intra-operative image registration and dynamic view expansion;

extend the current SLAM framework for MIS surgical scenes involving dynamic tissue deformation.

In **Chapter 2**, the basic concept of IGI is introduced. A brief overview of the key components of IGI is provided and the clinical benefits of IGI for MIS are outlined for cardiac and hepatic surgery. In this chapter, the technical challenges in delivering IGI for MIS are discussed. Following this, a comprehensive literature review of state-of-the-art vision algorithms is provided. A particular focus is placed on techniques that are potentially suitable for tissue deformation tracking and 3D structural recovery in laparoscopic and endoscopic images.

Chapter 3 investigates the application of region tracking for the purpose of estimating deformation of the tissue surface. The use of vision-based algorithms in MIS has attracted significant attention in recent years due to its potential of providing *in situ* 3D tissue deformation recovery for intra-operative surgical guidance and robotic navigation. However, a direct application of these techniques to MIS has revealed many problems, largely due to free-form tissue deformation and varying visual appearances of surgical scenes. This chapter evaluates the current state-of-the-art region descriptors in computer vision and outlines their respective performance issues when used for deformation tracking. A probabilistic framework for selecting the most discriminative descriptors is presented and a Bayesian fusion method is used to boost the accuracy and temporal persistence of tracked features. The performance of the proposed method is evaluated

using both simulated data with known ground truth and *in vivo* video sequences recorded from robotic assisted MIS procedures.

In **Chapter 4**, a context specific approach to region tracking is presented. The method learns region representations online without making assumptions regarding the type of image transformations and visual characteristics involved. These representations are updated continuously as the tracking progresses. In this chapter, the strength of the algorithm is validated with respect to drift, deformation, surgical smoke, occlusion and changes in scale and orientation. Decoupling and modelling cardiac and respiratory motion during robotic assisted surgery demonstrates the practical value of the method.

Chapter 5 describes a technique for building a global 3D map of the scene whilst recovering the camera motion using SLAM. A sequential vision-only approach is adopted which models 6 DOF camera movement. Image artefacts resulting from fibreoptics are removed with pre-processing. The method has been applied to *in vivo* MIS video sequences and validated using a simulated data-set and phantom data-set with CT known ground truth. The results indicate the strength of the proposed algorithm under complex reflectance properties of the scene and its potential for integration with existing MIS hardware.

Chapter 6 provides two, practical applications of the proposed SLAM framework. The first pertains to optical biopsy mapping involving multi-modality image registration and mapping to a single coordinate space. This facilitates intra-operative navigation and visualisation. A micro-confocal imaging probe is used to obtain point-based optical biopsy information. The probe is tracked in the image space to infer the position of the biopsy site, which is incorporated into the statistical framework of SLAM. The second application investigated in **Chapter 6** is dynamic view expansion during MIS. The SLAM framework is proposed to temporally register intra-operative images and to create a 3D textured model of the soft-tissue. The textured model is augmented to the current intra-operative image to extend the effective camera field-of-view for improving spatial awareness during navigation thus reducing disorientation. Methods to improve visual fidelity with texture selection and blending are proposed.

The effect of tissue deformation on the static SLAM framework is the focus of **Chapter 7**. In this chapter, a new formulation of the SLAM framework, with capabilities in

dynamic environments is proposed. The method relies on a high level model of periodic tissue deformation and explicitly incorporates this knowledge into the framework of SLAM. Detailed validation is performed in this chapter and the method is used to visualise pre-operative data in realistic MIS settings for IGI.

Finally, **Chapter 8** concludes the thesis by outlining possible future research directions and challenges for *in situ* 3D structural recovery. A detailed discussion concerning the relative merit and potential drawbacks of the techniques developed in this thesis are discussed.

The work presented in this thesis has resulted in the following publications in peer-reviewed international journals and conference proceedings:

Peter Mountney, Danail Stoyanov, Andrew J. Davison, Guang-Zhong Yang. "Simultaneous Stereoscope Localization and Soft-Tissue Mapping for Minimal Invasive Surgery". In proc *MICCAI* (1) 2006: pp. 347-354

Peter Mountney, Benny P. L. Lo, Surapa Thiemjarus, Danail Stoyanov, Guang-Zhong Yang. "A Probabilistic Framework for Tracking Deformable Soft-tissue in Minimally Invasive Surgery". In proc. *MICCAI* (2) 2007: pp. 34-41

Peter Mountney and Guang-Zhong Yang. "Soft-tissue Tracking for Minimally Invasive Surgery: Learning Local Deformation Online". In proc *MICCAI* (2) 2008: pp. 364-372.

Peter Mountney and Guang-Zhong Yang. "Dynamic View Expansion for Minimally Invasive Surgery using Simultaneous Localization And Mapping". In Proc *EMBC* 2009: pp. 1184-1187

Peter Mountney, Stamatia Giannarou, Daniel Elson and Guang-Zhong Yang. "Optical Biopsy Mapping for Minimally Invasive Cancer Screening". In proc *MICCAI* (1) 2009: pp. 483–490

David Noonan, Peter Mountney, Daniel Elson, Ara Darzi and Guang-Zhong Yang. "A Stereoscopic Fibroscope for Camera Motion and 3D Depth Recovery During Minimally Invasive Surgery". In proc *ICRA* 2009: pp. 4463-4468

Peter Mountney and Guang-Zhong Yang. “Motion Compensated SLAM for Image Guided Surgery”. In proc *MICCAI* (2) 2010: pp. 496–504

Peter Mountney, Danail Stoyanov and Guang-Zhong Yang. “Recovering Tissue Deformation and Laparoscope Motion for Minimally Invasive Surgery” IEE *Signal Processing Magazine*. 2010 June. Volume: 27. Issue 4. pp. 14-24

Peter Mountney and Guang-Zhong Yang. “Context Specific Descriptors for Tracking Deforming Tissue”. To appear in the International Journal of Medical Image Analysis

Mikael H Sodergren, Felipe Orihuela-Espina, Peter Mountney, James Clark, Julian Teare, Ara Darzi, Guang-Zhong Yang. “Orientation strategies in Natural Orifice Translumenal Endoscopic Surgery”. To appear in the Annals of Surgery

The original technical contribution of the thesis includes:

A boosted tracking-by-detection framework for recovering tissue deformation using systematic image descriptor evaluation, selection, and fusion;

An algorithm for learning contextually specific information to improve tissue tracking online using unlabeled data;

A SLAM system to simultaneously estimate laparoscope motion and 3D tissue structure using stereo cameras and robust region matching;

Optical Biopsy Mapping; A method for registering multi-modality images to a common coordinate system for Augmented Reality enhanced navigation;

Dynamic view expansion; Intra-operative image enhancement using photorealistic models generated via SLAM;

A novel Motion Compensated SLAM (MC-SLAM) algorithm for laparoscopic camera localisation and dynamic mapping in a periodically deforming environment.

Chapter 2

Image Guided Intervention and Minimally Invasive Surgery

Surgical procedure has changed dramatically during the past two decades due to technological innovations. Advances in medical imaging, computing, fibre optics, and robotics have created radical, new approaches to surgery with improved consistency and patient outcome. Two of the most significant advances are Minimally Invasive Surgery (MIS) and Image Guided Intervention (IGI). In MIS, the use of small incisions to gain access to internal organs has reduced patient trauma and recovery time. It is now a common practice for many procedures in arthroscopic, abdominal and thoracic surgeries.

The role of IGI is to use imaging and visualisation to guide the surgeon during operation. This generally involves the visualisation of pre- or intra-operative data by augmenting the normal surgical view to reveal structures below the tissue surface with *see-through* vision. The method has been adopted in many procedures for neurosurgery, orthopaedics and ENT (Ear, Nose and Throat) surgery. IGI has proven to be a valuable tool for localising critical structures and assessing anatomical pathways to reach target anatomy. However, its application is currently limited to rigid anatomy: image guidance for surgery involving large tissue deformation remains a significant challenge. IGI combined with MIS, promises to be a powerful tool that can further increase the functional capacities of MIS. In this chapter, an introduction to IGI and its application to MIS will

be provided. The current state-of-the-art techniques are reviewed and key technical challenges and clinical requirements highlighted.

2.1 Image Guided Intervention

For IGI, pre-operative imaging is used to identify the target anatomy and critical structures of a patient prior to surgical intervention. The model is then registered or aligned to the patient's anatomy at the beginning of the procedure and information from the model is made available to the surgeon during the intervention. This enables the surgeon to create a detailed surgical plan, identify appropriate incision points, and determine the optimal pathways for reaching the target area whilst avoiding critical structures.

Although the use of IGI for surgery is a recent practice, the basic idea can be dated back over 100 hundred years. Medical images were first used for surgery in 1895 [1] when X-ray images were used to remove a needle embedded in a patient's hand. However, it was not until 1908 that Horsley and Clark [2] introduced the stereotaxic frame. This is considered the earliest example of an IGI system using external reference landmarks to define co-aligned anatomical and instrument manipulation space. Progression was slow during the following years until the introduction of Computed Tomography (CT) in 1973 [3]. This developed 3D patient-specific data with resolution adequate for surgical guidance. CT was advanced during the early 1980's by the introduction of the personal workstation with high-end graphics, thus allowing IGI to be used in common operating theatres. These technological advances represent the advent of modern IGI.

Four research groups are attributed with the simultaneous invention of frameless stereotaxy [4] - Dartmouth [5, 6], Tokyo Police Hospital [7, 8], the Vanderbilt group [9] and Schloendorff's group [10]. These systems all incorporate the basic components of an IGI system; 3D pre-operative data, intra-operative tool localisation and tracking, and anatomical registration. Over the past two decades, these systems have expanded to incorporate a wide variety of sensing and imaging techniques. A schematic overview of IGI is provided in **Figure 2.1**, which defines the three main components of the system- *i.e.*, pre-operative planning, intra-operative guidance, and post-operative assessment. The pre-operative stage relates to the generation of anatomical models based largely upon

patient-specific imaging data, atlas information, and operation-specific constraints. These models can then be used for surgical planning and guidance during the procedure. The intra-operative stage includes steps during the operation such as the use of intra-operative imaging and sensing data for tracking surgical instruments and monitoring tissue deformation. Post-operatively, the efficacy of the procedure is assessed, yet again, with imaging and, more recently, with pervasive sensing [11].

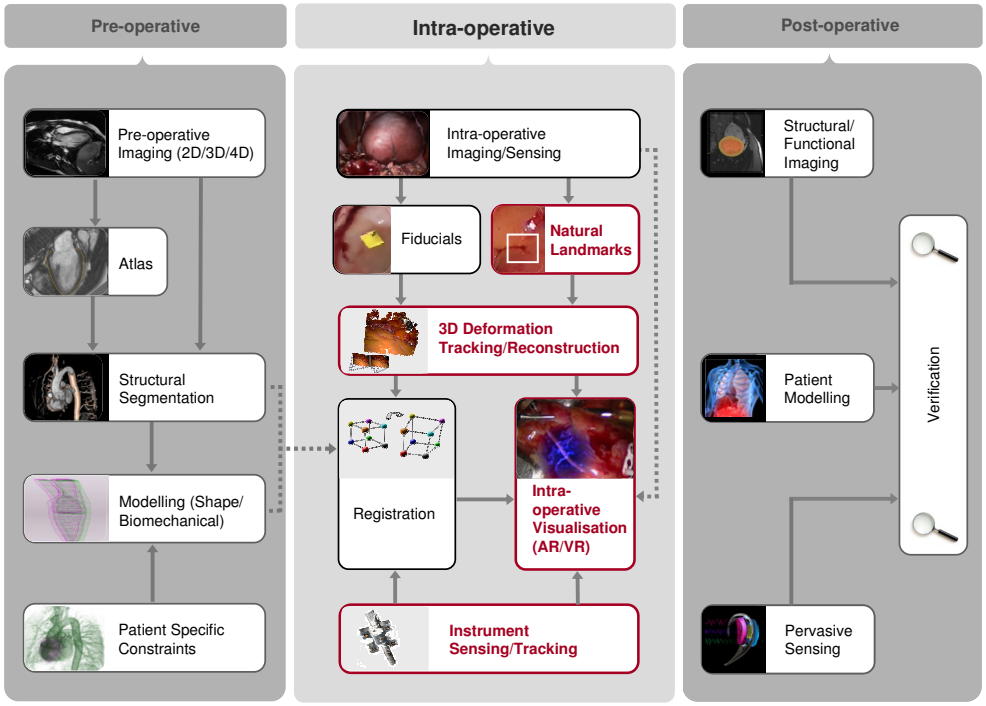


Figure 2.1 Overview of the three main stages of IGI for pre-operative planning, intra-operative guidance and post-operative assessment. The technical contribution of this thesis is mainly concerned with intra-operative guidance involving 3D tissue deformation recovery, instrument sensing/tracking, and intra-operative visualisation, as highlighted in red.

In the following sections, a broad overview of the three main components of IGI is provided. The benefits and hurdles for its successful application to minimally invasive cardiac, abdominal and gastrointestinal surgery are outlined. The work in this thesis will be mainly focused on intra-operative methods- specifically the handling of tissue deformation, camera localisation, and 3D anatomical mapping.

2.1.1 Pre-operative Planning

As previously mentioned, pre-operative data is used in IGI to generate patient-specific models. Information from the model, such as those associated with critical structures and abnormal tissue regions, are used to plan the procedure. The model can be created using multiple imaging modalities, *a priori* atlas information, structural segmentation, and patient specific anatomical constraints.

A variety of imaging modalities have been applied to IGI including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, and Positron Emission Tomography (PET). In practice, the choice of the imaging modality depends on the type of surgery to be performed, but the use of MRI and CT for pre-operative planning is popular because of the imaging resolution and functional details that these imaging modalities can provide. Each imaging modality has its recognised benefits and drawbacks, however, they are complementary and can be used together effectively. By combining multiple modalities, it is possible to incorporate different anatomical and functional aspects of the surgical site for improved guidance and clinical decision-making.

For IGI, structural segmentation is the process of identifying and extracting anatomical details from the per-operative data. The extracted anatomy is used to build a patient-specific model containing relevant information only. Retaining these details facilitates the surgical visualisation process. Segmentation can be performed manually, semi-automatically, or fully automatically. Manual segmentation, whilst accurate, is time consuming and requires extensive user input. Semi-automated approaches reduce the requirement for user interaction by employing iterative methods such as region growing, level sets, or active shape models. Statistical shape models and atlases can be used for automated segmentation. Atlases can contain functional or anatomical data and can be used to register different image modalities together. For example, in [12] statistical atlases are used to aid the placement of electrodes for deep brain stimulation. The use of atlases for IGI is a significant field of research: more details can be found in [13].

2.1.2 Intra-operative Guidance

2.1.2.1 Intra-operative Imaging Techniques

Intra-operative imaging typically involves those modalities that provide real-time or near real-time performance and can be used for interactive guidance. This enables the surgeon to visualise changes in anatomy as the procedure progresses, and to update the pre-operative model and the surgical plan appropriately. Important attributes of intra-operative imaging are safety and the practical use in the operating room. The imaging equipment needs to enable clear and unrestricted patient access without interfering with surgical tools. Common intra-operative imaging modalities include Fluoroscopy, MRI, ultrasound, and biophotonics imaging techniques offering cellular and molecular details *in situ*.

Fluoroscopy provides high spatial resolution with good temporal resolution. These characteristics make it a popular choice for tracking tools and catheters in interventional radiology, cardiology and electrophysiology. With continuing effort to reduce X-ray radiation for CT, intra-operative CT (iCT) is becoming a viable tool for IGI. The 3D volumetric data it provides enables increasingly accurate guidance and localisation of critical structures. It has been used in brachytherapy, spine surgery, and tissue biopsy. However, the temporal resolution is still limited (typically 2 Hz), and the segmentation of the 3D data in real-time is a significant technical challenge.

Intra-operative MRI (iMRI) is one of the most promising intra-operative imaging modalities. It is safe to use and can provide high contrast images of the soft-tissue offering both anatomical and functional information of the surgical site. A variety of iMRI configurations exist [14], ranging from low field mobile installations to high field systems. For IGI, the issues that need to be addressed include: 1) specialised MR compatible tools must be developed; 2) access space of the magnet is restricted, currently making it ergonomically and practically difficult for most MIS procedures; 3) the installation and maintenance cost is prohibitive for most hospitals.

Ultrasound is a popular alternative and is a well established imaging modality. It is relatively affordable and is capable of real-time image acquisition, while remaining safe and practical to use. It is also easily integrated into conventional operating rooms. The

disadvantage of ultrasound is that images are noisy and have low contrast [15]. It is also associated with very minor side-effects of enhanced inflammatory response and unwanted heating of soft-tissue.

Recent advances in biophotonics have lead to a number of cellular and molecular imaging modalities that are amenable to intra-operative use. These approaches to optical are capable of tissue characterisation and include confocal laser scanning microscopes, Optical Coherence Tomography (OCT), two photon excited fluorescence, and high magnification endoscopy [16]. These probes are generally introduced into the patient via the instrument channel of the endoscope, and thus can easily be deployed in routine surgical environments. Currently, the disadvantage of these techniques is that data provision is for a small, localised region only. Due to tissue deformation, large-scale tissue surface surveillance is difficult.

In most MIS procedures, endoscopic and laparoscopic cameras are used to visualise the operating site. These cameras use a variety of hardware including rod lenses, fibre optics, and Charge-Coupled Device (CCD) tip mounted chips. The choice of camera depends on the procedure. Rigid laparoscopes have a limited range of motion but are easy to navigate. Flexible endoscopes require more skill to manoeuvre and are used when access to target anatomy is difficult. Both high-speed and high-definition cameras can be employed, and these will be discussed in more detail in the following sections along with several imaging modalities that are being developed to measure tissue deformation based on structured illumination and time-of-flight principles.

2.1.2.2 Instrument Localisation

Tracking surgical instruments is an essential part of IGI. This data is used to register the position of the tools inside the patient and in relation to the pre-operative model. This enables the tools to be visualised even if they are not directly visible to the surgeon allowing safe navigation below the tissue surface. Tools can be handheld or robotically controlled and may employ geometric or triangulation [4] localisation methods. In robotics, kinematic modelling can be used to geometrically estimate the position and orientation of tools. Recent robotic technologies have been used to overcome the limitations of traditional hand-held devices and further extend the functional capabilities and manual dexterities of the surgeon. Robotic devices provide the control and

manoeuvrability required for precise, microsurgical tasks by offering motion scaling and tremor removal. In addition, robotics provide a platform for motion compensation and impose active constraints, which mark out *no go* zones, such as nerves and blood vessels, based on pre-operative data. In this regard, the use of surgical robots combined with haptic feedback offers a unique opportunity. Early robotic systems include ROBODOC, developed for orthopaedics (Total Hip Replacement and Total Knee Replacement) [17]; the PROBOT, for transurethral resection of the prostate [18, 19]; and the Acrobot [20], incorporating active constraints for knee surgery. Robot assisted surgery has recently moved towards master-slave systems where the surgeon sits at a control station and controls the robotic arms remotely. For example, the da Vinci™ system is a generic MIS soft-tissue robot capable of performing a range of procedures including urological, paediatric, gynaecological, and cardiothoracic surgery. It provides intuitive controls, high precision, tremor elimination, motion scaling, and up to seven degrees of freedom-of-movement for instruments.

Triangulation techniques can be applied to both handheld [21] and robotic tools [22] for instrument tracking and localisation These approaches use transmitters and receivers to localise the tool where either the transmitter or receiver is fixed. Early triangulation methods relied on sonic systems [23, 24], however, these were often affected by changes in temperature and humidity. Optical tracking methods are more reliable [21, 22, 25] due to higher accuracy and frame-rate. These methods use two or more cameras to track optical markers, which can be either active or passive. Active markers contain small Infrared Light Emitting Diodes (IRED). These markers are programmed to strobe at unique frequencies, enabling multiple markers to be tracked simultaneously in the same working volume. The main drawback is that the markers are mostly wired, which can restrict movement. Passive techniques employ reflective strips, patterns or balls. Accuracy and robustness can be improved by employing infrared light sources. For elongated rigid tools inserted into the body, the distal tip can be localised by placing markers at the proximal tip and performing hand-eye calibration [26]. This approach to localisation is, however, only applicable to rigid tools. Electromagnetic tracking may be used to track flexible tools such as endoscopes; however, large metal objects common in the surgical theatre can cause field distortion, and thus introduce errors to localisation accuracy.

2.1.2.3 Registration

One of the fundamental problems in IGI is registration: the process of spatial alignment of separate imaging modalities, or sensors. In atlas generation, for example, registration is used to align multiple patient data-sets. Intra-operatively, registration is used to align pre- and intra-operative data, as well as surgical tools, to a common frame of anatomical reference. Temporal registration must also be performed in addition to spatial registration. Temporal registration updates pre-operative models to reflect changes in anatomy throughout the procedure. This is particularly important for procedures involving tissue deformation.

Rigid registration methods attempt to estimate a 6 degrees-of-freedom transformation (*i.e.*, 3D position and pose) between data streams. This is a well-defined problem with a number of practical solutions. It is, therefore, a common approach for commercially available IGI platforms. In practice, approaches to registration can be categorised according to the dimensionality of the data as either 2D/3D or 3D/3D.

3D/3D registration techniques can be either geometric (feature) or intensity based. Geometry based approaches attempt to extract meaningful information (features) and use this to perform registration. This information can be conveyed as 3D points or surfaces. Point data can be created by fiducial markers or extracted from anatomical structures. Given two sets of point data, a *least squares* solution can be used to estimate the transformation. The surface of anatomical structures can be extracted by segmenting the data using algorithms such as marching cubes [27]. The surfaces represented as set of point data can be registered using the Iterative Closest Point (ICP) algorithm [28].

Intensity based 3D/3D registration uses the original intensity values from the imaging data or its derivatives. The registration problem is posed as an optimisation problem with an objective function and similarity measure. The objective function is defined in terms of the intensity values and the transformation parameters. A variety of similarity measures can be used for registration, which include Sum of Squared Difference (SSD), Normalised Cross Correlation (NCC), or Mutual Information (MI) [29]. Generally, an interpolation scheme is employed during registration since optimisation takes place in a continuous domain wherein the data is intrinsically discrete.

2D/3D registration is closely related to the problem of camera pose estimation in computer vision. The problem is generally formulated as an optimisation problem, which iteratively searches for the optimal transformation. This requires careful initialisation, which can be manually defined or based on external tracking sensors. Similar to 3D/3D registration, current methods can be categorised as geometric or intensity based. Geometric methods such as [30, 31] extend the ICP algorithm to 2D/3D. Intensity methods require an additional step wherein simulated images are created: a direct comparison between 2D and 3D information is not possible using intensity. Projecting the 3D volume data into a virtual perspective camera generates a simulated image, while moving the location of the virtual camera during optimisation yields multiple simulated images. A similarity measure is used to identify the simulated image corresponding to the actual 2D data. This has been applied to broncoscopic images [32] [33], however, intensity based registration requires computationally intensive volume rendering and a large set of simulated images, thus making online registration difficult.

2.1.2.4 Visualisation and Augmented Reality

Visualisation is the front-end of IGI that is presented to the surgeon. It integrates information from pre-operative data, intra-operative imaging, and instrument tracking into a common visualisation. 3D volumetric data may be visualised as slices, surfaces, or by direct volume rendering. Augmented Reality (AR) combines pre-operative and intra-operative data in a simple and intuitive manner. The pre-operative model is added to the physical world as viewed by the intra-operative imaging device. The intra-operative data is augmented with the computer-generated model. When the two data streams are merged, the result simulates see-through vision showing visually co-aligned 3D structures beyond the exposed tissue surface. AR has been applied to a variety of procedures including orthopaedics [34], neurosurgery [35, 36], and interventional radiology [37, 38]. A comprehensive review of medical AR is provided in [39] and AR in MIS will be discussed in further detail.

2.1.3 Post-Operative Assessment

During surgery, particularly MIS, assessing the efficacy of an interventional procedure is important. It provides an opportunity to detect comorbidity and refine patient management to maximise the therapeutic outcome. In practice, the type of post-operative

assessment is specific to the surgical procedure and can range from questionnaires and physical assessment to imaging and pervasive sensing. For example, post-operative assessment can be performed using the same pre-operative imaging modalities. This provides a direct comparison and validation of the efficacy of the operation by monitoring the patient's recovery. For example, CT is used in [40] to identify haematoma in patients after neurosurgical procedures. In many scenarios, post-operative assessment is performed in the patient's home via remote monitoring. In these cases, recently developed, miniaturised wireless sensors can be used to detect surrogate signs of post-surgical complication. This practice has attracted significant interest. In [41, 42], for example, the authors propose the use of ubiquitous sensing to monitor patient recovery. By detecting changes in gait and posture, it is possible to infer early signs of post-surgical infection and complication, as well as assess the general well-being and recovery of the patient.

2.2 Clinical and Technical Considerations of IGI for MIS

2.2.1 Clinical Considerations of IGI

Throughout this thesis, two types of surgery will be used to demonstrate the application of IGI for MIS – cardiac and hepatic surgery. Approximately half of the deaths caused by cardiovascular disease are related to coronary heart disease, which is strongly correlated to dietary habits, physical activity, and tobacco consumption. The illness is caused by a gradual build-up of fatty deposits (atheroma) in the coronary artery resulting in stenosis or narrowing of the artery. Bypass surgery is required for patients who cannot be treated with medication or angioplasty. During this treatment, an additional artery is used to bypass the blocked coronary artery. This enables the oxygen-deprived myocardium to receive an alternative blood supply. The bypass graft is usually harvested from the Left Internal Thoracic Artery (LITA) or the Left Internal Mammary Artery (LIMA).

Open bypass surgery requires the use of a median sternotomy to gain access to the thoracic cavity, thus causing severe trauma that may lead to extended recovery time, scarring and morbidity. This has motivated the development of Minimally-invasive Direct Coronary Artery Bypass (MIDCAB) and Totally Endoscopic Coronary Artery Bypass grafts (TECAB). Minimally invasive approaches to cardiac surgery pose numerous, significant challenges to the surgeon, and image guidance can be used to

visualise and identify target anatomy as illustrated in **Figure 2.2**. IGI for cardiac surgery is far more complex than rigid anatomy due to large-scale tissue deformation. The shape and position of the cardiac surface is affected by both the cardiac and respiratory cycles, both of which must be modelled for accurate registration. The use of a mechanical stabiliser minimises the motion of the heart, but it also alters the shape and motion characteristic of the organ, making the direct use of a pre-operative model difficult.

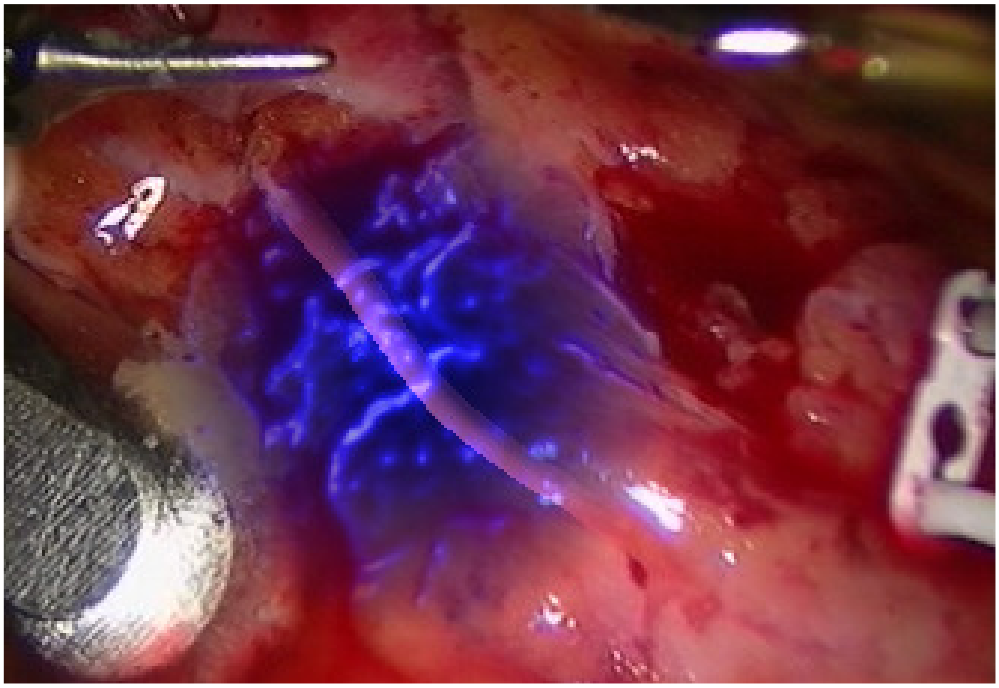


Figure 2.2 Illustration of IGI for cardiac MIS. A laparoscopic image of the cardiac surface augmented with a pre-operative model of a vessel visualised as Augmented Reality using Inverse Realism [43].

For hepatic surgery, IGI can enable visualisation and management of critical structures, such as the blood vessels and bile ducts and the definition of accurate tumour margins. Current practice for resection is to perform pre-operative surgical planning [44]. However, intra-operatively, the procedure may be enhanced by displaying pre-operative data to identify *no go* areas and guide resection margins as illustrated in **Figure 2.3**. When combined with robotic control, this practice enables active constraints that prevent the surgeon from moving tools into dangerous or critical anatomical areas. In practice, managing resection margins is crucial. If the resection is incorrectly performed, abnormal tissue may not be removed or excess healthy tissue may be damaged. This may cause increased recovery time and in some cases, liver failure.

There is strict guidance governing the current eligibility of patients for liver resection - only 5% to 15% of patients qualify [45]. Radio Frequency Ablation (RFA) offers an alternative treatment to patients who do not qualify for resection. Percutaneous RFA can be performed using laparoscopy or laparotomy with CT, MRI or ultrasound image guidance. Currently, errors in the delivery of RFA are largely due to the use of static pre-operative images for guiding tools used in a deforming environment. Such circumstances delineate the importance of the effective handling of tissue deformation during image guided MIS.

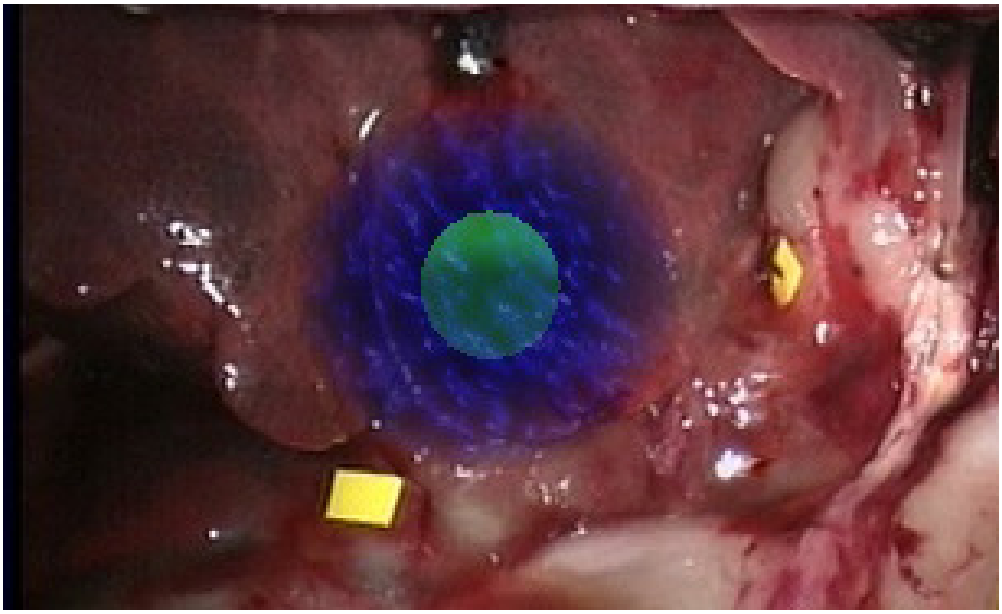


Figure 2.3 Illustration of IGI for hepatic MIS. A laparoscopic image of the liver augmented with the model of a tumour (green) and visualised as Augmented Reality using Inverse Realism [43] (blue).

2.2.2 Key Technical Challenges

2.2.2.1 Causes of Tissue Deformation

Tissue deformation is the most significant problem preventing IGI from being clinically adopted for cardiac, abdominal and gastrointestinal surgery. Tissue deformation creates inaccurate registration between pre-operative and intra-operative data. Deformation during the procedure causes pre-operative data to be misaligned against intra-operative data. During MIS, insufflation of the patient during the procedure can cause significant organ shift. Furthermore, respiration can cause tissue motion and deformation:

contraction of the diaphragm and surrounding muscles cause the air to be drawn into the lungs. Naturally, this deforms the shape of the lung, however, it also affects other organs. In fact, the effect of respiration is evident on all visceral structures although this effect varies between pre-operative data (where breathing is freely controlled by the patient) and intra-operative data (where breathing is controlled by a ventilator). For cardiac procedures, myocardial contraction is a significant obstacle. As discussed earlier, a stabilising clamp is often used to constrain the heart during MIS cardiac surgery . This clamp significantly alters the shape of the heart and introduces another form of tissue deformation. The deformation involved is generally difficult to predict and it is highly non-linear. During MIS, it is, therefore, necessary to consider instrument-tissue interaction. This is an extremely complex problem, which is likely to require biomechanical modelling with detailed, physical properties of the tissue – information that may not be readily available for each subject.

2.2.2.2 *In situ* Tissue Deformation Recovery

As mentioned earlier, the use of tomographic imaging for tissue deformation recovery has clear advantages, but it is difficult to incorporate into the normal surgical workflow. For MIS, using laparoscopic or endoscopic images to register pre-operative data is highly desirable because it requires no additional hardware in the operating theatre, and it provides a natural interface for an AR visualisation of the pre-operative data. Currently, the standard technology used for intra-operative imaging during MIS is monocular laparoscopic/endoscopic cameras. Stereo laparoscopes are becoming more common with the introduction of robotic systems such as the da Vinci. **Figure 2.4** shows six example images captured with laparoscopic/endoscopic cameras. These imaging systems provide 2D video visualisation of the internal organ surfaces. The 3D structure of the tissue can be measured using either stereo or monocular cameras as shown in **Figure 2.5**.

Extensive research has been carried out in computer vision in order to estimate 3D structure and tracking. However, its direct application to MIS involving deforming tissue has revealed numerous difficulties. Many of these challenges are shared with the broader computer vision community such as image blur, noise, artefact, non-linear illumination, occlusion, and changes in viewpoint. In MIS, tissue deformation is non-linear and does

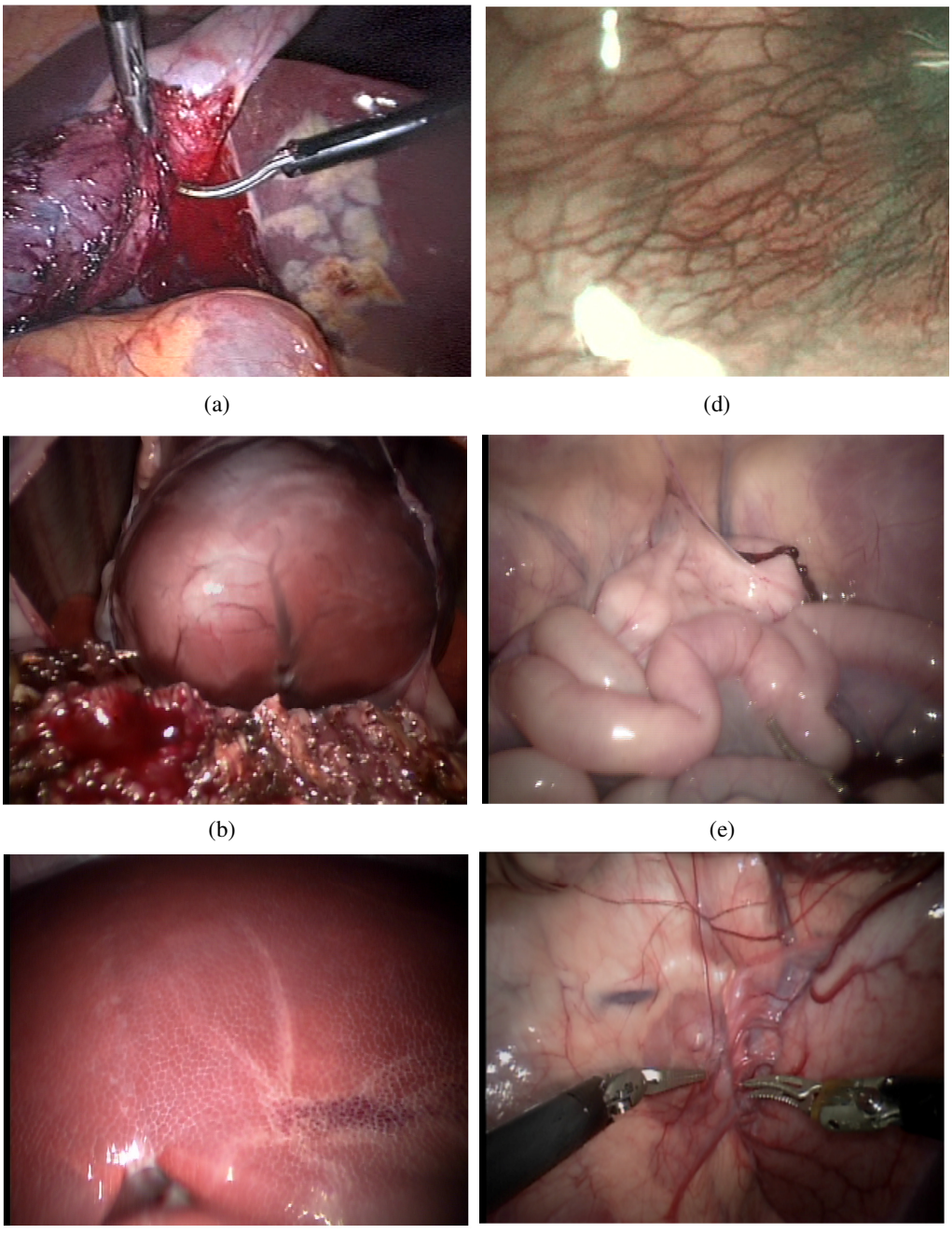


Figure 2.4 Endoscopic and laparoscopic images illustrating specular highlights, tissue-tool occlusion, homogenous surface, repetitive structures, saturation and non-linear illumination. **(a)** Liver and gall bladder, **(b)** heart, **(c)** liver, **(d)** oesophagus (using narrow band imaging), **(e)** bowel viewed from the abdominal cavity and **(f)** bladder viewed from the abdominal cavity.

not conform to affine image transformations or the rigid body assumption commonly used in computer vision.

The problem of tissue deformation tracking and structural estimation is further complicated by the visual appearance of tissue as shown in **Figure 2.4**. Specular highlights are common in MIS and are caused by reflections from a fine layer of mucus formed on the tissue surface. The appearance of tissue can vary greatly from organ to organ. It can appear homogenous, lacking in texture information, or highly repetitive and visually similar. Structures such as veins, arteries and vessels can be observed through the tissue. However, because these structures are below the surface of the tissue, their visual appearance alters depending on illumination and the camera's viewing angle. Furthermore, image artefacts, such as smoke caused by diathermy, can result in full or partial occlusion of the tissue surface. Ultrasound, illustrated in **Figure 2.5 (e)**, offers an affordable modality for visualising structures below the surface but has a poor signal-to-noise ratio. Due to these difficulties, several optical approaches have been proposed which use additional hardware to address the issue of tissue deformation tracking.

Structured light has been used to recover the shape of the surgical site for augmented reality [46]. More recently, the use of projected coded patterns, [47] [48], has also been investigated. This approach requires an additional surgical port but is not reliant on the natural visual appearance of the tissue and, as a result, it is robust where other optical techniques may fail. These methods are mainly limited by the modification of the surgical view caused by the active projection of a pattern onto the tissue surface. Although this may be done in a non-visible spectrum, such a process requires specialised optics. In addition, temporal tracking is difficult because the projected pattern may not correspond to the same region on the tissue over time.

Recently, time-of-flight cameras have been adapted for reconstructing 3D surfaces during MIS [49]. Time-of-flight cameras function on a similar principal to Light Detection And Ranging (LIDAR). The sensors consist of a pixel matrix and an illumination device, which projects incoherent near-infrared light with a modulating frequency. The pixel sensors are synchronised to the same modulation frequency and measure the phase delay between the emitted and measured light. The time required for the light to emit, reflect, and be measured corresponds to the distance it travels from the camera. In **Figure 2.5 (a-c)** a liver phantom is imaged using a MESA-Imaging SR-3000

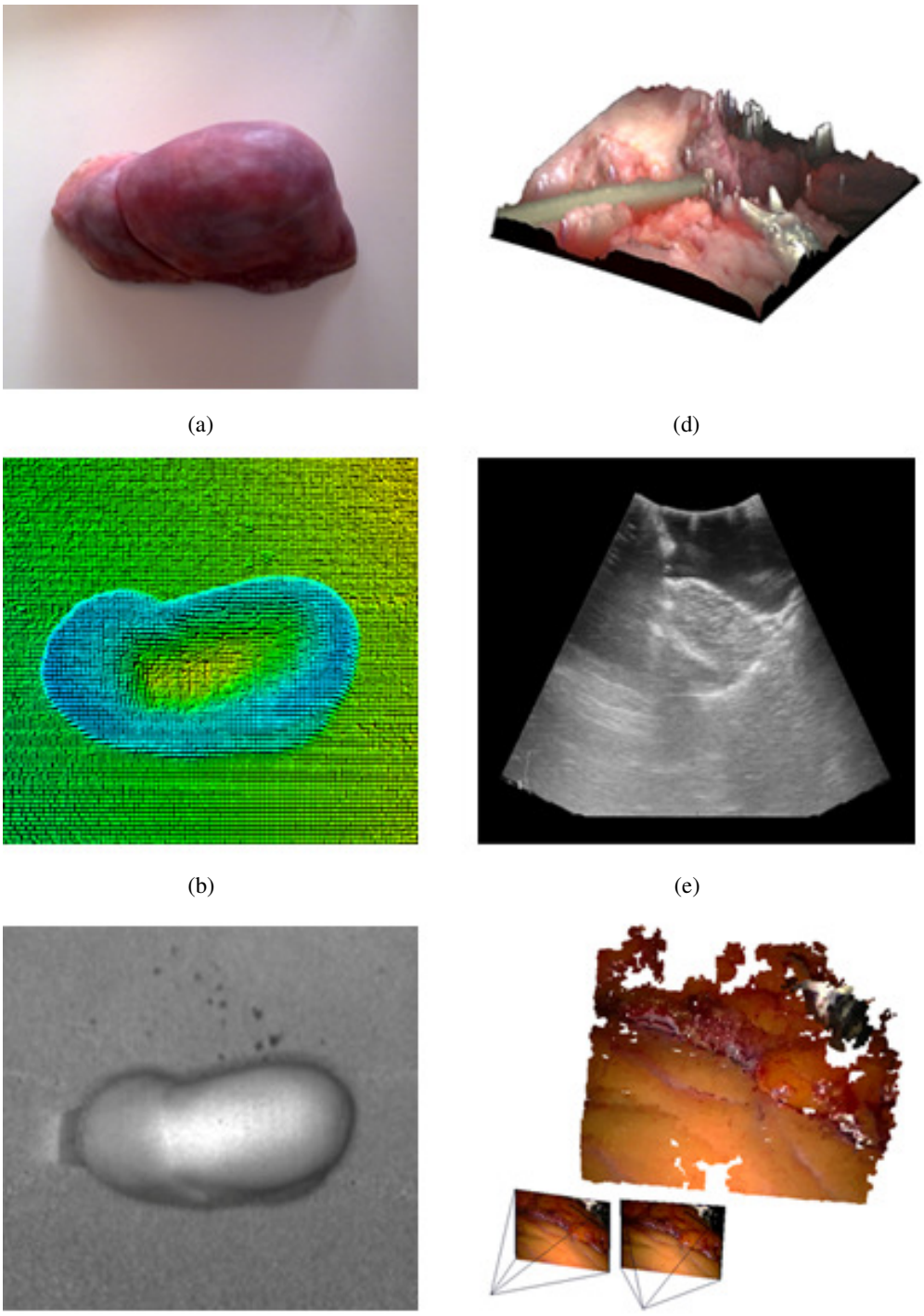


Figure 2.5 Intra-operative recovery of 3D tissue geometry. **(a-c)** Image of a phantom liver captured with **(a)** a laparoscope and **(b-c)** a MESA-Imaging SR-3000 *Time-of-Flight* camera showing the 3D depth map **(b)** and reflectance image **(c)**. **(d)** Shape from Shading reconstruction from monocular endoscopic images [50]. **(e)** A 2D ultrasound image and **(f)** a dense stereo reconstruction from stereo laparoscopic images [51].

time-of-flight camera. These devices are currently in their infancy and early results appear promising. Although existing devices have limited frame rate, resolution, and field-of-view, it is an interesting approach that may harbour viability.

2.2.2.3 Non-Rigid Registration

Rigid registration methods assume that a single global transformation, generally consisting of a rotation and translation, can be used to describe the spatial relationship between the patient's anatomy and pre-operative data. This assumption does not translate to non-linear tissue deformation. Registration with free-form deformation is a well-researched topic in atlas generation but a reliable, fully automated intra-operative method remains difficult to achieve. Approaches to non-rigid registration for IGI include 2D/3D or 3D/3D methods. As previously discussed, they can also be intensity based or feature based. For IGI, registration of patient anatomy to an atlas is frequently used.

Intra-operative, non-rigid image registration for IGI generally requires registration across different imaging modalities making the process more challenging. During hepatic surgery, for example, pre-operative CT can be registered to intra-operative 2D ultrasound data using a point based method [52]. The accuracy can be improved via surface registration [53]. During cardiac surgery, pre-operative CT and intra-operative ultrasound are used to register the heart using fiducial markers [54]. CT is combined with an angiogram [55, 56] to segment the coronary tree, which is manually registered to intra-operative laparoscopic images. A method has been proposed [57], to register a 4D cardiac data-set to stereo laparoscopic images, thus making it feasible for beating heart procedures.

2.2.2.4 Intra-operative Instrument Tracking

In IGI, it is essential to track intra-operative imaging devices, as well as surgical tools for navigation and guidance. The use of infrared markers for rigid instruments has been popular as described earlier. Flexible instruments used in gastrointestinal surgery and more recently in the exploration of Natural Orifice Transluminal Endoscopic Surgery (NOTES) require an alternative approach. Electromagnetic tracking systems can be employed as they do not require the instrument to be rigid or line of sight; however, their accuracy is affected by other electronic devices and ferromagnetic objects such as

surgical tools. The problem of defining the relative position of the laparoscopic or endoscopic camera in MIS, is similar to general camera pose estimation in computer vision. It is feasible to use the imaging device itself to estimate the camera pose. In this case, there are two parallel approaches to this problem: *i.e.*, Structure-from-Motion and Simultaneous Localisation And Mapping (SLAM). The exact technical details of these methods will be discussed later in this chapter.

2.2.2.5 Visualisation and Augmented Reality

In MIS, intra-operative images from a laparoscope/endoscope are usually visualised on a 2D monitor. Physical constraints in the operating theatre dictate the possible location of the monitor, often leading to off-axis visualisation. Loss of 3D vision and a lack of additional depth cues, such as shadows, make 3D navigation difficult. This problem is further complicated by the high magnification factor of the camera, which leads to localised field-of-view and disorientation. The visualisation of intra-operative data can be improved by stereo-vision, similar to that adopted by the da Vinci system, and alignment of visual-motor axes. A second light source can be introduced in MIS [58] to create an shadow barely visible to the human eye which is digitally enhanced to provide an additional depth cue. A tri-axial microelectromechanical system sensor is placed on the tip of an endoscope in [59] to detect its orientation. This information is used to rotationally correct the image during NOTES, thus improving the visual-spatial orientation of the surgeon.

The application of orientation correction has also been addressed using image based techniques [60-62]. Rotational image rectification assumes that the only camera motion is rotational, around the optical axis, or translational, along the optical axis. This is rarely the case in practice. Dynamic view expansion [63] is proposed to enhance intra-operative navigation and reduce visual-spatial disorientation. The method uses optical flow to increase the field-of-view of the camera through image mosaicing. Intra-operative image enhancement techniques such as these can provide valuable orientation information; however combining multi-modality imaging data require more sophisticated visualisation methods.

In MIS, it is intuitive to display pre-operative data to the surgeon in the laparoscopic/endoscopic images. This AR approach to visualisation removes the

requirement for addition visualisation equipment such as head mounted displays and enables smooth integrating with the existing surgical workflow. However, accurate visualisation is not solely dependent on registration and additional errors can be introduced by the camera optics and camera motion [39]. The optical lens of the camera introduces radial distortion into the image which must be catered for. The distortion can be estimated by camera calibration and either the intra-operative or pre-operative images altered accordingly. The camera motion must be tracked to register the position of the imaging device to the pre-operative data. Tracking techniques described earlier can be used in conjunction with a hand-eye calibration process. Both of these steps can introduce additional error into the system. AR has been used for a variety of applications in MIS and detailed reviews are provided in [15, 39, 64].

AR has been applied in [65, 66], where the position of a bronchoscope is visualised to aid navigation. Several systems have been proposed for laparoscopic surgery [38, 67-71]; however these methods do not take into account deformation and employ rigid registration methods. Major challenges of AR include the correct handling of occlusion and depth perception. Objects correctly registered and rendered in metric space, may still appear to be floating. Psycho-visual factors influencing depth perception are addressed in [43, 72]. A virtual mirror is introduced in [72] for interactive 3D visualisation and in [43] a method is proposed based on *pq*-space called Inverse Realism, which is suitable for real-time implementation with high fidelity depth perception.

2.3 Vision Based Techniques for Soft-tissue Deformation Recovery

For soft-tissue deformation recovery, the use of vision-based techniques will mainly be investigated. As previously mentioned, this is desirable as it exploits existing hardware in the surgical theatre. In practice, however, there are a number of challenges involved in vision-based techniques for MIS. An explanation of the physical configuration of the imaging devices in MIS, from camera models to calibration methods, is provided in the following sections. This is the basis for recovering 3D deformation and the relative poses of the laparoscopic cameras.

Figure 2.6 (a) illustrates a pinhole camera model and the projection of a 3D object on to the image plane. Two cameras simultaneously viewing the same 3D scene, as illustrated in **Figure 2.6 (c)**, are related by epipolar geometry. Epipolar geometry describes the relationship between the 3D points in the scene and projections into the cameras' image planes as defined by the cameras' intrinsic and extrinsic parameters. The centre of the left camera is taken to be the origin. In this case, the extrinsic parameters describe the translation and rotation of the right camera relative to the left camera. Epipolar geometry can be illustrated in **Figure 2.6 (c)** where $\mathbf{C}$ and $\mathbf{C}'$ are the camera centres. $\mathbf{M}$ is a point in 3D and m and m' are the projection of $\mathbf{M}$ into the left and right cameras. e and e' are the epipolar lines.

With multiple cameras, it is possible to estimate the 3D geometry of a scene. From **Figure 2.6 (c)**, it is clear that, if the positions of m and m' in the image are known, the position of $\mathbf{M}$ can be estimated. In practice, the back-projected rays may not intersect in 3D due to noise in the image and errors in the estimation of the intrinsic and extrinsic parameters. Therefore, $\mathbf{M}$ is usually taken to be the midpoint of the shortest distance between the rays, which is computed using Singular Value Decomposition (SVD).

The unknown intrinsic parameters of the camera model and the extrinsic parameters of the stereo cameras can be estimated using a calibration process: this is illustrated in **Figure 2.6 (b)**. Requiring only a few minutes, the stationary cameras observe a coplanar calibration grid at several (usually 10-12) orientations. There are a number of well-known calibration algorithms sourced within the computer vision communities [73-75]. Implementations of these methods are available online in several calibration toolboxes [76, 77]. Calibration is usually performed once pre-operatively. The extrinsic parameters of a stereo laparoscope are presumed fixed, and the intrinsic camera parameters are assumed to be constant during a MIS procedure. Otherwise, the parameters can be adaptively estimated [78].

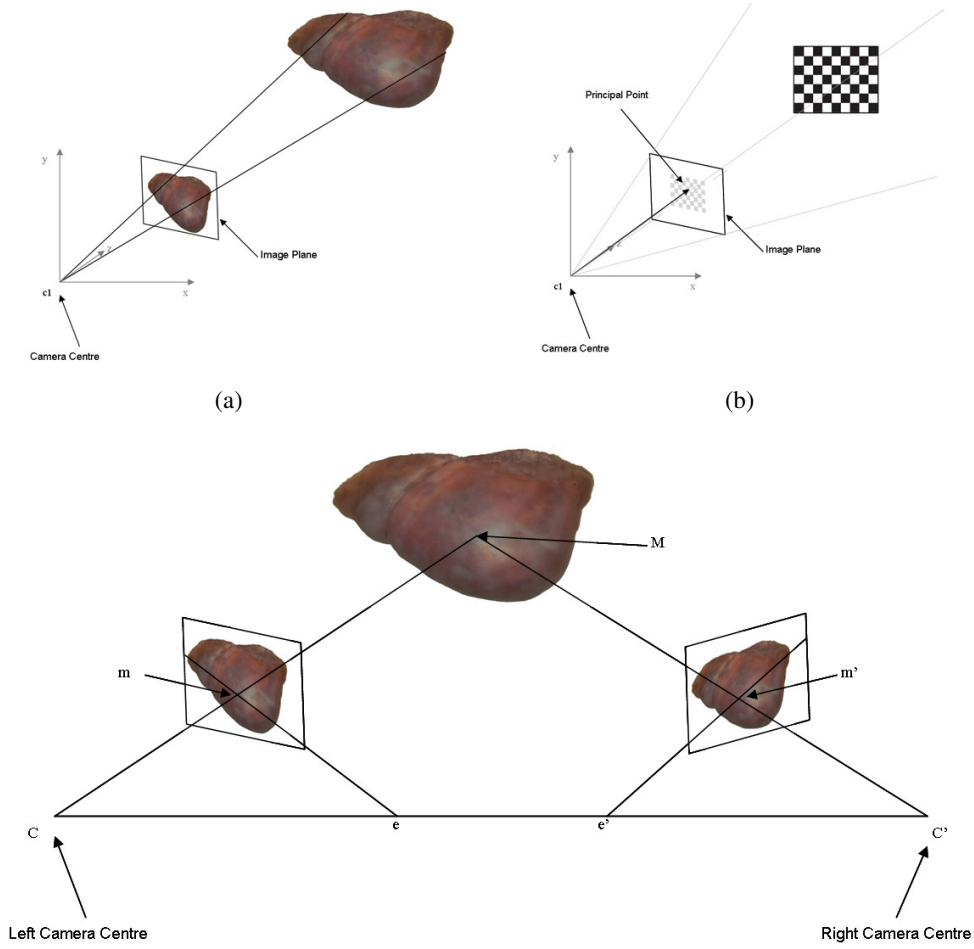


Figure 2.6 The physical configuration of a laparoscopic camera. **(a)** Monocular optical set-up illustrating an object in 3D projected onto the 2D image plane with respect to the camera centre C. **(b)** Calibration of monocular optics showing the principal point and calibration grid. **(c)** Stereo optical set-up with left camera centre C, right camera centre C' and two epipolar lines e and e'. The point M in 3D is projected onto the left and right image plane at locations m and m', respectively.

2.3.1 Recovering Soft-Tissue 3D Structure

For 3D shape recovery, there are many techniques that have been developed by the vision community over the years. In this section, the focus is on methods that have been applied in MIS, these are summarised in **Table 2.1**. The understanding of biological vision systems and the cues humans use to interpret images inspired early work in the field of computer vision. This led to the development of Shape-From-X algorithms,

inspired largely by the work of Marr [79]. In Shape-From-X, a variety of visual cues have been used including shading, texture, and stereo disparity. For MIS, approaches to 3D tissue reconstruction have mainly exploited two visual cues: shading and stereo.

Shape-From-Shading (SFS) is a technique used to estimate surface information by observing the surface under predefined lighting conditions. SFS can be used to estimate surface normals, gradient, slant, tilt, and local depth. The formulation of an image is dependent on the structure and properties of lighting and surface. Therefore, by making assumptions about lighting and surface properties, it is possible to infer information about the structure. The benefits of SFS include performance using a single camera and application to homogenous areas. In its simplest form, SFS assumes an infinitely distant light source, orthographic projection, the Bidirectional Reflectance Distribution Function (BRDF) and Lambertian reflectance. Under these conditions, SFS has been used to accurately estimate surface information.

Traditional SFS assumptions have proven overly constrained for MIS as these images are often affected by perspective and lens distortion. Furthermore, the environment does not strictly obey the BRDF, and specular highlights are not compatible with the Lambertian reflectance assumption. The conjoined light source and camera ,shown in **Figure 2.7** in a number of configurations, also breaks with the assumption of an infinitely distant light source. In order to use SFS for MIS, these constraints must to be relaxed.

Numerous authors to date have addressed these constraints. Lens distortion and perspective projection are addressed in [80] and [81, 82], respectively. In [83], a method is proposed by assuming the light source is positioned at the optical centre of the camera. Although this is not strictly true, it is a reasonable assumption in practice. Building on this work, a modified BRDF is introduced in [84], which monotonically decreases with respect to the cameras viewing angle. Although these methods increase the general applicability of SFS to MIS, the main problem remains that only the relative surface shape is recovered but not the metric 3D representation. To avoid this problem, stereo methods can be used.

Figure 2.7 Examples of the physical optical and lighting configuration of endoscopes and laparoscopes. **(a)** 30° laparoscope, **(b)** 0° laparoscope, **(c)** flexible endoscope, **(d)** stereo laparoscope with two light sources, **(e)** stereo laparoscope with a single light source and **(f)** the da Vinci robotically controlled laparoscope.

Figure 2.7 (d-e) shows the optical configuration of two stereo laparoscopes. Dense stereo algorithms are based on establishing pixels that correspond between stereo image pairs. An approach based on a simple, normalised cross correlation algorithm is used in [85] to estimate the surface of the heart and the authors demonstrate that such a system is capable of surface reconstruction and dealing with discontinuities introduced by surgical tools. Dense techniques such as this, however, require regions on the tissue surface to be distinguishable by the chosen correlation method. In MIS, specular highlights, homogenous tissue, poor illumination, and image noise can introduce significant errors. These errors can be magnified in subsequent depth estimation due to the small baseline of typical stereoscopic laparoscopes which is generally around 5mm and illustrated in **Figure 2.7 (d-e)**. Prior knowledge or strong geometrical assumptions can be introduced to improve surface reconstruction.

Splines have been used to approximate the surface of the heart [86-88] under a smoothness assumption. These approaches identify a number of control points on the tissue surface to provide an initial 3D estimation of the geometry. This is poorly suited to areas of discontinuity caused by surgical instruments. Region-based techniques, on the other hand, extract salient regions of interest and search for correspondences between the stereo images, [89]. The drawback of region-based techniques lies in surface reconstruction, which is likely to be sparse as it relies on salient regions existing in the image. To address this issue, methods based on sets of salient features have been used to recover a sparse 3D depth map and, subsequently, propagate this information to achieve a semi-dense 3D representation, [51]. The major strength of region-based approaches is their ability to track temporal tissue deformation. A plethora of region extraction and matching techniques exist and will be reviewed in detail within the next section.

Table 2.1 Summary of methods used in MIS for 3D reconstruction from image data.

SFS Assumptions	Stereo Approaches	Active Technique
Orthographic [83]	Computational [90, 91]	Fiducial [92-94]
Perspective [80-82]	Surface Priors [84, 86, 95, 96]	One-shot [46, 47]
Illumination [97]	Cue Fusion [50, 82]	Progressive [98-100]

2.3.2 Temporal Tissue Tracking and Modelling

2.3.2.1 Deformable Tissue Tracking

For MIS, the dynamic motion of the soft-tissue can be recovered and modelled by temporally tracking regions of interest through a video sequences. A summary of these methods is provided in **Table 2.2**. Relevant practical challenges include partial or full occlusion, motion blur, image noise, and changes in image scale and orientation. Additional challenges involving tissue deformation during MIS are illustrated in **Figure 2.8**. The surface appearance of tissue, for example, can vary greatly from homogenous to highly texturised. Artificial fiducials have been used to create distinct patterns in the image space: these are easily distinguished from their surroundings, thus simplifying the tracking problem and increasing robustness. Early work in this field [93] used markers containing a Light-Emitting Diode (LED), which is easily detected in the image space. In [92, 94], patterned square markers were attached to the tissue surface. The markers were of a known size enabling the 3D position to be estimated from a monocular camera. In practice, there are three significant drawbacks with the use of fiducial markers. First, the surgeon is required to attach the fiducials to the surface, which can be time-consuming. Secondly, the fiducials can obscure the tissue surface and alter the surgeon's field-of-view. Finally, the density of the surface estimate is limited by the number of fiducials used.

To address some of the problems associated with the use of fiducials, the surface of the tissue can be marked with diathermy [101, 102]. This is an efficient way of creating a distinct mark on the tissue and tracking homogenous regions. However, it is not an advisable technique for patient study as intentional tissue scaring may have long-term, adverse effects. Tracking methods which use naturally occurring features such as vessels, corners, or blobs are preferred. These regions can be selected manually or automatically. Manual selection, [101, 102], ensures the quality of the regions for tracking, however, automatic selection is preferable as it removes the need for user interaction. Automatic region detection (also known as feature detection) involved identification of salient regions in the image, which are distinguished from their surroundings. A large number of region detectors exist and a comprehensive review of their application to MIS is provided in [103].

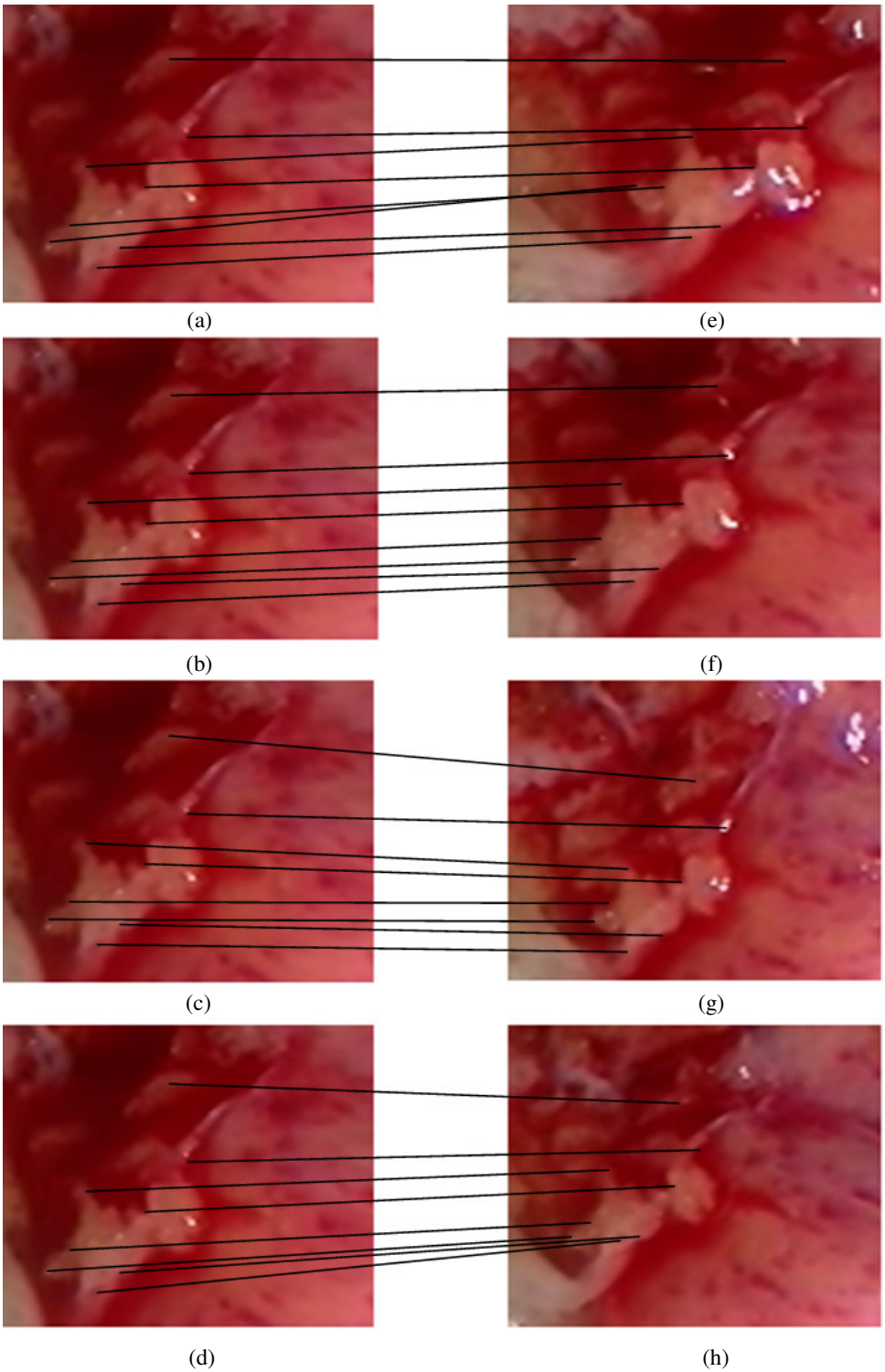


Figure 2.8 An example of non-linear deformation of the cardiac surface. The lines indicate corresponding regions between **(a-d)** the first frame in a video sequence and **(e-h)** images of the cardiac surface at a four temporal positions in the cardiac cycle.

Automatically detecting salient regions in MIS images is challenging due to the variety of imaging conditions and differing visual organ appearance. Classic corner detectors, such as Harris [104] and Shi and Tomasi [105]have been used to detect regions of interest on the epicardial surface [106-108]. These methods use the second moment matrix or the autocorrelation matrix, which describes the gradient distribution in a local neighbourhood. The corner strength of a region is determined by the magnitude of the eigenvalues. A region is detected if the corner strength is above a predefined threshold. Non-maximal suppression can be used to remove poorer corners and manage the number of detected features. However, the surface of tissue may not contain corner-like structures. Maximally Stable Extremal Regions (MSER) [109] have been used to extract blob structures on tissue [89]. The MSER approach is similar to watershed segmentation. MSER detects self-contained blobs in the image in areas where the intensity values inside the blob vary significantly from its surroundings. Corner detection and blob extraction are complementary, and thus can be combined to increase the number of detected regions, [89]. It should be noted that these methods do not explicitly model invariance to changes in scale.

Scale invariant region detectors extract salient regions of varying size in the image. The application of scale invariant trackers on soft-tissue has been recently studied in [103]. Scale invariant detectors, such as the Laplacian of Gaussian [110], Difference Of Gaussian (DOG) [111], and Fast Hessian [112], automatically select the scale allowing each region its own characteristic size. These methods vary in their implementation but commonly detect regions at several scales simultaneously, and the scale level is selected according to a maximum in the scale space. The scale space is either a set of images with different levels of resolution or a size-varying filter. Non-maximal suppression can be used to remove regions and prevent multiple detection of the same point at different scales. One advantage of scale invariant approaches is the associated size of detected regions. This enables the whole region to be used for tracking whereas a fixed size detector may only detect part of the region. The main benefit of this approach, however, is the facilitation of tracking-by-detection, which will be discussed in due course. Once a region has been detected, information from the image must be selected to represent the region.

Tracking algorithms can generally be categorised as recursive or tracking-by-detection [113] approaches. Recursive algorithms have been employed for over twenty years,

[114]. However, recent advances in computational power have enabled real-time tracking-by-detection, which has led to further research in this area. The following sections present a comparison of these approaches and reviews of how they have been used in MIS.

Recursive methods, such as Lucas Kanade (LK) [115], and mean-shift search locally for the best match that minimises a measurement function. The LK algorithm and its extensions are based on optical flow and generally work in image space. This approach aims to register a template of the region by warping or transforming it to align with the input data. The transformation model W defines a set of image warps (*e.g.* affine, homography). The LK tracker attempts to find the set of parameters p of the transformation model W by iteratively minimising a cost function. It is assumed that a current estimate of p is known. LK is based on three core assumptions, [116]: 1) *brightness constancy* - the pixels associated with a region do not change brightness from frame-to-frame; 2) *temporal persistency* - the image transformation between frames is small; 3) *spatial coherence* - neighbouring pixels belong to the same surface. During MIS, the light source attached to the camera can invalidate the brightness constancy assumption due to its intensity in the central field-of-view. Furthermore, the light source is co-located with the camera and therefore mobile. The assumption of temporal persistency is related to the camera's frame rate and organ motion. Motion parallax and occlusion caused by tools can affect the spatial coherence of the tracked regions. These assumptions are rarely held in real-world tracking problem; non the less, LK approaches have been widely used [114] and adapted to function in MIS tracking.

In [89], for example, the LK algorithm is modified to work in 3D and to track the surface of the heart. By formulating the problem in 3D, it enables computationally efficient tracking with multiple cameras. In [108], an affine model is used in conjunction with a linear illumination compensation model and a robust estimator to address partial occlusions. Both [89] [117] demonstrate the ability of the LK tracker to follow multiple regions on the surface. A recent study, [101, 118], has performed a comparative analysis of recursive methods for tracking the stomach, liver and gall bladder. The authors compared a variety of transformation models (affine and homography), cost functions (forward additive and inverse), optimisation methods (Gauss-Newton and Efficient Second-order Minimisation) [119], and parameterisations (greyscale, Red, Green and

Blue (RGB) colour space, and mean of gradients). The most effective model is an affine model with mean gradients, forward additive cost function, and Gauss-Newton minimisation. It has also been determined that changes in illumination present the most significant challenge to all trackers, particularly the inverse composition.

The main strength of the LK tracker is its ability to find the best fit between the data and the template. Although non-linear deformation is not explicitly represented in the transformation model, it finds the best transformation available. LK based approaches have been shown to work in constrained conditions. However, there are fundamental weaknesses here, which greatly affect their application performance for MIS.

It has been shown [101, 118] that non-linear changes in illumination can violate the brightness constancy assumption and cause the tracking process to fail. Temporal persistency, alternatively, introduces two problems - recovering from failure and error propagation (*i.e.* drift). The frame-to-frame tracking nature of LK requires strong prior knowledge of the region's current location and parameters. As a result, if region tracking fails, it cannot be reinitialised easily. This presents a problem for MIS as occlusion becomes frequent with the use of surgical instrument. Error propagation, in general, is a product of deviation in the estimation of p, which propagates over time leading to drift. This problem is exacerbated by image noise and changes in illumination. Thus, in practice, forgoing the update removes drift, but this limits the range of image transformation that can be tracked and makes convergence less likely - especially in the presence of large tissue deformation.

Mean-shift trackers [120, 121] require a region to be tracked and represented as a histogram of the RGB values of the pixels. This histogram is a probability distribution that provides a look up table. The tracker defines a search window based on the size and previous known location of the region for each new frame. Each pixel in the search window is assigned a weight according to the probability of it belonging to the tracked region. Next, the centre of mass is computed for the window. The centre of the window is then moved to the location of the centre of mass whereupon the probabilities are computed again. This is performed recursively until the window becomes stationary. Mean-shift tracking is attractive for MIS because it does not rely on spatial information, and it can accommodate deformation [122, 123].

The main draw-back of recursive methods is the requirement of strong priors. This prevents long-term robust tracking in the presence of occlusion, which is common in MIS. Tracking-by-detection on the other hand, performs region detection for each new frame. The regions are transformed into a feature space where they are matched and the corresponding regions are identified. This approach is robust to occlusion since temporal information is not required. This, however, introduces a different problem. The region must be represented such that it is invariant to large image transformation: this can be computationally demanding. Despite such complication, the recent success of SIFT [124] and SURF [112] has demonstrated the potential of real-time implementation.

Tracking-by-detection systems are generally composed of a region detector, a descriptor, and a matching strategy. Scale invariant region detectors, as previously mentioned, represent a fundamental component of tracking-by-detection. The detectors enable region matching to be performed across scales and, in some cases, they correct for affine transformations, [103]. Region detection selects a subset of regions in the image for consideration for matching. This is important because descriptors can be computationally expensive. However, because only a subset of the image is considered for tracking, it is important that the detector has high repeatability. High repeatability means the same regions will be detected under different image transformations. Low repeatability can lead to tracking failure, regardless of the matching power of the descriptor.

In the tracking process, descriptors are used to represent the region of interest in feature space. Prior to converting the region to feature space, it may be warped to make it invariant to image transformation. The scale, rotation, and affine parameters of a detected region can be estimated, and the region is normalised in image space. Region descriptors select what information from the image will be used (*e.g.* greyscale, colour, gradient) and how this information will be represented (*e.g.*. energy in the co-occurrence matrix [125], non-uniformity of the run length matrix [125], histograms of gradients [126]). These choices dictate how effectively a region is distinguished from its surroundings and, consequently, the success of the tracker.

Many matching strategies exist to determine the corresponding regions between two images. Given two sets of regions encoded in feature space, the aim of the matching strategy is to find corresponding regions within these images. This establishes material correspondence over time during tissue tracking. Matching strategies can be one-to-one

(*e.g.* nearest neighbour), one-to-many (*e.g.* nearest neighbour ratio), or many-to-many (*e.g.* Random Sample Consensus (RANSAC) [127]). Global matching techniques, such as RANSAC, are used extensively for object tracking under a rigid model assumption.

The strengths of tracking-by-detection for object tracking in rigid man-made environments are well demonstrated: it is naturally able to deal with occlusion and is robust to large image transformations. Notable problems with this approach are localisation accuracy, region density, and outliers. Localisation accuracy can be affected by the discretisation of scale space and noise in the image. Region density can be affected by repetitive patterns. This is not a problem with object tracking where multiple regions are used to represent the object.

The main drawback of tracking-by-detection is the *ad hoc* modelling of image transformations and the assumptions used. These techniques are only designed to work within the range of image transformations they model. Their assumptions have been well tuned for man-made static environments. However, modelling non-linear tissue deformation is challenging. The choice of what information is suitable for discriminating a region from its surrounding is context-specific. The performance of descriptors can be affected by low contrast images, changes in illumination, and specular highlights. This makes the selection of a robust descriptor challenging especially for MIS.

2.3.2.2 Tissue Deformation Modelling

In addition to tracking, explicit deformation modelling is also important. Modelling can be done either statistically or parametrically, and it enables the prediction and anticipation of tissue motion. For example, deformation resulting from respiratory and cardiac cycles can be modelled as periodic or quasi-periodic signals [128] as illustrated in **Figure 2.9**. The dynamic motion of abdominal organs, such as the liver [129], is correlated to the periodic motion of the diaphragm. In [130], it is shown that organ positioning during free breathing motion is not cyclic. However, during MIS, a ventilator often controls respiration: it regulates the frequency of breathing and renders the motion periodic [131]. Modelling the respiratory cycle from CT and MRI data has been well studied. These imaging modalities have been used to demonstrate how a typical respiratory cycle is asymmetrically periodic with a longer dwell time at exhalation [132]. In MIS, the spatial arrangement of the organs is different to that of pre-operative setting

due to carbon dioxide insufflation. Organs shift, and the inflated cavity provides more room for this motion. The motion of the epicardial surface is more complex because it contains deformation caused by both cardiac and respiratory cycles. The deformations can be decoupled [89, 133] into their intrinsic components. A number of approaches are suggested for modelling cardiac motion, including Fourier series [87], vector autoregressive models [87], Taken's Theorem [134], and Linear Parameter Variant Finite Impulse Response Models [131].

Modelling non-periodic tissue deformation caused by instrument-tissue interaction or muscular contraction is more challenging. This is likely to require prior knowledge of the physical characteristics of the organ. These characteristics are patient and organ-specific and require the application of statistical shape models and finite element biomechanical analysis, such as those used in needle steering and surgical simulators [135].

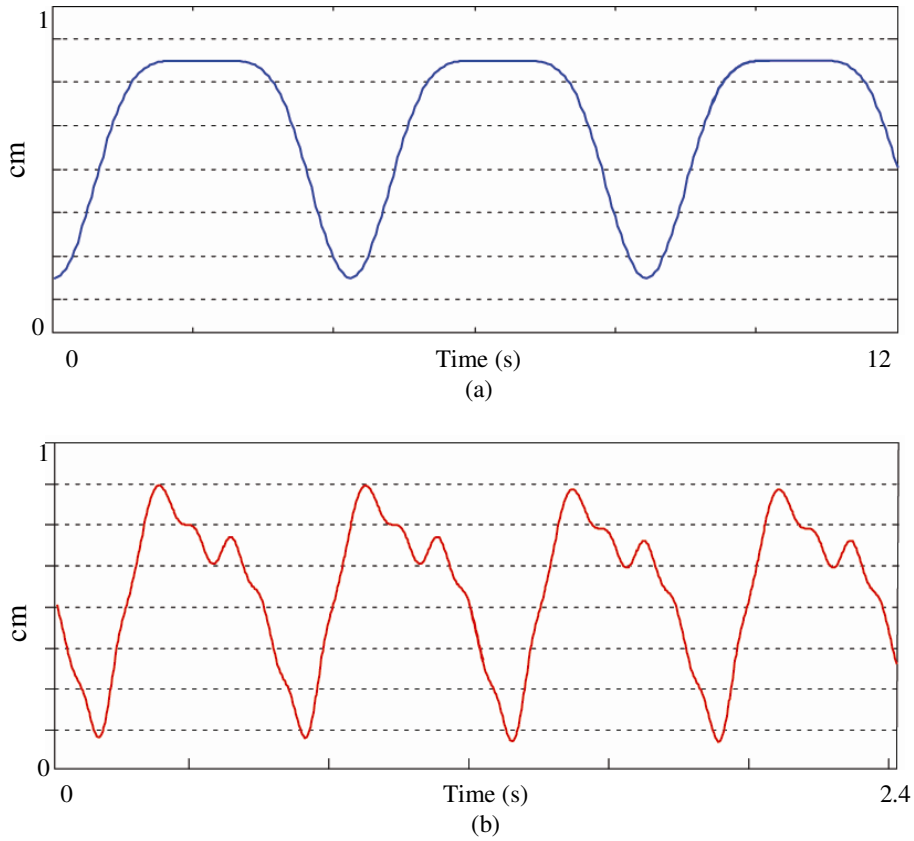


Figure 2.9 Tissue deformation is caused by the cardiac and respiratory cycles. Example signals of the **(a)** respiratory and **(b)** cardiac cycles illustrating their periodic and quasi-periodic nature.

Table 2.2 Summary of tissue morphology and structure estimation methods applied in MIS.

Organ	Recovered Scene Geometry	
	Static	**Deforming**
Heart	[85, 90, 107, 136]	[86, 88, 89, 92, 95, 108, 134, 137-142]
Abdomen / Liver / Gall Bladder / Kidney	[25, 143-146]	[101, 118, 122, 123]
Colon	[147-149]	-
Bladder	[150-153]	-
Oesophagus	[154-156]	-
Sinus	[126, 157-161]	-

2.3.3 Structure and Camera Motion Estimation

As previously mentioned, two common methods for structure and camera motion estimation are: Structure-from-Motion and SLAM. **Figure 2.10** compares these methodologies schematically. Both of these approaches assume the environment is structurally static: a weighty assumption for MIS. They have been employed, regardless, to situations wherein there are small deformations only. The theoretical frameworks of the two approaches, their limitations, and the extension of these techniques to non-static environments is described in the following section.

2.3.3.1 Structure-from-Motion

Structure–from-Motion has its origin in computer vision. It is a general term for methods that recover the structure of a scene and the motion of the camera (the reader is directed to [73] for a comprehensive mathematical formulation of the problem). A variety of Structure-from-Motion approaches exist, however, the basic framework is consistent and illustrated in **Figure 2.10**. It includes: 1) image registration and frame-to-frame camera motion estimation; 2) global optimisation or bundle adjustment where multiple images are registered; and 3) scene reconstruction.

The scene model describes the assumptions made regarding the structure and geometry of the environment and defines the relationship between pixels in different images. The pixel motion between images can be modelled as a projective (8 DOF), affine (6 DOF), similarity (4 DOF), Euclidean (3 DOF), or translation (2 DOF) transformation. The type of model selected depends on the assumptions made regarding the shape of the organ.

A planar assumption, as employed in [144, 151-153, 162, 163], enables the use of simple scene models. This assumption is acceptable if the endoscope is far from the organ and motion parallax is not observed. The planar assumption is not held in the oesophagus or colon because the anatomy is intrinsically tubular. Several authors have exploited this prior anatomical information [149, 154, 155] to model the surface of the oesophagus and colon as a generalised cylinder. Recovering structure and motion in an unknown, unstructured environment without prior knowledge or assumptions requires a projective scene model. Such models have been applied on the mouth [61], abdomen [60, 150, 164], colon [147], static heart [165], and bladder [166]. The full projective model can work in the presence of motion parallax and can accurately recover the camera motion in unstructured environments. However, the choice of the scene model is application specific. For example, if the desired end product is a 2D mosaic, a full projective model is not required.

Image registration is the method by which the relationship between two or more images is estimated according to the scene model. This usually involves estimating the fundamental matrix, which describes the change in position of the camera between images. However, camera motion estimation is a by-product of registration, and, if an inappropriate scene model is chosen (*i.e.* planar for a non-planar environment), the recovered camera motion will be inaccurate. Registration can be performed using direct alignment or region matching.

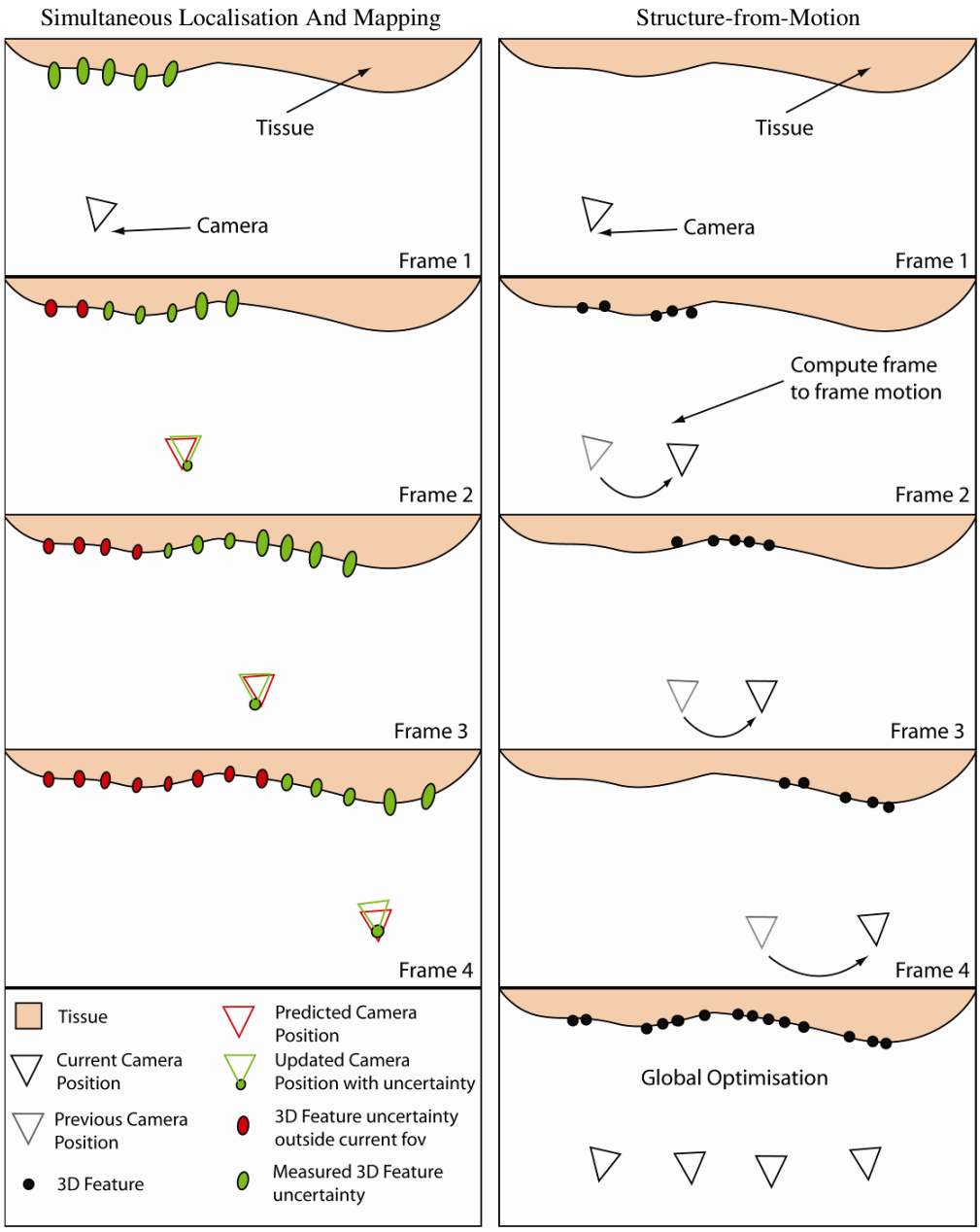


Figure 2.10 Illustration of structure and camera motion estimation. Simultaneous Localisation And Mapping (left) demonstrating sequential and incremental long term mapping with uncertainty estimates, motion prediction and updates. Structure-from-Motion (right) showing frame to frame pose estimation and global optimisation.

Direct alignment works with the pixel values in the image [149, 154, 155] and attempts to find a match between every pixel in a pair of images by warping one of the images. The set of warps applied to the image can be exhaustive, and this can be computationally expensive. It is, therefore, common to apply iterative techniques, such as LK, which minimise the error metric. Error metrics are usually the sum of squared difference, the sum of residual value, or the sum of absolute difference. Direct methods require the images to overlap and be sufficiently similar in order to converge and correctly perform alignment. This makes them well suited to sequential online applications. Performance directly on the pixel values makes these approaches robust to blur, however, they can suffer from the aperture problem and do not model view dependent specular highlights commonly found in MIS.

Region matching registration [60, 147, 150, 163, 165] extracts salient regions of interest and attempts to match these regions between pairs of images. This approach is advantageous because it does not require the images to be processed sequentially or have large, overlapping image areas, thus making it well suited to loop closing. Region detection and matching approaches described earlier can be used. Based on the region matches between images, the motion transformation can be estimated using a minimisation algorithm such as Least Means Squared or Levenberg Marquardt. Regions on the surface of tissue may appear visually similar and outliers can be removed using global techniques such as RANSAC, assuming the scene is rigid. Unlike direct alignment, region matching registration requires the surface of the organ to have detectable regions of interest within the image and for said regions to be accurately matched. Regions close to specular highlights may be ignored. Region matching enables more sophisticated global optimisation due to its non-sequential performance ability. A significant drawback of Structure-from-Motion is error propagation caused by frame-to-frame camera motion estimation. Small errors accumulate over time and can cause inaccuracies in the estimation of camera pose and structure. This problem can be addressed using global optimisation.

Global optimisation refers to the use of batch operations, or bundle adjustment, to register multiple images together, remove error propagation, and find an optimal set of transformations. Bundle adjustment is an iterative method, which searches for a non-linear model. To improve the chances of convergence, bundle adjustment is usually

initialised using estimations from a sequential approach. It generally is performed offline and is computationally expensive - making it inappropriate for online *in vivo*, *in situ* applications [60, 149, 164]. When used in applications such as post-operative diagnosis [151, 155, 162, 163], where high quality reconstructions may be beneficial, this extra offline computation does not pose a problem. Global registration provides an opportunity to ensure that loops are closed during navigation and the removal of specular highlights.

Scene reconstruction may be performed once the camera motion has been estimated. Correspondences from the regions matched between frames can be projected into 3D and triangulated to estimate their 3D position relative to the camera. This requires camera calibration information, which can be obtained in advance of, or after, the procedure. Scene reconstruction creates a sparse set of points, which represent the tissue structure. These points can be meshed together to create a solid object onto which textures can be applied.

Compositing determines the visual appearance of the 3D reconstruction. It is important to note that the surface is an approximation where alignment errors may be manifest, which, in turn, can cause blurring. Varied illumination in the images (caused by the point light source on the laparoscope) can also lead to visible seams, further disrupting the reconstruction. The compositing process includes mapping the pixels to the surface, selecting which images will contribute, and selecting how the pixels from these images will contribute to the final image. A variety of techniques have been proposed including sub-sampling, optimal seam selection, exposure compensation and multiband blending. These techniques have generally been developed to produce aesthetically pleasing results. For *in vivo* MIS applications, it is critical that techniques do not interfere with the current surgical field-of-view, this means incremental techniques must be registered to the current image. For post-operative diagnosis, it is important that when combining images, information that could affect the diagnosis is not removed.

Structure-from-Motion is a well established technique; however, it has a number of draw-backs in its application to MIS. *In vivo, in situ* MIS applications require a sequential implementation, which leads to error propagation and drift. This limits the accuracy and long-term application of the technique for MIS. The work described above is based on the assumption that the MIS environment is static. Non-rigid Structure-from-Motion has, thus far, been used for tracking faces [167, 168] and clothing [169]. This

technique is based on the factorisation method and shape basis representation. Conceptually, they are not suitable for real-time applications as the deformation is dealt with in an offline global optimisation step. Non-rigid Structure-from-Motion has also been applied to the heart [136]. However, it is used to deal with residual motion when constructing a static cardiac surface at a pre-selected point rather than a deforming surface during the cardiac cycle.

2.3.3.2 Simultaneous Localisation and Mapping (SLAM)

SLAM originated from research in autonomous robotic navigation: a detailed review of SLAM can be found in [170, 171]. SLAM was developed to solve two problems; consistent incremental environment mapping and localisation of a robot within a map. Prior to SLAM formulation, these problems were treated separately where either the map or robot location was assumed. This approach was unsuccessful as neither value could be assumed due to noise in sensor measurement. In [172], Smith, *et al,* present a seminal paper, which is credited with the development of the basic framework for simultaneously solving the localisation and mapping problem. This framework proposed the use of stochastic maps to represent the environment as a series of spatially related landmarks or features. A feature constitutes a position in the environment and a covariance matrix to model the positional uncertainty. Early work in SLAM on mobile systems used laser range finders [173], radar [174], sonar [175], and odometry sensors [174]. The framework has since been extended to work with stereo [176, 177] or monocular cameras [178].

SLAM has not been widely applied to MIS. Current approaches have mainly been based on two systems; V-GPS [179], developed for robotic localisation, and monoSLAM [178], developed for localisation of handheld cameras. V-GPS is classified as a SLAM system because it incrementally and sequentially builds a long-term map of the environment whilst simultaneously estimating the pose of the camera. However the localisation problem is solved using a Structure-from-Motion approach based on corresponding points between two monocular images. This has been applied to the sinus [159, 180] where deformation and tissue motion is minimal. The monoSLAM system is closely aligned with the original formulation of SLAM presented in Smith, *et al*. A probabilistic framework is used to represent the state of the system and noise in sensor measurement. MonoSLAM was first extended to work with stereo cameras (**Chapter 5**)

for MIS and has recently been applied to the abdomen [145] and oesophagus [156]. In MIS, the goal is to localise the laparoscopic camera and build a map of the tissue surface. A typical feature-based SLAM system [178] is illustrated in **Figure 2.10**, which contains the following key steps: system initialisation, feature tracking, prediction, camera pose estimation and initialisation of new features.

In [159, 160, 181], the V-GPS system builds a sparse 3D map of the sinus structure, which is registered to pre-operative CT. The motion of the camera is estimated between the current and next frame, similar to Structure-from-Motion. The intra-operative 3D sinus map has an arbitrary scale, and a scale invariant registration method is proposed for registration. The SLAM system contains the following steps: system initialisation, feature tracking, camera pose estimation, initialisation of new features, and registration to pre-operative data.

In the system initialisation step, features are first identified in the image plane to form an estimation vector (projected from the camera centre through the image plan). The camera depth is calculated using an arbitrary scale using monocular images. Two system initialisation methods are proposed in [159, 160, 181]. The first is manual initialisation, which requires the user to select a point in the image space and the corresponding points in the pre-operative data. This, can be error-prone: visual correspondence is difficult. The second approach uses the eight point algorithm [182] to estimate the essential matrix (rotation and translation) between two images. This approach requires camera motion between the images. System initialisation creates a map of features on the surface of the sinus, which is then used to localise the endoscopic camera in subsequent frames. The map is stored as direction vectors with distance magnitude scaled relative to a camera pose but not as 3D positions in a world coordinate frame.

The initialised features are temporally tracked in the 2D image space. The system is not specific to a feature tracking system. In [159, 160, 181], the authors track features on cadaver data by pre-processing the image with a gradient filter before finding correspondences using a sum of squared difference approach [117], which compensates for rotation, translation and illumination changes. Coloured spots are added to the surface for phantom data and are segmented in the hue plane for tracking.

The ego-motion of the camera is estimated between the two frames using the position of the tracked features in image space and the estimated distance magnitude. The approach [179] requires only three point correspondences between image frames. The camera motion is calculated by posing the problem as a least square-fitting problem solved iteratively by SVD. SVD estimates the rotation first and the translation between frames second. It requires initial estimates of rotation, translation, and depth of features from the camera. These are taken from the previous estimate or system initialisation, and SVD iteratively converges to an accurate estimate of these parameters.

New features can be added once the system is initialised. An approach is employed that uses the rigid body assumption and estimates the depth of a new feature's distance from the camera. This can be performed using the estimated ego-motion of the camera between two frames, or three frames for increased robustness. The depth of the new feature is estimated using the same scale as features in the existing map, thus maintaining a uniform scale. The 3D reconstruction is registered to pre-operative data in order to visualise the position of the endoscope in pre-operative. The resulting 3D reconstruction is only known up to an arbitrary scale, thus making registration difficult. The authors recovered the scale of the 3D map by comparing the relative structure of the map to that of the pre-operative model by computing the covariance matrix of point clouds to identify the dominant direction of the surfaces. The normalised eigenvectors were compared, and a rotation and scale transformation was computed to perform registration using Iterative Closest Point (ICP).

The monoSLAM system, on the other hand, alternates between a prediction step, where the motion of the camera is blindly predicted, and an update step, where the map is measured relative to the camera. A vision SLAM system consists of a state vector, a probabilistic framework, feature initialisation, a prediction model, and a measurement model.

The state vector contains the position of the laparoscope camera and a map of the tissue. The camera position is represented in the state by an XYZ position and roll, pitch, and yaw rotations. The state vector contains the velocity and angular velocity of the camera motion. The map is made up of the 3D XYZ position of a set of features or points. SLAM has been demonstrated with real-time performance in [178] wherein sparse maps

are generated, tracking up to 12 features at each frame with a total map size of 100 features.

The probabilistic framework in SLAM models the noise or uncertainty in the system. SLAM represents the joint probability between the features in the map and the position of the camera at a given point in time. It corresponds to the current estimate of the state and the associated uncertainty or noise. In MIS, [145, 156], the Extend Kalman Filter (EKF) has been employed which assumes Gaussian noise. In EKF, the uncertainty in the state estimates is represented in a covariance matrix as the variance from the estimate state. In the wider SLAM community, a variety of approaches have been implemented included Unscented Kalman Filters and Rao-Blackwellised particle filters (FastSLAM) [183, 184].

Features initialisation is specific to the optical configuration, monocular or stereo. In stereoscopic systems, features are matched in the stereo images. The 3D position of the feature is triangulated relative to the camera. In monocular systems, the 3D position is estimated by matching features temporally and partially initialising the feature using inverse depth, [145, 156]. SLAM ensures convergence of the map by employing a full covariance matrix between all features in the map.

The prediction or motion model defines how the camera is expected to move. This model involves two elements; 1) the deterministic element - the motion is estimated based on an assumption or a sensor (*e.g.* odometry); 2) the stochastic element – this is a probability distribution represented by a collection of particles or Gaussian. It represents the unknown motion of the camera which is non-trivial to model. A constant velocity, constant acceleration motion model, assumes smooth camera motion. This assumption may not hold in both handheld MIS and robotic assisted MIS, thus leading to system failure.

In the update step of SLAM, the measured state is compared to the predicted state. The measurement model provides a means of measuring the current system state. SLAM measures the location of the map features relative to the camera. The optical set-up defines the measurement model. In stereo SLAM, map features are measured in 3D using stereo feature matching and triangulation. In monocular SLAM, visible features are

projected into the image plane of the camera. Features are matched in image space and the measurement is performed in the 2D image plane.

SLAM has been studied extensively in the past decade, and the research area is now relatively mature with many of the fundamental problems solved. Its success is largely due to its probabilistic foundations and real-time capability. The key advantage of SLAM over Structure-from-Motion is online, long-term consistent mapping. This prevents error propagation and drift, making it well suited to revisiting previously observed areas.

Future research in the practical application of SLAM in MIS must focus on identifying more robust, long-term features, creating increasingly dense maps that cover larger areas, developing motion models better suited to rapid motion, recovering from failure, and incorporating information from additional sensors and robotic devices. The main challenge, however, is the theoretical treatment of deformation and dynamic soft-tissue motion. Within the wider vision community, SLAM has found application within non-static civil environments where motion occurs due to people and transportation. Non-static motions in the environment are treated as outliers, which can be identified using approaches such as RANSAC or Joint Compatibility Branch and Bound. These assume a global rigidity model and outliers are identified as features, which do not align with the rigid model. This approach requires part of the environment to be static, which may not be the case in MIS. The static environment assumption lies at the core of SLAM, yet removing this assumption is fundamental to the successful application of SLAM to soft-tissue environments.

2.4 Conclusion

MIS is a well established practice due to reduced hospitalisation, patient trauma, and recovery time. However, there remain many instrumentation, ergonomic design and visualisation challenges. This chapter has discussed the clinical demand for IGI for MIS and the need for handling soft-tissue deformation. The requirements have been outlined for intra-operative deformation recovery for non-rigid registration of pre- and intra-operative data. Meeting this requirement is complicated when the intra-operative imaging device is mobile. The current methods for estimating the position of endoscopes and laparoscopes in the operating theatres generally use additional hardware such as

electromagnetic trackers and optical tracking. Neither of these techniques is ideal because electromagnetic trackers suffer from interference, and optical trackers require line-of-sight and cannot be used for non-rigid devices such as endoscopes.

Estimating the motion and position of an endoscopic/laparoscopic camera can be achieved using the images collected by the camera itself. This will be the focus of the thesis' technical development: a challenging topic due to the visual appearance of tissue, and the potential for the robustness of the developed algorithms to be affected by a number of factors including small baseline stereo-optics, paucity of features, specular highlights, rapid camera motion, and large-scale tissue deformation. Estimating tissue deformation from a static laparoscopic camera is difficult, but the problem is further complicated when the laparoscope camera is mobile. In this case, motion observed by the laparoscopic camera may be a result of tissue deformation or camera motion, and they must be separated in order to enable accurate registration.

In the following chapters, each key component of the proposed framework will be addressed, in turn. First the robustness of existing region descriptors for tracking deformable tissue will be examined and a new method proposed to boost its performance by fusing multiple descriptors. It will then be demonstrated how online learning methods can be used to identify unique, visual region characteristics that can be employed to further improve tissue-tracking performance. Finally, the use of SLAM for MIS will be investigated and a new formulation of SLAM proposed without the static environment assumption.

Chapter 3

A Probabilistic Framework for Tracking Deformable Tissue

The last chapter reviewed existing methods for tracking deformable tissue and outlined their current, respective technical difficulties. The purpose of this chapter is to examine the use of detect-and-match tracking techniques for estimating tissue deformation. State-of-the-art-region descriptors are evaluated on MIS data and a supervised feature selection algorithm is used to systematically identify descriptors robust to deformation. A probabilistic framework is proposed for fusing the most discriminative descriptors to boost matching performance based on a Bayesian network. The performance of the proposed method is quantitatively evaluated on both simulated data and *in vivo* MIS data-sets.

3.1 Tissue Tracking

3.1.1 Region Descriptors and Matching

As described in the last chapter, tracking-by-detection systems consist of a region detector, a descriptor and a chosen matching strategy. Generally, it is the region detector that models invariance to scale, and the descriptor encodes invariance to rotation and to deformation. A total of 21 descriptors have been evaluated on MIS data to examine the relative performance of state-of-the-art descriptors. These include nine descriptors,

which represent spatial information, and four descriptors, which represent colour information. Seven spatial descriptors have been extended to work in colour space using techniques outlined in [185]. These descriptors are listed in **Table 3.1**, Most of these descriptors, such as template matching and image moments, are widely used in image analysis. To be comprehensive, brief explanations of some complex descriptors are provided in this chapter. The source code for the descriptors is available at [186], [187], or was made available by the authors.

Table 3.1 A summary of the region descriptors evaluated in this study. Colour descriptors are identified by a 'C' prefix.

ID	Descriptor
CC, CCC	Cross correlation, a 9×9 uniform sample template of the smoothed feature.
MOM [188], CMOM	Moment invariants computed up to the 2nd order and 2nd degree.
DI [189], CDI	Differential Invariants, Gaussian derivatives are computed up to the 4th order.
SF [190], CSF	Steerable Filters, Gaussian derivatives are computed up to the 4th order.
Spin [191], CSpin	Spin images, a 2D histogram of pixel intensity measured by the distance from the centre of the feature.
GIH [192]	Geodesic-Intensity Histogram, A 2D surface embedded in 3D space is used to create a descriptor which is robust to deformation.
SIFT[124], CSIFT [193]	Scale Invariant Feature Transform, robust to scale and rotation changes.
GLOH [194], CGLOH	Gradient Location Orientation Histogram, SIFT with log polar location grid.
SURF [112] CSURF	Speeded Up Robust Features, robust to scale and rotation changes.
CCCI [195]	Colour Constant Colour Indexing, A colour based descriptor invariant to illumination which uses histogram of colour angle.
BR-CCCI [196]	Sensitivity of CCCI to blur is reduced.
CBOR [197]	Colour Based Object Recognition, a similar approach to CCCI using alternative colour angle.
BR-CBOR [196]	Sensitivity of CBOR to blur is reduced.

3.1.2 Geodesic-Intensity Histogram (GIH)

A deformation invariant descriptor is proposed in [192] where the image is treated as a 2D distribution embedded in 3D space. Deformations are homeomorphisms between two images that allow pixel intensities to change location but not value. The descriptor is based on the Spin image descriptor, but Geodesic distance is instead used for calculating distances.

On the embedded surface, the first two coordinates are proportional to the (x,y) coordinates in the image, and the third coordinate of the embedded surface is proportional to intensity, with an aspect weight α. The descriptor is built as a Geodesic-Intensity Histogram (GIH) by sampling points on the embedded surface. The geodesic distance on the embedded surface becomes less sensitive to deformation as α increases. When $\alpha = 1$, it is exactly deformation invariant as α controls the weight given to intensity and the image coordinates. A large value for α means intensity is more important than spatial information in the image coordinates. Therefore when $\alpha = 1$, only intensity is considered. Since the surface is assumed to be homeomorphic it can be considered invariant to deformation.

GIH differs from most descriptors because it does not compute a descriptor based on the size and shape of the region. Instead, it uses the detected region location as a starting point to define its own local support region. Geodesic-level curves are extracted around the local neighbourhood for a given α (the authors use 0.98) at set intervals. Points are then sampled at intervals along the curves. These sampled points are used to build the GIH in a similar manner to that of the Spin image. It should be noted that an assumption of the deformation as homeomorphic is not necessarily true, and specular highlights can cause problems. In addition, this approach requires well-defined local support regions with high contrast, which are not always available for MIS.

3.1.3 Scale Invariant Feature Transform (SIFT)

The Scale Invariant Feature Transform, or SIFT [124] descriptor, has been shown in [194] to perform well under large image transformations. It is one of the most cited works in relevant literature. It captures a large amount of information relating to spatial intensity patterns while being robust to small deformation and localisation errors. The

descriptor has been specifically designed to be invariant to rotation and scaling while being partially invariant to illumination. It is inspired by the neurons in the primary visual cortex, which respond to gradient at a particular orientation and spatial frequency but allow small positional shifts in the gradient.

With SIFT, a region's orientation is determined by computing the gradient magnitude and orientation for each point in the region. The scale of the region is obtained from the region detector, which is usually a Difference Of Gaussian (DOG). Detection is performed on images convolved with Gaussian filters of varying scales. The scale is determined as a local maximum in the filtered images across scales. The histogram of orientated gradients is computed on the filtered image associated with a given scale. Adjacent pixel difference is used to compute the gradient $m(x,y)$ and orientation, $\theta(x,y)$ for each image sample, *i.e.*,

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad \textbf{(3.1)}$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1)) \,/\, (L(x+1,y) - L(x-1,y))) \quad \textbf{(3.2)}$$

An orientation histogram with 36 bins is used to represent the 360 degrees of possible orientation. Each sample point is added to the histogram weighted by its gradient magnitude along with a Gaussian window 1.5 times the scale of the region. This places emphasis on gradients at the centre of the region. The highest peak in the orientation histogram is taken to be the orientation of the region.

SIFT explicitly models changes in orientation and scale making it robust to large image transformations. The use of tri-linear interpolation enables SIFT to cope with small deformations resulting from changes in view-point. The *ad hoc* decision concerning which image transformations to model, limits the application of this technique to deforming environments. Large spatial changes are not explicitly modelled, and in addition, SIFT assumes that changes in gradient are the most important information to encode. In MIS images, specular highlights and low contrast regions can affect the calculation of gradients and consequently, the robustness of the descriptor.

3.1.4 Gradient Location-Orientation Histogram (GLOH)

Gradient Location-Orientation Histogram (GLOH) [194] is a variation of the SIFT descriptor which improves matching while maintaining the same descriptor dimensionality (*i.e.* 128). The gradient-orientation and magnitude are computed in the same manner as SIFT, however, the region is not divided into sub areas using a Cartesian grid: a log polar grid is used. The log-polar grid divides the region into eight bins in the angular direction and three bins in the radial direction. Although GLOH is more distinctive than SIFT, it requires an offline training phase using PCA to determine the dominant elements of the histogram. The use of PCA to remove less dominant elements of the histogram is context-specific and requires the training data to be an accurate representation of the test data [194].

3.1.5 Speeded Up Robust Features (SURF)

Speeded Up Robust Features (SURF) are presented in [112]. It is conceptually similar to SIFT but with a fast and simple implementation. A fast Hessian region detector is used to determine the orientation of the region and compute a descriptor based on Haar-wavelet responses. In order to define the orientation of the region, Haar-wavelets are used. The neighbourhood is taken as a circle with a radius six times the scale of the region. The wavelet responses are weighted with a Gaussian model and represented as vectors based on horizontal and vertical response strengths. This allows the dominant orientation to be estimated by calculating the sum of the responses with a sliding window. It has been demonstrated that, when combined with a fast Hessian region detector, the SURF descriptor can outperform SIFT and GLOH on images with changing viewpoints or scale. SURF can also handle image blur, changes in brightness, and JPEG compression.

3.1.6 Colour Model

Colour information can be combined with spatial information to improve descriptor performance. The spatial descriptors outlined above compute representations using intensity information only. The SIFT descriptor can be extended to work in the colour space as proposed in [193]. Other descriptors can also be extended, in a similar fashion, to work in a colour space as shown in **Table 3.1**. The colour model [193] is derived from the Kubelka-Munk model for physical reflectance. A Gaussian colour model is used to represent spectral and structural information. The spectral differential quotients are computed using a linear transform from Red, Green and Blue (RGB) space, and the

spatial differential quotients are computed using a Gaussian convolution. This colour model has been shown to be robust to shadow, changes in illumination, specular highlights, and noise.

3.1.7 Colour Constant Colour Indexing (CCCI)

Colour Constant Colour Indexing (CCCI) [195] is a technique developed to perform object retrieval using colour information alone. The approach is based on the work of Swain and Ballard [199], which is extended to be invariant to changes in illumination using the retinex theory. CCCI is able to cope with spatial variance in illumination, intensity, and colour. The descriptor outlined in [199] represents objects in a colour histogram that counts the number of pixels of a given value in opponent colour space. Colour histograms are well suited to representing deforming objects. They make no use of geometric or structural information and therefore can cope with changes in orientation, viewpoint, and non-linear deformation. During MIS, however, illumination varies spatially as a result of the point light source used. CCCI is selected due to its robustness to variation in illumination. The main drawback of the colour histogram approach is the assumption that colour is sufficient to distinguish different regions.

3.1.8 Colour Based Object Recognition (CBOR)

Colour Based Object Recognition (CBOR) [197] is a colour histogram descriptor which extends CCCI for 3D objects. The CCCI algorithm is invariant to changes in illumination and is based on the Mondrian, or flat world, assumption. This assumes that neighbouring locations in the image have the same surface normal. Such an assumption can be problematic for complex geometry exhibiting significant change in surface orientation. In [197], the authors propose a new colour constant ratio independent of illumination, colour variation, and surface geometry. The approach introduces a dichromatic reflection model, and the descriptor is based on the ratio of the surface albedo. CBOR is robust to spatial changes in the intensity and spectral distribution of illumination. The main strength of this method is its invariance to surface orientation; however, it is designed to work with narrow band images, white illumination and matt surfaces.

3.1.9 Blur Robust (BR) Colour Ratios

Blur Robust (BR) colour ratios [196] extend CCCI and CBOR, enabling them to work in with image blur. Colour constant image descriptors, such as CCCI and CBOR, are based

on image derivatives, the use of which make these methods sensitive to blur, which can be caused by rapid motion, out of focus, or rapid camera motion – all common influences during MIS. The colour ratios are computed using a Gaussian derivative at a given scale giving the ratios have an associated scale. The intrinsic robustness of the method is equivalent to robustness to changes in the scale of ratios. The application of the BR approach to the colour ratios of CCCI and CBOR results in BR-CCCI and BR-CBOR.

3.2 Descriptor Selection and Fusion

With such a large number of descriptors available, the key issue remains: how to select the appropriate descriptor for specific tracking applications. Some of the descriptors may provide similar information while others may provide complementary information that may be fused to enhance the overall quality of the tracking process. This section outlines an offline method for selecting a sub-set of the most discriminate descriptors and an online approach for fusing these descriptors to improve tissue tracking. The first step is to perform descriptor selection with known ground truth data. The method for acquiring the training data is outlined in **Figure 3.1**. Training data is acquired by detecting regions of interest on each frame in the sequence. Corresponding regions between the first frame and subsequent frames are identified by an expert user. In the following sections the key steps of descriptor selection and fusion will be explained in detail.

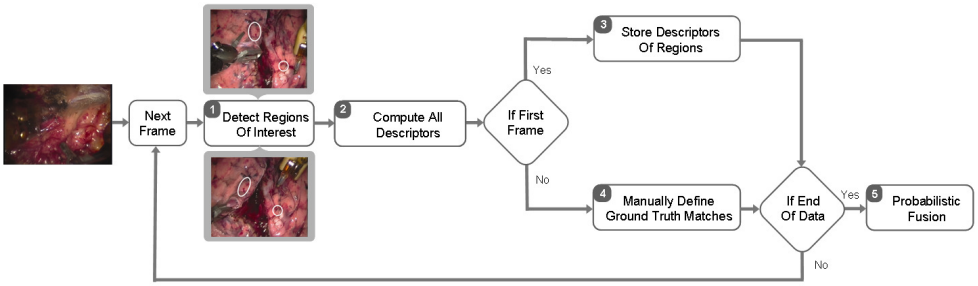


Figure 3.1 Flow chart illustrating the six steps in the generation of training data from laparoscopic video data. A region detector is applied to each frame of the video, regions of interest are detected and descriptors are computed. Tracking is performed relative to the first frame and corresponding regions in subsequent images are manually defined.

The purpose of descriptor selection is to identify a sub-set of the descriptors described above and to acquire the best tracking performance. This problem can be formulated within a machine-learning framework as feature selection. The aim of feature selection is to identify a small subset of features, or descriptors, that can be efficiently computed

without reducing the overall discriminatory power. This is achieved by eliminating irrelevant or redundant descriptors that contribute little to the classification accuracy. Feature selection can be performed by individual performance ranking or sub-set evaluation. Individual performance ranking can lead to redundancy making a sub-set evaluation approach preferable. A Bayesian Framework for Feature Selection (BFFS) as proposed by [200] is used.

3.2.1 Bayesian Framework for Feature Selection (BFFS)

The BFFS, illustrated in **Figure 3.2**, is a machine-learning algorithm formulated as a filter approach. It maintains the inference accuracy while reducing the complexity of multiple descriptors. The approach benefits from a descriptor selection that is based on data distribution rendering it unbiased towards a specific model. The BFFS defines a search strategy and an objective function to evaluate performance and can be implemented with either forward or backward search strategies. In forward search, the algorithm starts with an empty set and, in each iteration, the descriptor contributing the largest increase in discriminative power is added to the set. Backwards search starts with a set containing all descriptors and, in each iteration, the descriptor responsible for the smallest reduction in discriminative power is removed. It is determined [201-203] that backward elimination is less prone to feature interaction and is used in this thesis.

Criteria for selecting optimal descriptors is based on the expected Area Under Curve (AUC) where the curve is the Receiver Operating Characteristic (ROC). The ROC curve describes the relationship between the *sensitivity* and *1 - specificity* of a classifier, making the value of the AUC a direct measure of the descriptors. Within the BFFS framework, the expected AUC is taken to be a metric for the intrinsic discriminability of the descriptors classification performance.

For clarity, s*ensitivity* is defined as the ratio of correctly matched regions to the total number of corresponding regions between two images such that

$$sensitivity = \frac{\#Correctly \quad Matched}{\#Correspondences} \tag{3.3}$$

1-specificity is defined as the ratio of incorrectly matched regions to the total number of non corresponding regions such that

$$1 - specificity = \frac{\#\,Incorrectly \quad Matched}{\#\,Non \quad Correspondences} \tag{3.4}$$

wherein a region is defined as corresponding if it exists and is detected in both images, and a region is classified as correctly or incorrectly matched according to a matching strategy. In this work, threshold-based matching is used to compare different regions. Two regions are matched if the similarity between the descriptors is below a threshold. In keeping with [204], the Mahalanobis distance is used to compare SF, CSF, DI, CDI, MOM, and CMOM. The Euclidean distance is used for SIFT, CSIFT, GLOH, CGLOH, SURF, CSUFR, CCCI, BR-CCCI, CBOR, BR-CBOR, GIH, Spin, and CSpin.

The BFFS framework is based on the definition of irrelevance derived from Bayesian theory. Irrelevance is used as a metric in the objective function to remove descriptors that contribute little to the overall combined performance. This is based on conditional independence of the posterior probability

$$P(y \mid \boldsymbol{\mathcal{G}}^{(1)}) = P(y \mid \boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)}) \tag{3.5}$$

where $P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)}) \neq 0$, $\boldsymbol{\mathcal{G}}^{(1)}$ and $\boldsymbol{\mathcal{G}}^{(2)}$ are sets of descriptors. This definition asserts that given $\boldsymbol{\mathcal{G}}^{(1)}$ event y is conditionally independent of $\boldsymbol{\mathcal{G}}^{(2)}$.

An alternative definition of relevance can be created using the likelihood ratio such that

$$L(\boldsymbol{\mathcal{G}}^{(1)} \mid\mid y = a, y \neq a) = L(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid\mid y = a, y \neq a) \tag{3.6}$$

This states that given $\boldsymbol{\mathcal{G}}^{(1)}$ event y is conditionally independent of $\boldsymbol{\mathcal{G}}^{(2)}$ for any assignment of $y = a$ and $P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)}) \neq 0$ where

$$L(\boldsymbol{\mathcal{G}} \mid\mid y = a, y \neq a) = \frac{P(\boldsymbol{\mathcal{G}} \mid y = a)}{P(\boldsymbol{\mathcal{G}} \mid y \neq a)} \tag{3.7}$$

Therefore the definition of the conditional independence of the posterior probability can be rewritten as,

$$\frac{P(\boldsymbol{\mathcal{G}}^{(1)} \mid y = a)P(y = a)}{P(\boldsymbol{\mathcal{G}}^{(1)} \mid y = a)P(y = a) + P(\boldsymbol{\mathcal{G}}^{(1)}) \mid y \neq a)P(y \neq a)} \tag{3.8}$$

which is equal to,

$$\frac{P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid y = a)P(y = a)}{P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid y = a)P(y = a) + P(\boldsymbol{\mathcal{G}}^{(1}, \boldsymbol{\mathcal{G}}^{(2)}) \mid y \neq a)P(y \neq a)} \tag{3.9}$$

therefore,

$$\frac{P(\boldsymbol{\mathcal{G}}^{(1)} \mid y = a)}{P(\boldsymbol{\mathcal{G}}^{(1)} \mid y \neq a)} = \frac{P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid y = a)}{P(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid y \neq a)} \tag{3.10}$$

The likelihood ratio is intrinsically linked to the ROC curve, which can be created by plotting the *sensitivity* and *1-specificity* of a descriptor as threshold β of the likelihood ration $L(\boldsymbol{\mathcal{G}}^{(1)} \mid\mid y = a, y \neq a)$ varies. This can be used to derive an additional definition of irrelevance using the ROC.

The likelihood ratio is equivalent to the slope of the ROC at a given value of β. Therefore, in accordance with [205], the *sensitivity* P_{sen} and 1-specificity P_{1-spec} can be defined as

$$\begin{cases} P_{sen} = \int_{L(\boldsymbol{\mathcal{G}} \mid\mid y=a, u \neq a) > \beta} P(\boldsymbol{\mathcal{G}} \mid y = a) d\boldsymbol{\mathcal{G}} \\ P_{1-spec} = \int_{L(\boldsymbol{\mathcal{G}} \mid\mid y=a, u \neq a) > \beta} P(\boldsymbol{\mathcal{G}} \mid y \neq a) d\boldsymbol{\mathcal{G}} \end{cases} \tag{3.11}$$

such that, as the threshold β is varied from 0 to ∞, P_{sen} and P_{1-spec} vary from 0 to 1. A definition of irrelevance may be created based on the ROC as

$$ROC(\boldsymbol{\mathcal{G}}^{(1)} \mid\mid y = a, y \neq a) = ROC(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \mid\mid y = a, y \neq a) \tag{3.12}$$

The ROC curve can be used to compare the individual performance of given descriptors, and the AUC can be used as a metric to evaluate their relative performance. The larger the AUC of a descriptor, the higher its discriminative power - although the shape of two ROC curves with equal AUC may vary. The AUC provides a metric for individual descriptor evaluation, however, within a fusion framework, the combined discriminatory power of the descriptors must be evaluated.

In order to evaluate the irrelevance in a multi-class situation, the expected AUC E_{AUC} can be used. It is observed that the ROC curve increases monotonically with the addition of descriptors. Therefore, if the discriminating power of $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ completely overlap (*i.e.* $\mathcal{G}^{(2)}$ cannot match any regions of interest that cannot be matched using $\mathcal{G}^{(1)}$), then it may be removed without affecting the ROC curve or the AUC such that;

$$AUC(\mathcal{G}^{(1)} \,||\, y = a, y \neq a) = AUC(\mathcal{G}^{(1)}, \mathcal{G}^{(2)} \,||\, y = a, y \neq a) \tag{3.13}$$

irrelevance can be defined for the multi-class situation such that

$$E_{AUC}(\mathcal{G}) = \sum_{i=1}^{N} P(y = a_i) \sum_{k=1\cdots N}^{k \neq i} H_{ki} AUC(\mathcal{G} \,||\, y = a_i, y = a_k) \tag{3.14}$$

where

$$H_{ki} = \frac{p(\mathcal{G}) \mid y = a_k)}{\sum_{j=1\cdots N}^{j \neq i} p(\mathcal{G}) \mid y = a_j)} for(i \neq k) \tag{3.15}$$

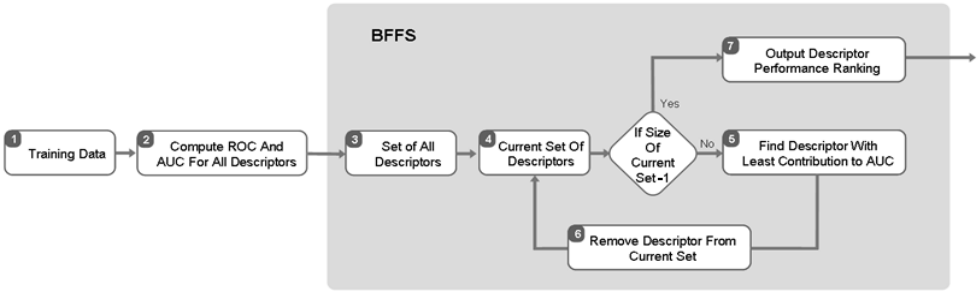


Figure 3.2 Flow chart illustrating the use of training data to perform descriptor selection with a BFFS framework. The backwards search strategy is shown where the process starts with the set of all descriptors and iteratively removes the worst performing until the set contains one descriptor.

3.2.1.1 BFFS Objective Function

At each step in the descriptor selection framework, a descriptor d_i is eliminated from the descriptor set $\mathcal{G}^{(k)}$, resulting in a new set $\mathcal{G}^{(k)} - \{d_i\}$. The eliminated descriptor d_i minimises the objective function $D(d_i)$. The performance metric is the expected AUC,

and the BFFS maximised performance by discarding, at each step, the most irrelevant or redundant descriptor that adds the smallest changes in the expected AUC.

$$D\left(d_i\right) = E_{AUC}\left(\mathcal{G}^{(k)}\right) - E_{AUC}\left(\mathcal{G}^{(k)} - \left\{d_i\right\}\right) \quad \textbf{(3.16)}$$

where $\mathcal{G}^{(k)} = \left\{d_j,\ 1 \le j \le n-k+1\right\}$ is the descriptor set at the beginning with k being the number of iterations, and the function $E_{AUC}()$ returns the expected AUC.

Unlike irrelevant descriptors, which are uninformative, redundant descriptors may offer useful information despite having little impact on the expected AUC. Redundant descriptors may be discriminative in their own right but are correlated with another descriptor. It is preferable to retain these discriminative, redundant descriptors and remove uninformative, irrelevant descriptors as these redundant descriptors may perform well on test data. In the evaluation function of **Equation (3.16)** redundant and irrelevant descriptors are treated equally. This is because they both make a small contribution to the overall performance of the model. In order to retain redundant descriptors and discard irrelevant descriptors, the following objective function has been proposed:

$$\mathrm{D}_r\left(d_i\right) = -\left(1-\omega_1\right) \times E_{AUC}\left(\mathcal{G}^{(k)} - \left\{d_i\right\}\right) + \omega_1 \times E_{AUC}\left(d_i\right) \quad \textbf{(3.17)}$$

where the weighting factor ω_1 ranges between 0 and 1, [206]. This function minimises the discriminability of the eliminated descriptors whilst maximising the discriminability of the selected descriptor set.

3.2.2 Probabilistic Descriptor Fusion for Tissue Tracking

In order to combine the selected features together, a fusion framework as outlined in **Figure 3.3** is proposed. The key component of this framework is a Naïve Bayesian Network (NBN) used to fuse the descriptors selected by the BFFS. This provides probabilistic fusion of the subset of descriptors that can be used for region matching. The NBN classifies two regions as either matching or not matching by fusing the similarity measurements between the descriptors and estimating the posterior probabilities.

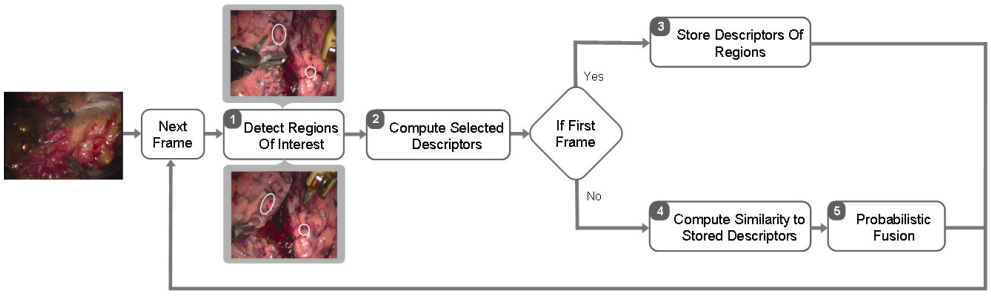


Figure 3.3 Flow chart illustrating the steps in online regions tracking using descriptor fusion. In the tracking-by-detection framework a region detector is first applied. Image descriptors are computed for the detected regions and fused in a NBN to improved tracking performance.

The NBN is a probabilistic classification technique that can fuse multiple sources of information. The NBN assumes all evidences are statistically independent and it classifies matching and non matching regions according to the posterior probabilities of the fused similarity measures. In accordance with Bayes' theorem, the fused posterior probabilities can be estimated as:

$$P(C \mid D_1, D_2, ..., D_k) = \frac{P(D_1, D_2, ..., D_k \mid C)P(C)}{P(D_1, D_2, ..., D_k)} \approx \alpha P(D_1, D_2, ..., D_k \mid C)P(C) \quad \textbf{(3.18)}$$

where $D_1, \ldots, D_k$ are the distance measurement between the two regions, k is the number of descriptors selected by the BFFS, C is the hypothesis of match or non-match, and α denotes the normalising constant. If all the descriptors are statistically independent, **Equation (3.18)** can be rewritten as

$$P(C \mid D_1, D_2, ..., D_k) \approx \alpha P(C)P(D_1 \mid C)P(D_2 \mid C), ..., P(D_k \mid C) \quad \textbf{(3.19)}$$

where $P(D_k \mid C)$ is the conditional probability of evidence D_k given hypothesis C. **Equation (3.19)** formulates the posterior estimation of the NBN, and can be modelled using a Directed Acyclic Graph (DAG) as shown in **Figure 3.4**.

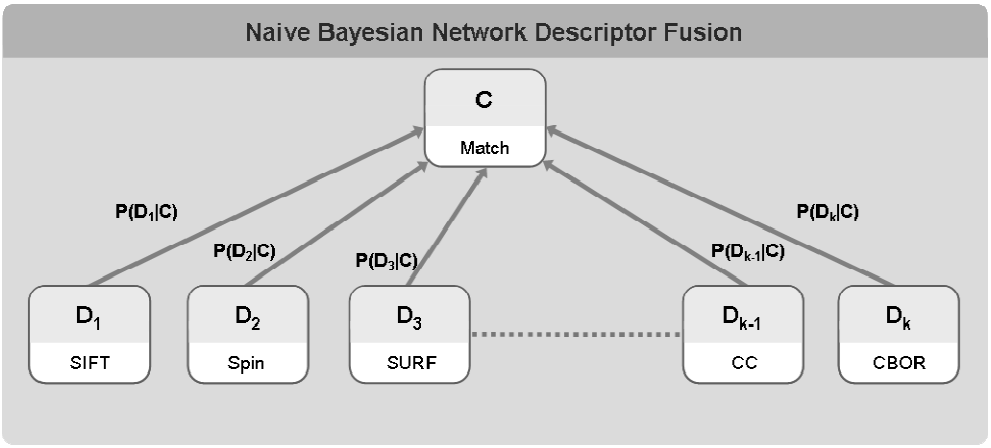
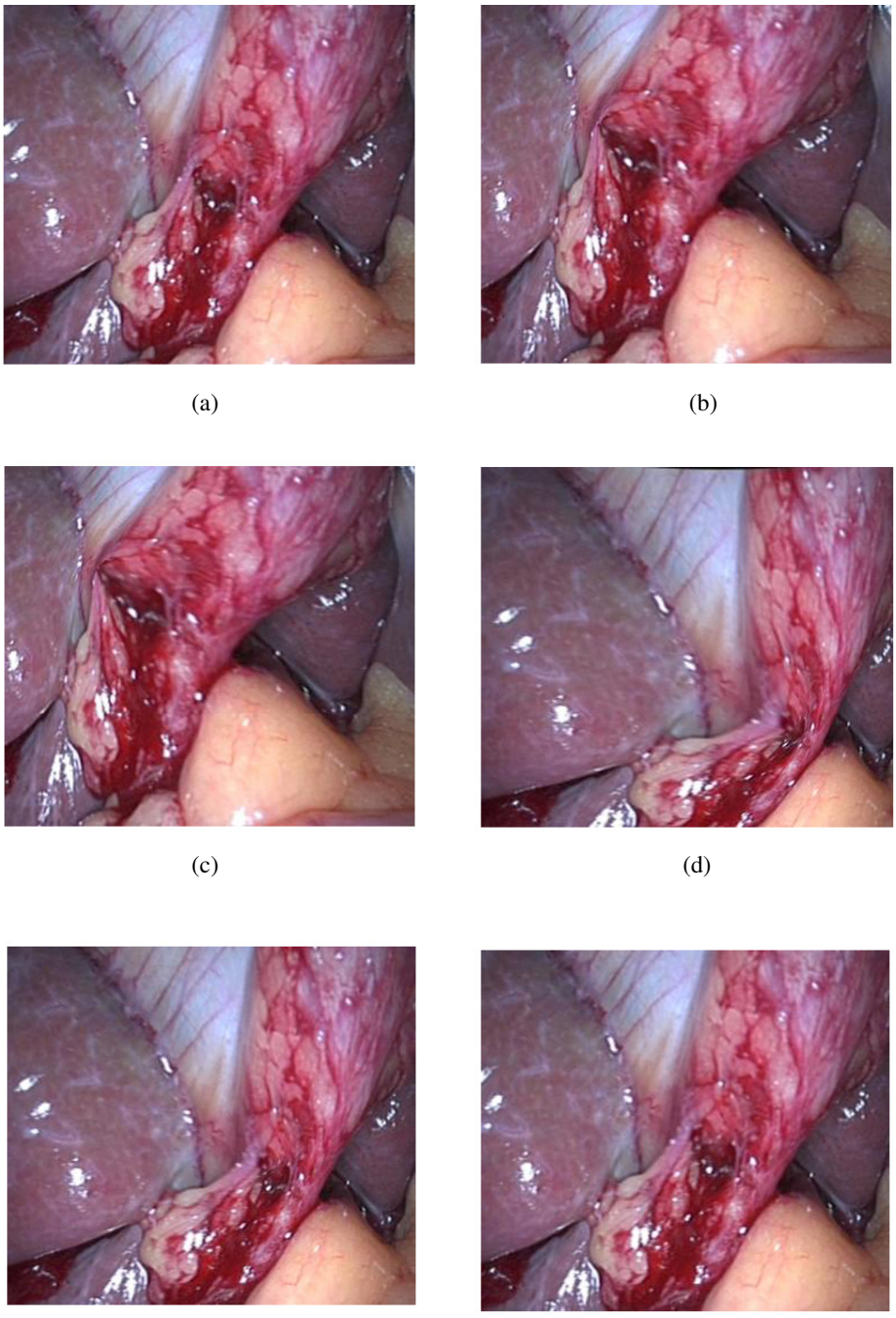


Figure 3.4 DAG visualisation of a NBN for descriptor fusion for classifying a region. The DAG contains nodes representing descriptors D and classification C. The nodes are joined together by directed arcs which represent the conditional probability between the nodes.

3.3 Experiments and Results

The above descriptor selection and fusion framework is evaluated with both simulated and *in vivo* data. The chosen sequences exhibit large tissue deformation resulting from instrument-tissue interaction. In essence, the proposed framework can be categorised as a tracking-by-detection approach. The framework's ability to increase the number of correctly tracked regions, and their persistence with respect to tissue deformation, is evaluated in addition to its capability to reinitialise tracking after failure.

The quantitative evaluation herein performed involves three metrics, descriptor *sensitivity*, *1-specificity* (results are presented in the form of ROC curves), and detector repeatability. The s*ensitivity* or recall is computed according to [114] and [204] and **Equation (3.3)** where correspondence is the number of regions of interest that are successfully redetected at each frame, and thus, have the potential to be matched. This method, of computing the number of correspondences, ignores the repeatability of the region detector, enabling the descriptors' performance to be evaluated independently. *Sensitivity* is a measure of the density of tracked regions: a good region tracker will have high *sensitivity*. *1-specificity* is the ratio of incorrectly matched regions to the number of non-corresponding regions as defined in **Equation (3.4)** Non-correspondences are regions that are not redetected by the region detector or fall outside of the current field-of-view. In general, a good region tracker will have low (*1-specificity*) values.



(a) (b)

(c) (d)

(e) (f)

Figure 3.5 Simulated data. An image, acquired during a laparoscopic cholecystectomy illustrating the gall bladder and liver, is textured onto a 3D deformable mesh. The mesh is deformed with a mixture of Gaussians. (**a-f**) Show the deformed surface used to validate the tracking algorithm.

3.3.1 Simulated Experiments

The simulated sequence for evaluating the performance of the proposed method is shown in **Figure 3.5**. The data-set is generated by texturing a 3D mesh with a real laparoscopic image showing the gall bladder, liver and cystic duct. The 3D mesh is deformed with a Gaussian mixture model and noise is added to the image. The parameters of deformation were selected to create images that visually replicate tissue deformation resulting from tissue-instrument interaction. The total length of the data stream is 100 frames. Points are tracked on the surface of the mesh at 10 frame intervals and re-projected onto the image plan to provide a ground truth data-set.

The results from the BFFS descriptor selection framework are shown in the AUC graph in **Figure 3.6 (f)**. The AUC curve indicates the descriptor IDs of the top performing descriptors in descending order. The most discriminative descriptor, based on the training data, is Spin, and it is evident that the overall discriminability of the system is improved by incorporating additional descriptors. The framework selects a subset (Spin, GIH, CSIFT, SIFT, GLOH, SURF and CGLOH) of the 21 descriptors. The order of the selected descriptors does not directly correspond to the descriptors' individual performance: note the order is defined by how much additional information the inclusion of a descriptor creates. This is demonstrated by the inclusion of the CSIFT descriptor before the SIFT descriptor. The individual ROC analysis of the two descriptors in **Figure 3.6 (a)** and **Figure 3.6 (c)** shows that SIFT outperforms CSIFT. The inclusion of CSIFT substantiates its ability to providing new information that is not captured by the Spin, GIH, or SIFT.

The performance of the descriptor fusion framework is evaluated using ROC curves shown in **Figure 3.6 (a-e)**. The ROC curves represent time-collated data and provide a simple metric of performance for the duration of an entire video sequence. The fused descriptor is shown graphically as point Fn where *n* represents the number of additional descriptors (*i.e.* F1 is Spin and GIH). The graphs illustrate, for an acceptable *specificity*, descriptor fusion can obtain a higher level of *sensitivity* than any individual descriptors. This enables the fusion technique to match more regions robustly. The top performing individual descriptor is Spin. For the level of *specificity* achieved with fusion F5, Spin's *sensitivity* is 11.96% less making its region density lower. Alternatively, to obtain the same level of *sensitivity* achieved with descriptor fusion, *specificity* must be

compromised. With the Spin descriptor, this results in an increase of 19.16% and a reduction in region matching robustness. However, in this data-set, not all of the $\mathbf{F}_n$ fusions offer an overall improvement. It has been shown that: 1) the addition of more descriptors, whilst theoretically improving performance, may not improve actual performance; and 2) not all descriptor fusions outperform individual descriptors. The former is demonstrated in the data by F2 outperforming F3. The latter is evident in the performance of F1, which is below the ROC curve of several individual descriptors. These issues are attributed to differences in the training data and the test data. However, it should be noted that the chosen fusion data-set, F5, has out-performed all individual descriptors.

The performance of individual descriptors varies. It was found that Spin, GIH, SIFT, SURF, and GLOH performed well on the detected regions. However, all descriptors converted to work in colour space performed unsatisfactorily compared to the original descriptors. The colour space is designed to enable descriptors to discriminate between significantly different colours (*e.g.* red and blue). In MIS the variance in colour is small and colour descriptors cannot easily distinguish between small changes in colour. Although the performances of colour descriptors were inferior, these descriptors still provide useful information in the fusion framework.

The colour histogram descriptors are evaluated in **Figure 3.6 (e)**. These descriptors do not contain structural information and only use colour to represent the region. It is clear, as indicated by the ROC curves, that these are the most inferior performing descriptors. Their performance is close to $sensitivity = 1 - specificity$. The descriptors with blur reduction perform marginally better. These results demonstrate that colour information alone does not contain sufficient variance to discriminate between different local regions in MIS data.

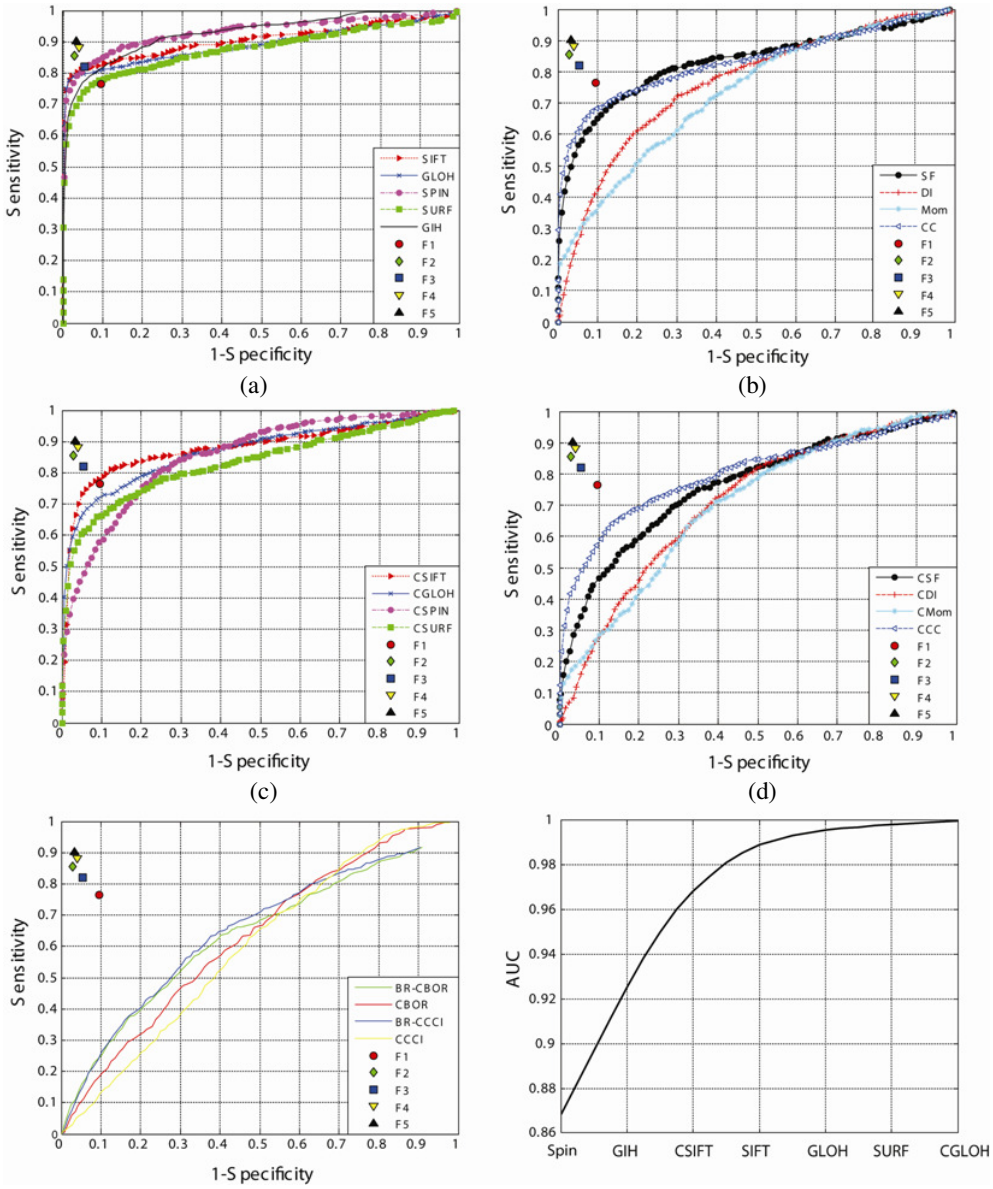
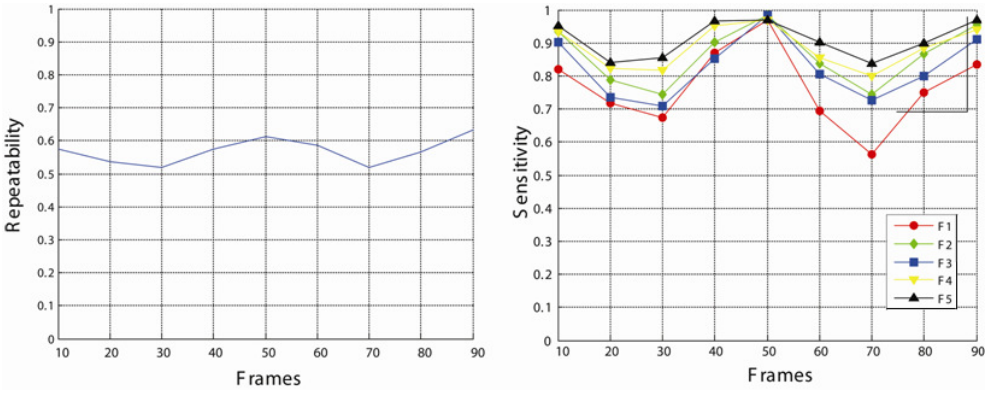


Figure 3.6 Simulated data. **(a-e)** ROC (*sensitivity vs. 1-specificity*) graphs for individual descriptors and fused descriptors F1-F5. The matching threshold is varied to obtain the curves. **(f)** AUC graph generate by BFFS selection framework.



(a) (b)

Figure 3.7 Simulated data. **(a)** Detector repeatability and **(b)** *sensitivity* of fused descriptors with respect to time.

The proposed descriptor fusion framework is evaluated further with respect to time and deformation in **Figure 3.7 (a-b)**. The ROC curves provide a simple metric for evaluating performance of the entire video sequence. However, it does not show changes in performance over time. **Figure 3.7 (b)** shows the *sensitivity* over time for the descriptor fusion framework. On this graph, a clear pattern is visible with three peaks and two troughs. The troughs correspond to extreme deformation, and the peaks correspond to images that are visually similar to the first image and exhibit a small amount of deformation. It is evident that the tracking performance is compromised during deformation. The proposed fusion approach successfully increases the number of regions that can be tracked. This graph also demonstrates how the tracking-by-detection approach is capable of reinitialising tracking after failure. However; this graph and the ROC curves, are computed without taking the repeatability of the region detector into consideration.

The repeatability of the region detectors is shown in **Figure 3.7 (a)**. The average repeatability of the detectors is 56.9%, which is lower than desired. The repeatability of the detectors is affected by surface deformation, the projection of the texture onto a curved surface and a small-scale change introduced by the simulation and addition of noise. A similar pattern of performance is observed in the *sensitivity* graph in **Figure 3.7 (b)**. Performance corresponds to deformation where repeatability is higher when less deformation is exhibited and lower at the extremes of deformation. This has a compounding effect on the region-tracking density. It is observed that regions not redetected are those undergoing the largest deformation and, therefore, will be the hardest to match using the descriptors.

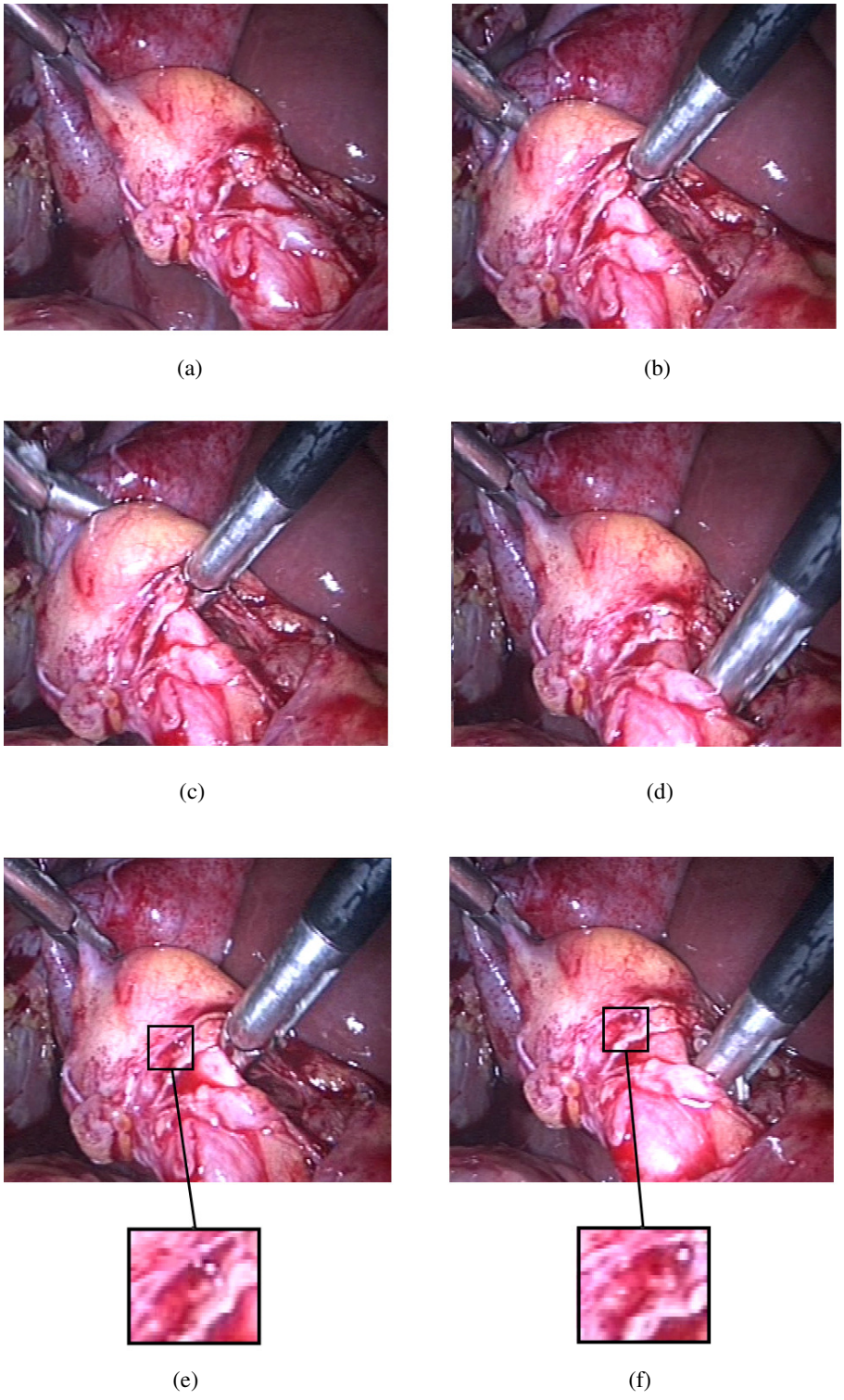
3.3.2 *In Vivo* Experiments

Figure 3.8 shows an *in vivo* sequence from a laparoscopic cholecystectomy procedure. The sequence consists of 1700 frames, and the ground truth data is defined by an expert user at 100 frame intervals. A total of 40 regions of the tissue surface were matched throughout the sequence. Only regions of the tissue in contact with the tools were selected. Specular highlights are identified by a threshold in the saturation channel, and regions of interest near specular highlights are ignored.

The AUC graph for *in vivo* descriptor selection using the BFFS framework is shown in **Figure 3.9 (f)**. The BFFS identifies GIH as the most discriminant descriptor and the best subset includes (GIH, SURF, Spin, SIFT, GLOH, CC). This set is similar to the set identified in the simulated data. Once again, it is evident that additional descriptors can improve the overall tracking performance. GIH and Spin share similar approaches and GIH can be seen as an extension of Spin. **Figure 3.9 (a)**, demonstrates their performance on this data-set as similar, and both outperform SURF. However, the BFFS framework prioritises SURF above Spin, illustrating that the feature selection prioritises contribution to overall performance above redundant data.

It is evident from the graphs in **Figure 3.9 (a-e)** that descriptor fusion improves the overall discriminative power of descriptors. The fused descriptors, F5, obtain a specificity of 0.235, this is a 30.63% improvement in *sensitivity* over GIH (the best performing descriptor) at the given s*pecificity*. This shows that the descriptor fusion framework is capable of matching more regions than any individual descriptor for deforming tissue. For this data-set, the five fused descriptors F1-F5 all outperform the individual descriptor performance in *sensitivity* and (*1 –specificity)*.

In **Figure 3.9**, the fused descriptors F3, F4, and F5 are tightly clustered in the top left of the ROC graph indicating a good performance. It can be seen from F3 to F5, there is a decrease in the 1 - *specificity* and, therefore, a reduction in the number of mismatched features. However, this is accompanied by a reduction in s*ensitivity*. This implies that the inclusion of GLOH and CC prevents erroneous matching at the expense of s*ensitivity*. The variance in s*ensitivity* and (1 – *specificity*) between F3, F4, and F5 are small, and the tight clustering suggests that a good result can be achieved with F3, which includes four descriptors.



(a) (b)

(c) (d)

(e) (f)

.

Figure 3.8 *In vivo* data. **(a-e)** A selection of laparoscopic images collected during a laparoscopic cholecystectomy. The images show deformation resulting from tissue-tool interaction. **(e-f)** Show local deformation of a region of interest.

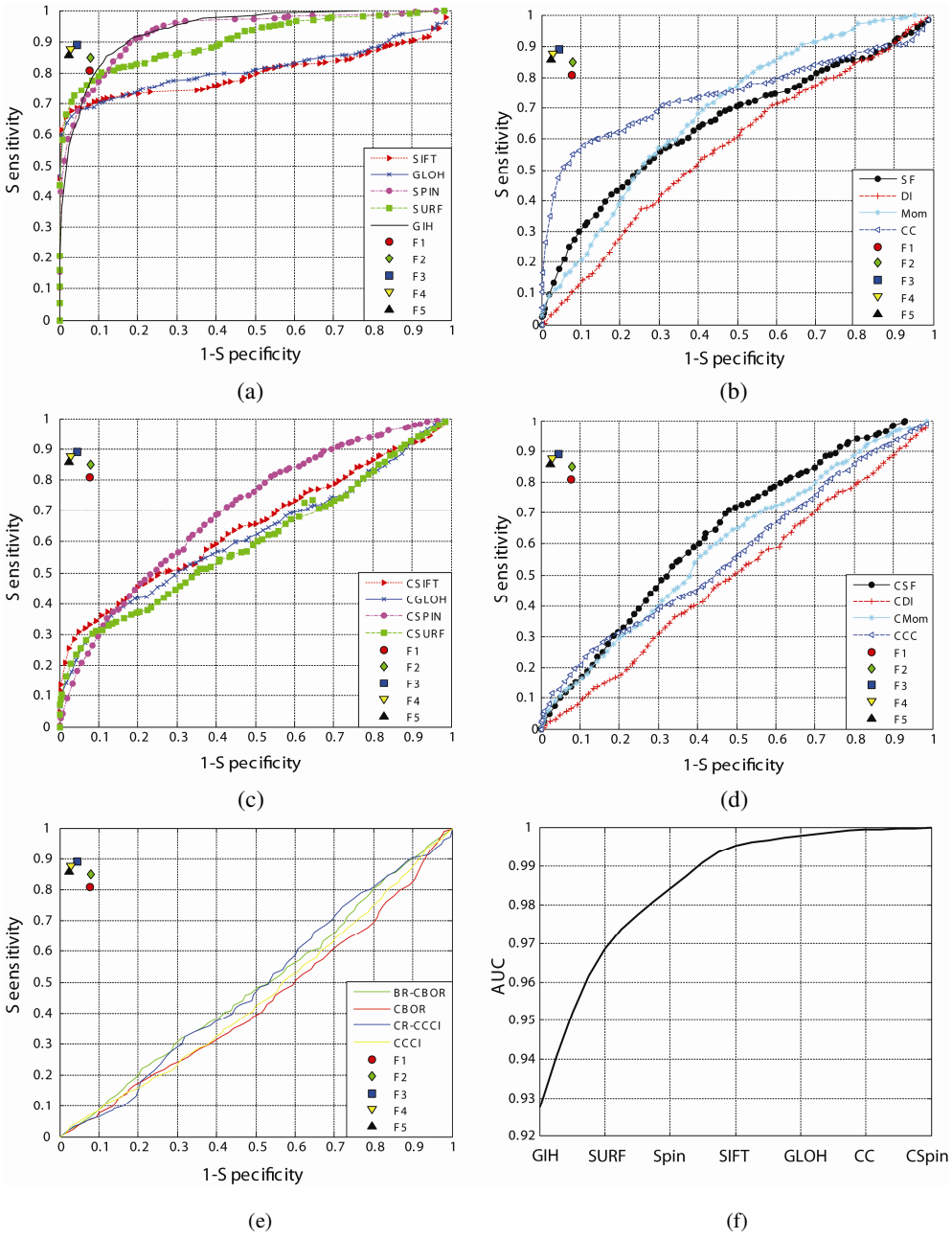


Figure 3.9 *In vivo* data. **(a-e)** ROC (*sensitivity vs. 1-specificity*) graphs for descriptors. **(f)** AUC graph generated by BFFS selection framework.

For this *in vivo* data-set, the relative individual performance of the descriptors is similar to that of the simulated data. The best performing descriptors are Spin, SIFT, SURF, GIH, and GLOH. These approaches encode structural information. Once again, when modified to work in colour invariant space, they perform worse than the original descriptors. The reduction in performance, however, is more pronounced than the simulated data. This is attributed to the complex illumination conditions in MIS. The colour histograms CBOR, BR-CBOR, CCCI, and BR-CCCI perform poorly and, at some points, drop below the line of $sensitivity = 1 - specificity$. The images from the *in vivo* sequence are shown in **Figure 3.8**, and it is evident that regions of interest on the tissue surface do not have unique colour distributions to allow for consistent tracking.

Figure 3.10 (b) shows the *sensitivity* performance of the descriptor fusion framework with respect to time. It can be seen that F1 to F5 perform relatively well with small dips in performance when deformation is increased. However, the repeatability of the detectors shown in **Figure 3.10 (a)** is low. At the end of the sequence, the repeatability is below 0.4 and more than 60% of the regions are no longer trackable. Low repeatability is caused by tissue deformation and changes in scale, illumination, and surface artefact due to bleeding and specular highlights.

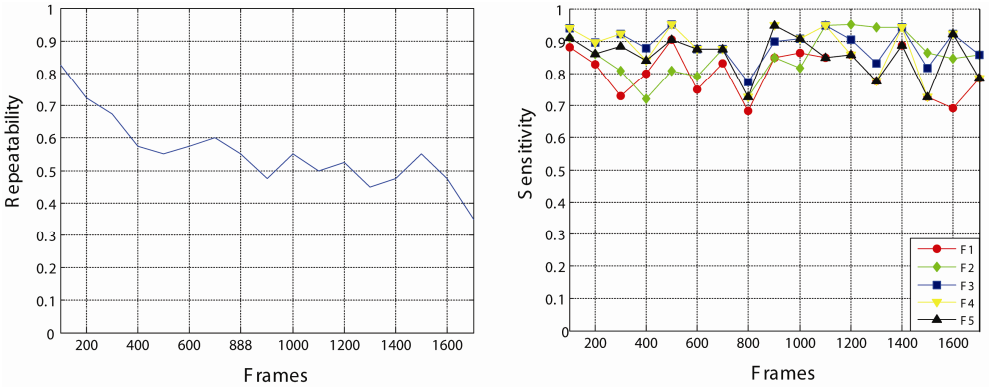


Figure 3.10 *In vivo* data. **(a)** Detector repeatability and **(b)** *sensitivity* of fused descriptors with respect to time.

The practical value of the proposed framework is further demonstrated in **Figure 3.11**. The fused descriptor F5 is used to track an *in vivo* sequence collected during a lung lobectomy procedure performed with the da Vinci robot. A 3D reconstruction of the

scene is generated using the stereoscopic laparoscope: camera calibration is carried out prior to the procedure. Regions of interest are detected in the first frame of the video and matched across the entire image sequence for temporal deformation recovery. Only regions that are successfully tracked through both time and space are used for 3D depth reconstruction. The sparse 3D reconstruction is overlaid on a dense reconstruction, created using only the stereo images without temporal tracking to provide a context for the tracking. The fused method is compared with that of the SIFT descriptor as a baseline. The SIFT features are matched using the nearest neighbour ratio matching. It is evident that the proposed fusion method has greater temporal persistence and density.

3.4 Discussions and Conclusion

This chapter has investigated the use of vision-based tracking algorithms to estimate tissue deformation in MIS. A probabilistic framework based on tracking-by-detection is proposed. Quantitative evaluation of the performance of selected state-of-the-art image descriptors has been performed and a selection framework proposed for determining the best-performing descriptors for MIS sequences. Descriptor selection is executed prior to tracking in an offline training phase where descriptor performance is quantitatively evaluated with known ground truth data. The relative performance of the descriptors is computed with the BFFS to create sub-groups of complementary descriptors with increased discriminative power. The sub-group of descriptors is then fused online using a Bayesian network to improve overall tracking performance. The performance of the proposed framework is quantitatively evaluated on both simulated and *in vivo* data-sets.

Tracking-by-detection removes the assumption of temporal persistency; however, it requires the application of a region detector at each frame. This chapter has shown that region tracking can be re-initialised after failure. This is important in MIS because tracking failure can be caused by occlusion due to surgical instruments. Although deformation is generally not explicitly represented by image descriptors, this framework is capable of boosting tracking performance with increased density and persistence compared to conventional approaches. This represents a step towards making efficient and effective use of visual cues for deformation recovery.

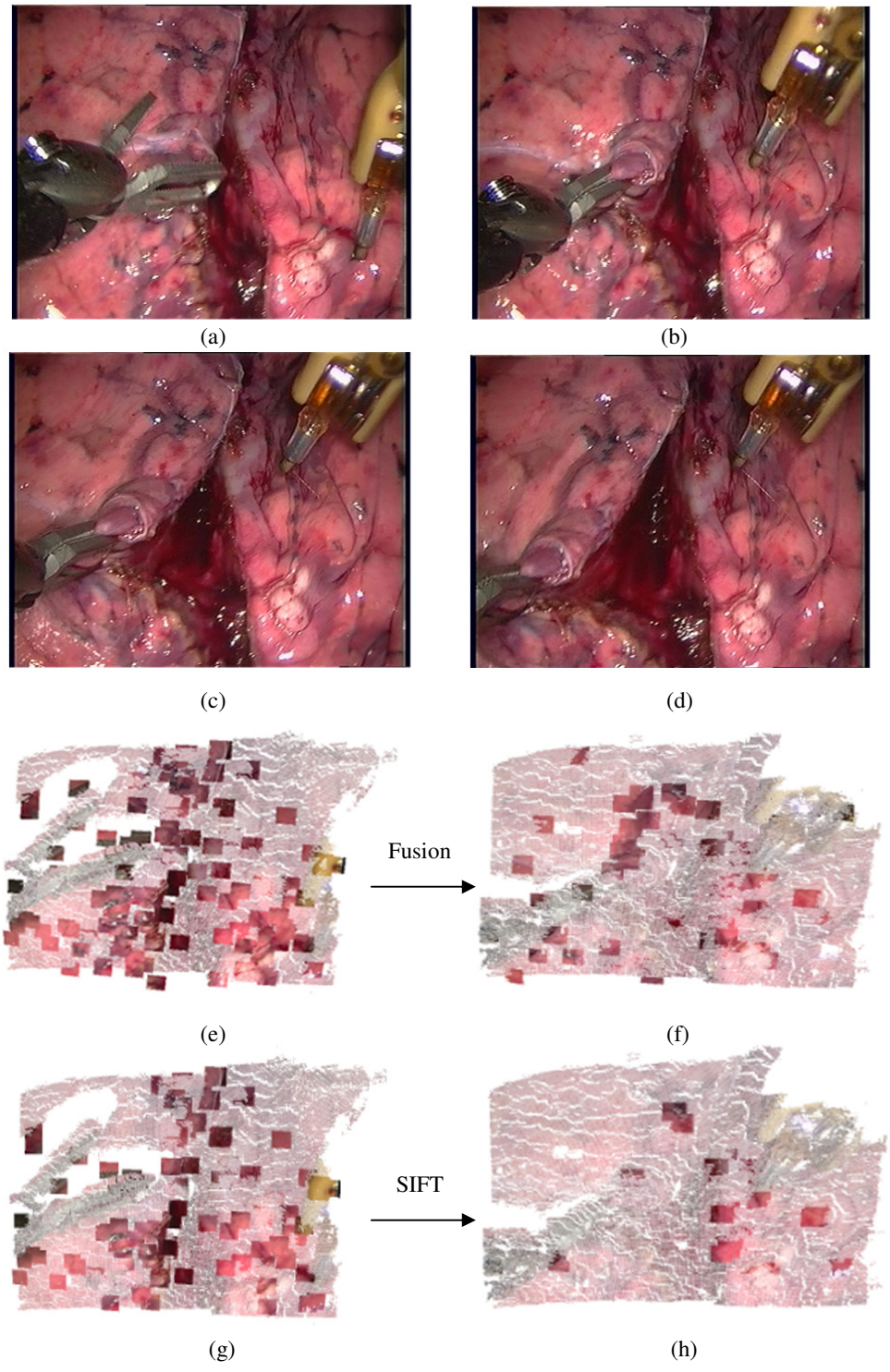


Figure 3.11 (a-d) Laparoscopic footage of tissue deformation resulting from tool interaction. The footage was acquired during a robotic assisted lung lobectomy procedure. 3D deformation tracking and depth reconstruction based on computational stereo. **(e-f)** Descriptor fusion and **(g-h)** SIFT. SIFT was identified by the BFFS as the most discriminative descriptor for this image sequence.

The presented results demonstrate how the repeatability of region detectors was not sufficiently high to ensure continuous tracking. This may be improved in practice via manual parameter tuning; however, the variation in visual appearance of soft-tissue renders these parameters context-specific. The current framework is based on pre-defined image descriptors, which make *ad hoc* decisions concerning what information will be used for matching. Although the selection framework identifies which descriptors are the most informative, the capability of the system is limited by the pre-defined descriptors. The next chapter proposes a systematic framework for automatically identifying context specific information online.

Chapter 4

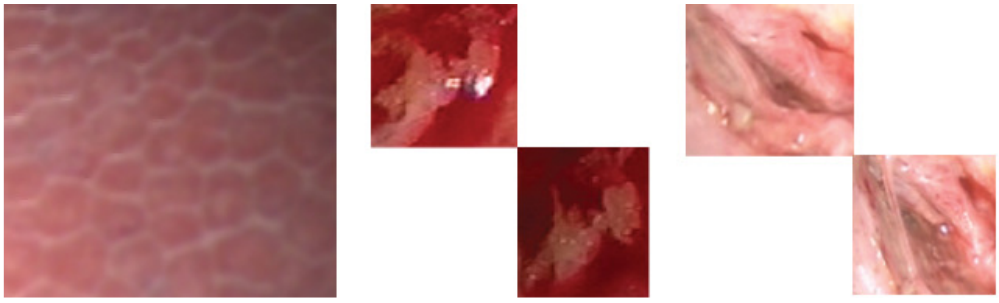
An Online Learning Approach to Tissue Tracking

In the previous chapter a method was described for tracking tissue to recover deformation from laparoscopic images. This method used a supervised machine learning approach for optimal descriptor. The research-findings demonstrate the feasibility of using a learning framework to improve region tracking. It demonstrated that learning discriminative visual cues and fusing information can increase the overall tracking performance with respect to *sensitivity* and *specificity*.

In Minimally Invasive Surgery (MIS), however, the appearance of the surgical scene varies greatly and is subject to constant change. In this case, the descriptors selected offline may not provide the optimal performance. To ensure region density and persistency it is necessary to learn the visual representations online and to adjust the tracking process appropriately. The purpose of this chapter is to present an algorithm which adapts to the scene context, learns a representation for deformation, and is robust to drift, occlusion, scaling, rotation and artefacts (such as smoke resulting from diathermy during MIS).

4.1 Introduction

The work presented in the last chapter explored which subset of descriptors yields optimal performance. An alternative methodology is to learn what makes a region distinguishable from surrounding regions of interest. This approach has been used in handwriting recognition [207], image classification, object detection [113] and corner detection [208]. These approaches pose the region-matching problem as a classification issue and consequently require training sets. PCA and kernel PCA are used to provide efficient representation of the objects offline prior to tracking [209-212] or with online adaptation [213]. It is not possible to generate training data in advance of an MIS procedure, and offline training is non trivial. This leaves an online training approach and the matter of how to deal with *in vivo* situations, such as those shown in **Figure 4.1**, where tissue undergoes non-linear deformation, has repetitive texture or pattern, or is occluded by specular highlights.



(a) (b) (c)

Figure 4.1 Specular highlights, non-linear tissue deformation and variation in the visual appearance of tissue makes tissue tracking challenging. A segment of the liver is shown in **(a)** with repetitive surface pattern. Non-linear tissue deformation on the cardiac surface is shown in **(b)** with occlusion caused by specular highlights. Tissue deformation resulting from respiration is shown in **(c)**.

In this chapter a method for learning robust representations for regions is described. To demonstrate the practical and clinical application this method is used to extract the 3D motion of tissue. The intrinsic components of the tissue motion are extracted into respiration and cardiac motion which is subsequently modelled.

4.2 Learning Region Descriptors

As previously mentioned, region tracking during MIS is influenced by the non-linearity of tissue deformation, changes in scale and orientation, and variation in lighting and occlusion. In addition, MIS is affected by organ appearance, which may lack distinctive anatomical landmarks (*e.g.* surfaces of liver and kidney). The ability to track a region successfully is governed by how distinguishable it is from surroundings regions. This is largely determined by the representation of the region, which consists of two elements: 1) what information is encoded and 2) how the information is encoded. For the former, such information includes colour, edges, lines, corners, scale, orientation, intensity, texture, and gradient. This information should contain sufficient variance to enable regions to be distinguished whilst allowing them to undergo changes in appearance due to image transformation. For the latter, information may be encoded as probability density histograms, histograms of gradients, templates, points, contours, and active appearance models. A successful encoding method should represent the relevant and discriminative information whilst ignoring irrelevant data.

The choice of the matching strategy is closely related to the encoding method. The matching strategy is a means of ascertaining whether a region exists in a new image by comparing the encoded information of a region with encoded information from a new image. The matching strategy includes how the encoded information is compared (*e.g.* cross correlation, minimisation, earth mover distance, and sum of squared difference) and how a match may be determined (*e.g.* threshold, nearest neighbour, or nearest neighbour ratio).

The choice of what information to encode and how it is represented can be context-specific. As shown in [120], colour information can provide sufficient variance between an object and its background such that a deformable object can be tracked using mean-shift. In [214], the Lucas Kanade (LK) tracker encodes structural information assuming brightness constancy, temporal persistence, and spatial coherence. By constantly updating the encoded information, the temporal persistence assumption can be maintained, thus enabling tracking using a variety of image transformations. As described earlier approaches such as Scale Invariant Feature Transform (SIFT) [124] detect scaled regions and encode gradient information as spatially oriented histograms and imposes geometric constraints on the visual appearance of regions. This makes *ad*

hoc assumptions about the most discriminative information and how to best encode it. Assumptions are also made regarding what type of image transformation the encoded information will be invariant to. These methods perform well when the underlying assumptions hold; however, as described above, the MIS environment is subject to constant change, and *ad hoc* modelling of tissue appearance can be problematic.

An alternative approach is to learn what makes a region distinguishable from its surroundings, what information is most discriminative, and how best to encode and match the region (this learning approach was discussed in the previous section). These methods require offline learning, access to prior information or knowledge of expected image transformations. In addition, robust estimators commonly used to remove outliers during matching are needed. The method proposed in this section identifies the most discriminative information for tracking a specific region and updates adaptively as the tracking process progresses. Training data is extracted online by bootstrapping an LK tracker and synthetically generating data. This enables the approach to accommodate unknown tissue deformation and standard image transformations such as scale and rotation. The proposed algorithm consists of six main steps as shown in **Figure 4.2**: 1-2) region tracking is initially performed using a LK algorithm and, subsequently, the online approach, 3) generation of synthetic data, 4) building the online tracker and learning a representation for a region, 5) adapting and updating the region representation, and 6) tracker evaluation.

4.2.1 Building the Online Tracker

The region tracking problem can be formalised as a classification problem [113, 213] within the learning framework. The aim is to classify a given region in a new image as a true match and classify all other regions as false. Training the classifier requires a set of data with true and false labels. To this end, the data can be represented as a set of image patches which can be acquired via either manual labelling for offline learning or, if the appearance of the region can be well modelled, synthetically generating data. In this chapter, the method combines online learning with synthetically generated data to improve robustness to tissue deformation and changes in scale and orientation.

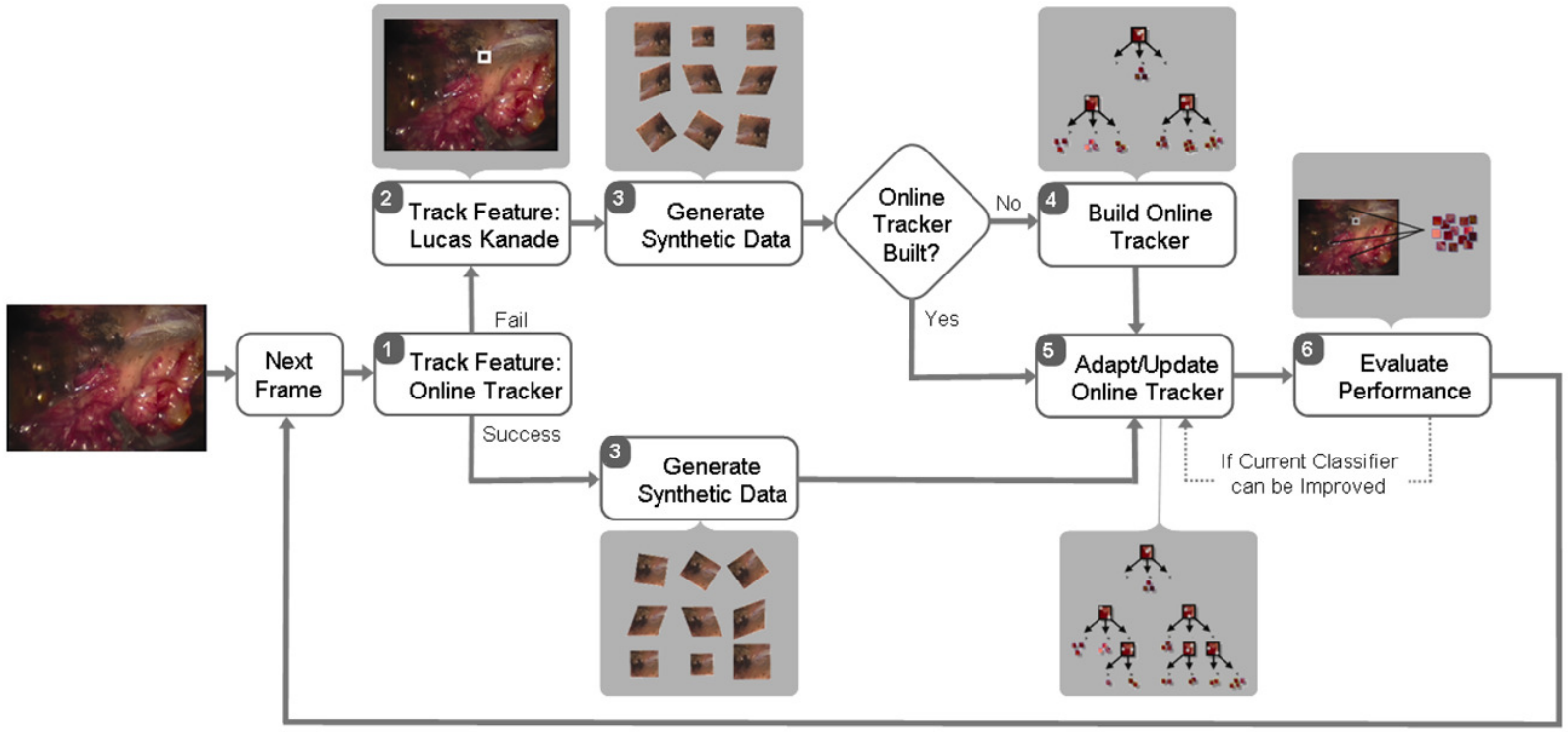


Figure 4.2 A diagrammatic overview of the proposed learning based online tracking system. The six steps of the system are shown and illustrate how the system learns online from real data, the generation of synthetic data, the construction of decision trees and the evaluation and update of the classifier.

4.2.1.1 Online Training Data Generation

This approach proposes to extract the training data for non-linear tissue deformation online while the regions are tracked and learnt from unlabeled data. It is shown in [89] that an LK tracker may be used to track regions on the deforming surface of the heart across cardiac cycles before eventually succumbing to drift. Regions are initially tracked on the surface of the tissue using an LK tracker thus creating a set of training data, which contains non-linear tissue deformation examples, and enables the learning of local deformation online. The set of false labelled data could be obtained by taking patches centred on every pixel in the image, which is not the region; however, this would create a set of training data which is computationally expensive to process. Instead, the set of false labelled data is obtained by randomly selecting patches and performing a local template matching to find similar regions. The true and false labelled data is added to a training set called S

4.2.1.2 Synthetic Training Data Generation

Synthetically generated data, used to model non-linear transformation, is particularly useful given current progress made in high-fidelity physical and appearance-based modelling. Simple projective image transformations include scale, rotation, skew, and perspective. For each image patch extracted in the online learning phase, four warp functions are individually applied to generate additional synthetic data. It should be noted here that the tracker will be updated online as the tracking process progresses. It is not necessary to apply an exhaustive set of transformations, rather, only a subset of transformations that are temporally persistent are applied. Rotational transformations used in this study range from -45° to 45° at 2° intervals. The scale transformations are applied from a factor of 0.8 to 3, thus enabling the tracker to handle similar transformations to [124]. The transformed patches are added to the training set S with a true label. Since synthetic data is used to build the classifier, note that initialising with an LK tracker is not required under these transformations.

For appearance based modelling, synthetically generated images with diathermy-induced smoke are used. The online learnt tracker is robust to occlusion; however, translucent smoke causes the visual appearance of tissue to change and region tracking to fail. Accurate modelling would require the detection of the smoke source location in 3D, estimation of the smoke density, and a 3D model of the peritoneal cavity with knowledge

of the input and output of carbon dioxide in the cavity. The use of a practical smoke model is proposed to model the effect of surgical smoke on the visual appearance of tissue. The model comprises three variables; colour, density, and 2D spatial distribution of smoke. These variables are combined in **Equation (4.1)-(4.3)** with the original pixel values from the image, in order to synthesise the visual appearance of smoke:

$$P_{r'} = s * P_r + (1-s)C_r \quad \textbf{(4.1)}$$

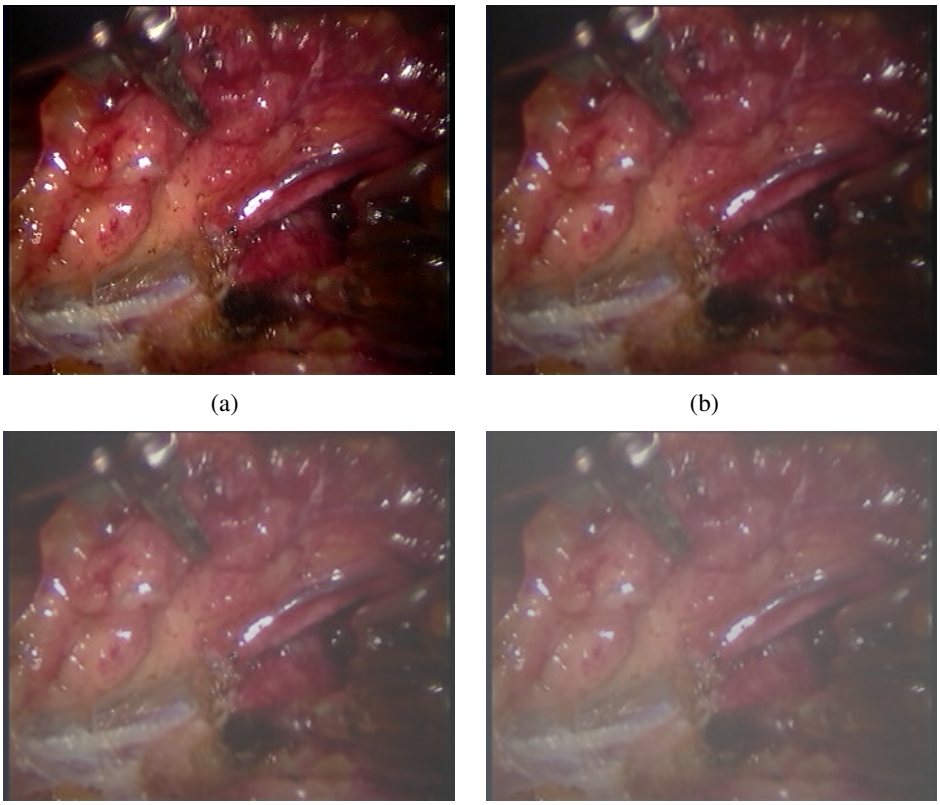
$$P_{g'} = s * P_g + (1-s)C_g \quad \textbf{(4.2)}$$

$$P_{b'} = s * P_b + (1-s)C_b \quad \textbf{(4.3)}$$

where P_r is the original red component of the pixel, C_r is the colour of the smoke (a random variable of Gaussian distribution with mean of 0.6 and standard deviation of 0.1), s is a random variable representing smoke density, and $P_{r'}$ is the transformed red component of the pixel. In this study, three different smoke density distributions are used with means 0.15, 0.25, and 0.4 with standard deviations of 0.05. These values are chosen because they represent translucent smoke. Values below 0.1 have little effect on visual appearance while those above 0.5 may result in occlusion. A Gaussian filter is applied to the resulting images to create smooth spatial distribution as illustrated in **Figure 4.3**. A training data with true and false labels can subsequently be generated and added to S.

4.2.2 Training the Classifier

Training the classifier for region tracking is equivalent to determining what information needs be encoded. The training process will learn the most discriminate information for classifying both true and false matches correctly. Given the labelled set of training data S, the classifier is trained to partition S into two sets; S_t and S_f, representing true and false matches. It has been shown that decision trees [113] can be used to effectively partition image patches into sets. Within this research, an ID3 [215] decision tree is used to iteratively partition S. A test is selected at each step in the tree generation, which partitions the set and creates a junction in the tree. The selected test is the one which best partitions the set S and determines what information is valuable for encoding.



(a) (b)

(c) (d)

Figure 4.3. The visual effect of smoke modelling based on **Equation (4.1)-(4.3)**. **(a)** Original image, **(b)** $s = 0.15$, **(c)** $s = 0.25$ and **(d)** $s = 0.4$ where s is variable representing the modelled smoke density.

This technique is different from many tracking techniques where either *ad hoc* decisions are made regarding what information is encoded or all information is treated equally. Each test examines a pair of pixels for every patch in set S . The test identifies if the first pixel is greater, similar or less in value than the second pixel and puts the patch into one of three subsets according to the result. A selection criterion function is used to identify the test or pair of pixels which best partitions the data. The selection criterion function should provide the maximum information allowing the entropy of each subset to be measured such that;

$$H(S) = \left|S\right| \log_2 \left|S\right| - \left|S_t\right| \log_2 \left|S_t\right| - \left|S_f\right| \log_2 \left|S_f\right| \qquad \textbf{(4.4)}$$

The optimal selected test is the subset where entropy is zero or the test with minimum entropy. The stopping criterion for tree building is zero or repeated minimum entropy.

The optimal test can be found by using an exhaustive search strategy: an approach that can be computationally prohibitive for large data-sets. Given a patch of size $j = x * y$, and a set of patches S of size k, an exhaustive search requires up to $(j-1) * j / 2 * k$ operations. This performance can be improved by sub-sampling j and k, however, this can deteriorate performance. Instead, a search step is introduced without the $j / 2$ component, which identifies the distribution of pixel at individual locations. This step searches for individual pixels in the set S, which yield a good separation between the sets S_t and S_f. These pixels are more likely to result in tests that provide optimal partition. A selection criterion is required to identify intra- and inter-class variance. Linear discriminant analysis may be used if the distribution of the sets is uni-modal, as illustrated in **Figure 4.4 (a)**, however, as shown in **Figure 4.4 (c)**, the distribution can be multimodal. It is proven in [216] that the log likelihood ratio is well-suited to the evaluation of multimodal distributions. At each pixel location, two histograms are created, $t(x,y)$ and $f(x,y)$, which correspond to S_t and S_f. The log likelihood is computed

$$L(x,y) = \log \frac{\max(t(x,y),\delta)}{\max(f(x,y),\delta)} \tag{4.5}$$

where δ is set to be 0.001 avoiding dividing by zero. The variance ratio of the log likelihood is used to measure the intra- and inter-class variance, *i.e.*

$$V(L;x,y) = \frac{\mathrm{var}(L;(t+f)/2)}{[\mathrm{var}(L;t) + \mathrm{var}(L;f)]} \tag{4.6}$$

where

$$\mathrm{var}\left(L;(a)\right) = \sum_i a(i)L^2(i) - \left(\sum_i a(i)L(i)\right)^2 \tag{4.7}$$

given the discrete probability density function a_i. This selection criterion function rewards low intra-class variance and high inter-class variance by identifying pixels with good separability. Using pixels with low intra-class variance and high inter-class variance, **Equation (4.5)**, is used to identify the optimal test.

4.2.3 Region Matching

In this thesis, two strategies have been considered for identifying the position of the region in a new frame: 1) detect and match - a region detection step extracts regions of interest for evaluation, and 2) exhaustive search - a region of the image is exhaustively searched by evaluating an image patch around each pixel. It is computationally efficient to evaluate regions and, therefore, an exhaustive search approach is appropriate. An exhaustive search method is also more conducive to continuous tracking as it is not susceptible to the repeatability of the region detection method.

To identify the location of a region in a new image frame, a patch centred on each point in a search region is classified using the decision tree. The search area of a fixed size was used with a Gaussian kernel weighting centred on the previous known position. This is an effective approach, however, in order to ensure the new position is within the search area, an oversized area was chosen.

The classification of patches in the search region can be performed quickly as the tests are simple and the false matches can be readily identified with only a few tests. It is common for more than one true match to be identified by the classification process. Matches are usually clustered within a few pixels of the true position of the region. This is due to the training data containing examples of the patch undergoing image transformations. The region is localised by examining the probability distribution, $P(N_j) = \left|S_{tj}\right|\left|S_t\right|^{-1}$, at the tree node, N_j ,to determine if it is a correct match, where $\left|S_{tj}\right|$ is the number of true matches classified by node j, and $\left|S_t\right|$ is the number of true matches classified by the entire tree. The best candidate point, $p_{x,y}$, in the search area is selected.

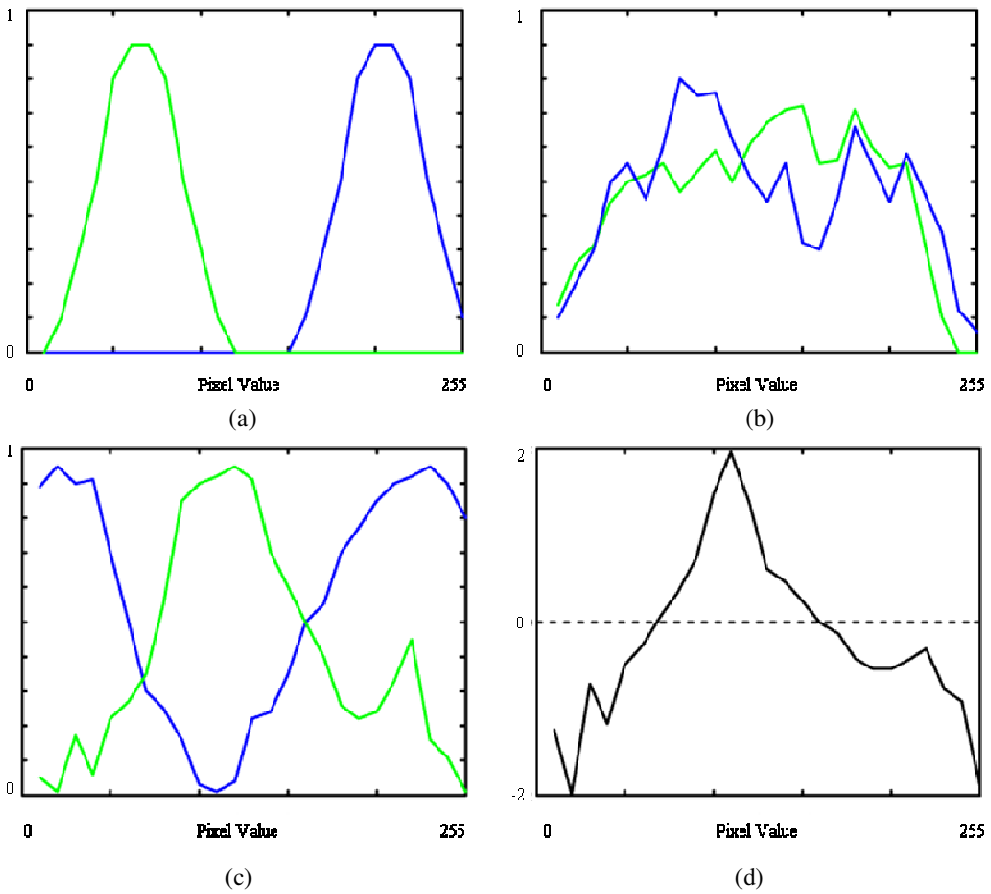


Figure 4.4 Hypothetical example distributions of training data-sets S_t (green) and S_f (blue) used to create the classifier. **(a)** Uni-modal distribution with low intra-class variance and high inter-class variance, **(b)** distributions with high intra-class variance and high inter-class variance, **(c)** multimodal distributions with low inter-class variance, **(d)** log likelihood ratio of multimodal distribution *(c)*.

4.2.4 Evaluating and Improving Online Tracking Performance

In this work, the decision tree is built incrementally online in order to optimise performance and reduce build time. Learning the decision tree can be computationally intensive if compared to testing the classifier, which is relatively fast. In order to exploit the speed of testing, a small set of training data is initially generated, S, as described above. S contains synthetic data and examples obtained online. It is also a subset of all the data available, $S_{complete}$. $S_{complete}$ is a set of patches from each point in each image and all possible synthetic image transformations. An initial decision tree is built with the

small data-set S. This classifier is tested on the current image from the laparoscope. The classifier is likely to perform poorly on the data as it has been trained on a small subset. This can lead to a high number of false positives indicating that the classifier may be improved. Patches around the false positive points detected in the image are added to the set S and the classifier is retrained. The entire tree does not require retraining; only the patches in the node N_j are responsible for incorrect classification. N_j will be retrained, based on the new distribution of the data, to obtain a more informative test. By retraining the tree, it is possible to over-fit the classifier. The metric indicating over-fitting is the number of false negatives. If false negatives are observed, the nodes of the tree are retrained.

The final adaptive step in the update is the selection of the most discriminative colour space for tracking. This follows the criterion set out in [216] where forty-nine colour spaces are searched in order to identify the most discriminative. This is a linear combination of Red, Green and Blue (RGB) defined as

$$F_1 = \{w_1 R, w_2 G, w_3 B \mid w_* \in [-2, -1, 0, 1, 2]\} \quad \textbf{(4.8)}$$

thus creating a set of colour spaces including RGB, intensity, approximate chrominance, and excess colour. The most discriminative colour space is identified using the variance ratio outlined in **Equation (4.6**).

4.3 Modelling Tissue Motion

The 3D position of the region is recovered by using stereo geometry. A region is detected in the left image while the epipolar line in the right image is searched in pursuit of a correspondence using the region tracking approach described above. The centre of the camera rig is the left camera and the origin of the world is taken as the centre of the camera rig on the first frame. This approach requires camera calibration, which is performed using a closed form solution [74] using the assumption of a pinhole camera model. The stereo laparoscope is calibrated before the procedure and remains unchanged. The baseline between cameras is approximately 5mm.

4.3.1 Extracting Intrinsic Global Tissue Motion

Tissue motion can be a result of respiratory motion, cardiac motion, and tissue-tool interaction. Tissue-tool interaction is difficult to predict; however, respiratory and cardiac motion is generally periodic or quasi-periodic. This intrinsic periodic motion can be used to predict and track the tissue surface.

The respiration cycle is a periodic 1D signal that causes a change in the 3D position of the tissue. It is the change in 3D position that is observed by tracking the tissue and subsequently, the 1D signal embedded in a 3D space. This signal can be extracted by transforming the data to a new coordinate system aligned to the largest variance in the data. This can be estimated using Principal Component Analysis (PCA), which provides an orthogonal linear transformation of data $m^T = (x, y, z)$ to a new coordinate system Y^T such that the first coordinate holds the largest variance in the data, or the principal component of the data. This transformation is defined as $Y^T = m^T W$. The principal axis of the motion of the liver is in the superior-inferior direction [132] and is described in the principal component, or the first coordinate of Y^T.

Extracting tissue motion caused by both the cardiac and respiration cycles is more complex. It is proven in [89, 137] that the motion of the heart is a coupled result of cardiac and respiratory motions. Such tissue motions are extracted by performing Independent Component Analysis (ICA) - a statistical technique for separating signals into additive subcomponents while maximising mutual statistical independence. ICA can be formulated to consider the recovered 3D motion of the surface of the tissue to be the latent variables $m = (x, y, z)$ and the components of intrinsic motion as $s = (h, r)$. It attempts to find the transformation W such that $s = Wm + n$ where n is zero mean Gaussian noise. The components of m can be written as the weighted sum of the independent components, *i.e.*, $m = \sum a_k s_k$, where a_k is a vector of mixing weights which make up the mixing matrix $A = (a_1 \ldots a_n)$, where $W = A^{-1}$. The source s and the mixing matrix A are estimated adaptively with cost function $s_k = w^T m$ to maximise non-Gaussianality.

4.3.2 Tissue Motion Models

Modelling the global motion of tissue resulting from respiration or the cardiac cycle can be used for motion prediction and compensation. For example, [129] identifies that motion of the liver is correlated to the periodic motion of the diaphragm. It should be noted here that there are two types of periodic respiration resulting from free breathing and assisted breathing, with the use of a ventilator. Assisted breathing is standard practice for MIS procedures and regulates the frequency of breathing cycles. Although the breathing cycle is periodic, it is asymmetric: more time is spent in the exhale position. Lujan proposes an asymmetric model in [132]

$$z(t) = z_0 - b\cos^{2n}\left(\frac{\pi t}{\tau} - \phi\right) \tag{4.9}$$

where z_0 is the position of the liver at the exhale, b is the amplitude, τ is the breathing cycle period, ϕ is the phase and n describes the shape or gradient of the model.

The motion of the heart can be described as quasi-periodic. In [217], the authors show that this quasi-periodic motion can be modelled using a Fourier series such that

$$y(t) = c + \sum_{i=1}^{m} r_i \sin(iwt + \phi_i) \tag{4.10}$$

where w is the cardiac period, ϕ is the harmonic phase, r_i is the harmonic amplitude and c is the DC offset. The Levenberg-Marquardt (LM) algorithm is used to estimate the model parameters. The LM algorithm is a non-linear, least squares minimisation algorithm which interpolates between the Gauss-Newton algorithm and gradient descent to optimise a set of parameters: β of the model $f(x_i, \beta)$ to minimise the square of the deviations such that

$$S(\beta) = \sum_{i=1}^{m} [y_i - f(x_i, \beta)]^2 \tag{4.11}$$

The methods described above not only extract the temporal respiratory and cardiac cycles, but they also provide spatial information about the global position of tissue in the MIS environment at any point in the cycle.

4.4 Experiments and Results

The performance of the proposed approach under scale change, orientation change, surgical smoke, occlusion, and deformation resulting from the respiratory and cardiac cycles has been evaluated. It is assessed on simulated and *in vivo* MIS data-sets of the liver, heart, and abdomen. The approach is compared to four conventional tracking techniques SIFT [124], LK [214] (with template update), and two mean-shift algorithms [216]. A brief description of these methods and their implementation is provided in the following section and more details are provided in **Chapter 2**.

SIFT is a method for wide baseline feature matching. It can be used for tracking in a tracking-by-detection framework where features are redetected at each frame and no temporal information is used. SIFT detects scale invariant Different of Gaussian (DOG) regions and encodes greyscale patches around the region as a histogram of oriented gradients. It is a tracking-by-detection approach using a nearest neighbour ratio matching, 0.6, and no temporal information between frames. This makes it well-suited to dealing with occlusion, however, it requires the region to be detected in each frame and for said region to be globally unique in the image. In rigid scenes, matching can be improved by using global matching constraints such as Random Sampling Consensus (RANSAC), however, these techniques cannot be easily applied in unknown, non-rigid environments.

LK is a pyramid optical flow method for region tracking. LK encodes greyscale spatial information in a template and iteratively attempts to minimise the difference between the template and the observed data. LK is based on three key assumptions [116]: 1) brightness constancy, 2) temporal persistence, or small, differential changes between frames, and 3) spatial coherence, or the assumption that pixels belong to the same surface and follow the same motion. LK tracking can suffer from the aperture problem and region drift due to template updating required for the temporal persistence criterion.

Mean-shift is a non-parametric, statistically robust method for locating local maxima in a probability density distribution. In this application, a DOG region is detected in the first frame and mean-shift is used to track in subsequent frames. Mean-shift encodes the colour values of the pixels in the patch around the region as a histogram and makes no use of structural information, thus making it capable of tracking deformation. Mean-shift

is well-suited to large self-contained blobs and is less suited to lines or corner-like structures. Unlike SIFT, regions only have to be locally unique since temporal information is used in tracking. It is assumed that the displacement of a region is small, and a spatial overlap of the image patches exists between frames. As mean-shift relies only on colour and not structure, it is able to deal with partial occlusion of a region, however, the requirement for overlap means it is not able to deal with the recovery from full occlusion if the region moved during occlusion. This work uses two mean-shift approaches as outlined in [216]. These mean-shift trackers are more discriminative than the standard approach and attempt to find the optimal colour representation of the region. The first approach compares the pixels in the detected region to the pixels in the surrounding area and searches for the three most discriminate colour spaces. These colour spaces are subsequently used in the mean-shift tracker. The second approach incorporates an additional step to reduce the effect of distracters, or pixels, in the surrounding area, which may cause the mean-shift algorithm to converge on the wrong point.

To evaluate the performance of the tracking methods the evaluation criteria outlined in [114] and [204] are used. Two performance metrics are computed. The *sensitivity* (also known as recall) is computed as

$$sensitivity = \frac{\#\,correct\ matches}{\#\,correspondences} \tag{4.12}$$

where *correspondences* is the number of trackable regions existing in the current image. A good region tracking method will have high *sensitivity* as this is a measure of the density of regions. The second performance metric is (*1 − precision*). This is a measure of the total number of incorrectly tracked regions with respect to the total number of tracked regions such that

$$\Lambda = 1 - precision = \frac{\#\,incorrect\ matches}{\#\,correct\ matches + \#\,incorrect\ matches} \tag{4.13}$$

A good region tracker will have a low Λ value. These metrics are evaluated individually with respect to time. This provides an evaluation of the temporal persistency of the region trackers.

4.4.1 Simulated Experiments

To quantitatively evaluate the performance of the proposed method with respect to deformation, a simulation was created to generate synthetic data. An image of the heart was textured onto a 3D surface, as shown in **Figure 4.5 (a-d)**. The surface was periodically deformed with a mixture of Gaussians so as to simulate cardiac and respiratory tissue deformation as different frequencies. The mesh was deformed for 4000 frames and generated the ground truth position of 100 regions for quantitative analysis. Cardiac, respiratory motion, and noise are simulated, but not specular reflection and changing light conditions as light-tissue interaction is non-trivial to model.

In **Figure 4.5 (e)**, the trackers are quantitatively evaluated with respect to *sensitivity* over time. It can clearly be seen that there is oscillation in the performance of all the trackers. This oscillation corresponds to the periodic nature of the applied deformation. The reduction in performance occurs at the extremes of deformation. At this point, the regions have changed shape, and localising the centre of the region accurately may fail given that the region tracking methods are drawn towards prominent information in the patch (such as edges or corners which may not be at the centre of the patch).

This oscillation affect is particularly noticeable on SIFT. The change in shape of the region makes accurate repeatability of the DOG detector challenging and violates the geometric constraints imposed by the SIFT descriptor. As deformation of the surface increases, the number of regions, which may accurately be matched using a histogram of gradients, is reduced. This oscillation makes SIFT less attractive for continuous tracking in MIS confirming the results in **Chapter 4**. **Figure 4.5 (f)** shows SIFT has the lowest Λ. This is a result of the matching strategy. The nearest neighbour ratio test prevents a region from matching if it is visually similar to another detected region to prevent false matches. Tracking may be improved with more sophisticated matching incorporating temporal information or prior knowledge of the global scene structure.

The LK tracker performs well at the beginning of the experiment with high *sensitivity* and low values of Λ. The temporal persistence and spatial coherence assumptions are held in the simulated data. The noise added to the system violates the brightness constancy assumption. Combined with non-linear deformation, the noise causes error propagation when the template is updated: the tracker's performance degrades over time.

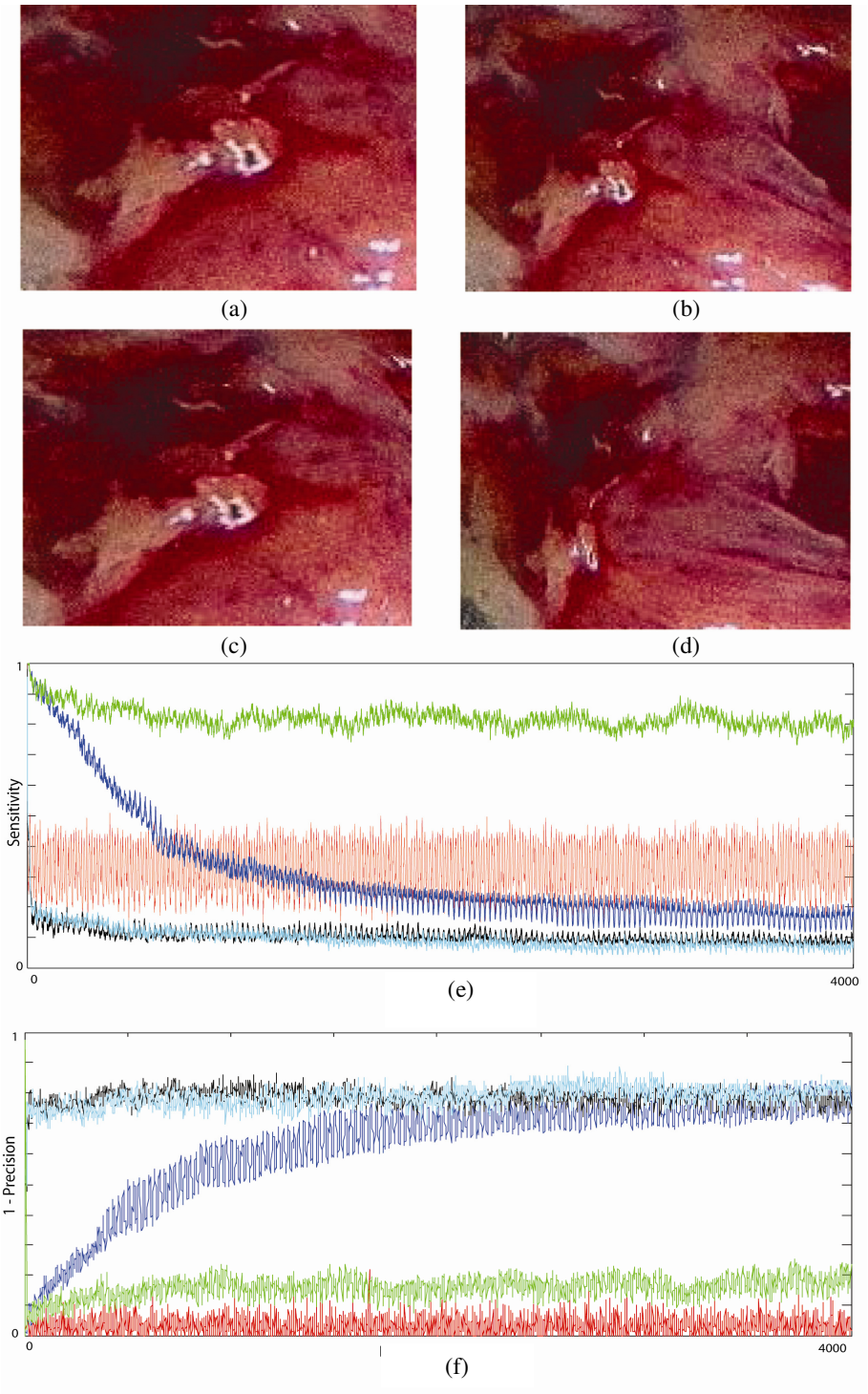


Figure 4.5 Quantitative tracking performance for simulated data with the five tracking algorithms considered. **(a-d)** The simulated data is created by warping an image taken from a MIS procedure with known ground truth deformation characteristics. **(e)** and **(f)** Quantitative performance evaluation for the five different tracking techniques compared; green – online learnt tracker, red – SIFT, dark blue – Lucas Kanade, black – mean-shift 1, and light blue – mean-shift 2.

The two mean-shift trackers perform similarly. The algorithms have low *sensitivity* and are only able to track a small number of regions accurately. This poor performance can be attributed to three factors: 1) for the majority of detected regions, colour information in MIS images is not sufficient to distinguish a region from its surroundings. This results in a high value of Λ and a high number of wrongly matched regions; 2) DOG detects edges, corners and blobs in greyscale images, however, mean-shift works well with large self-contained regions of distinct colour. If the region is not self-contained, the mean-shift will drift; and 3) mean-shift assumes a spatial overlap in a region's location between frames. If the overlap is too small, or does not exist, the pixels will fall outside the basin of attraction and lead to tracking failure.

The proposed tracker, with online learning, maintains a good performance in the presence of deformation with a derived *sensitivity* outperforming alternative approaches. The Λ is low and only outperformed by SIFT. This proves the proposed approach is capable of learning unique qualities in a region and encodes that information enabling the region to be successfully tracked in the presence of visually similar regions. The robustness to synthetic deformation can be attributed to learning from example data. By learning from example data generated by the LK tracker, the classifier is built on real image transformations, which will be subsequently observed due to the periodic nature of the deformation. Although the online learnt tracker performs well, there remain points that cannot be tracked, and there is fluctuation in performance. This is attributable to noise in the image or poor initialisation from the LK tracking.

4.4.2 *In vivo* Experiments

The performance of the proposed technique was quantitatively evaluated on *in vivo* data. 50 regions were detected in the first frame for each sequence. An expert user manually obtained ground truth data for each region in these sequences at 50 frame intervals.

4.4.2.1 Tissue Deformation

Figure 4.6 shows the three sequences used and the corresponding tracking results. **Table 4.1** provides additional results. **Figure 4.6 (a-f)** show sequences taken from two Totally Endoscopic Coronary Artery Bypass graft (TECAB) surgeries. Centred in the endoscopic

image is the epicardial surface deforming with cardiac and respiratory motion. Rapid tissue deformation and specular reflections make tracking challenging in these sequences. **Figure 4.6 (g-i)** show the sequence and tracking results for footage of the liver deforming due to respiratory motion. This sequence is challenging due to changes in illumination. The point light source causes large changes in illumination and alters the visual appearance of the tissue depending on the distance and orientation of the tissue to the light source.

The relative performance of the trackers is similar to the synthetic data. In the TECAB sequences as shown in **Figure 4.6 (a)** and **Figure 4.6 (d)**, the LK tracker performs well initially, however, as drift occurs, the performance degrades over time. Brightness constancy is violated here due to specular reflections and the change in visual appearance as a result of the point light source. The mean-shift tracker performance in the first sequence, shown in **Figure 4.6 (a)**, was inferior to that shown in **Figure 4.6 (d)**. This is due deformation, which is more pronounced in the first sequence leading to larger inter-frame motions. SIFT performs poorly in this sequence because larger deformation leads to larger changes in visual appearance. The SIFT descriptor is affected by the specular highlights which cause sharp gradients. Region density and persistency are higher using the online learning method as shown in **Figure 4.6 (b)** and **Figure 4.6 (e)**. The tracking deteriorates towards the end of the first sequence due to tissue-tool interaction. The robustness to drift of the online learnt tracking algorithm is compared to LK and illustrated in **Figure 4.8 (a)** as a 3D spatio-temporal plot.

Table 4.1 In vivo data. Summary of the tracking performance of five algorithms with respect to tissue deformation.

	LK	**MS1**	**MS2**	**SIFT**	**Learning**
Sensitivity	0.333	0.245	0.241	0.287	0.752
Λ	0.555	0.545	0.586	0.049	0.170

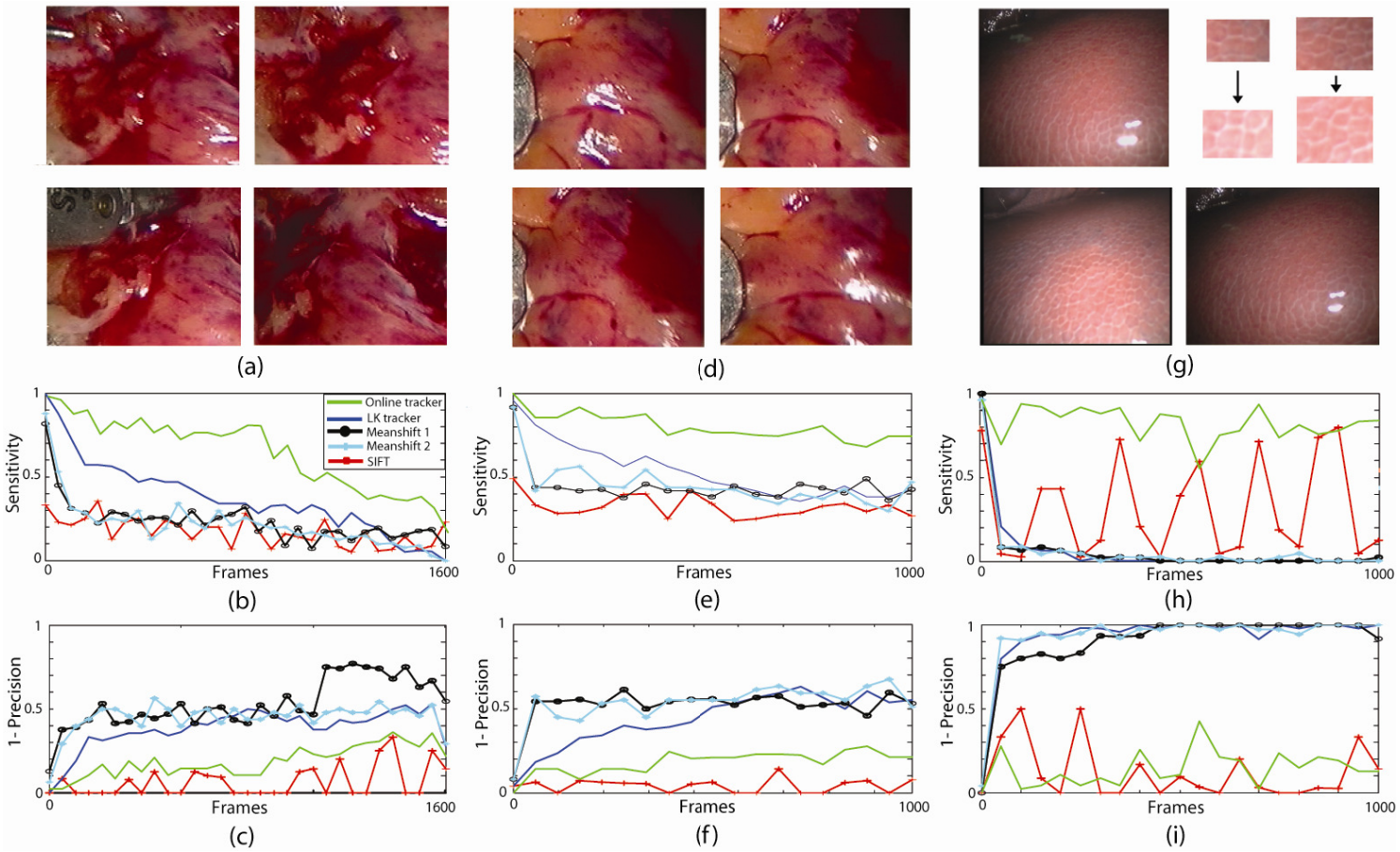


Figure 4.6 Quantitative tracking performance for *in vivo* deformation sequences. **(a-c)** *In vivo* cardiac data-set and tracking analysis. **(d-f)** Second *in vivo* cardiac data-set and tracking analysis, **(g-i)** Porcine liver data-set and tracking analysis. Five trackers are compared; green – the online learnt tracker, red – SIFT, dark blue – Lucas Kanade, black – mean-shift 1, and light blue – mean-shift 2.

Figure 4.6 (g) shows footage of the liver deforming as a result of respiration. Deformation is less pronounced here than on the cardiac surface, however, illumination of the surface changes significantly as the organ moves toward, and away, from the laparoscope and light source. The LK tracker performs poorly in this footage. This is attributed to changes in scale and illumination, which violate the brightness constancy assumption. The cyclical performance of SIFT is seen again in this sequence. The DOG detector has relatively low repeatability on this sequence because the structures on the liver are small. These structures disappear when the Gaussian kernel is applied, thus reducing the number of detected regions. The mean-shift algorithms do not perform well due to the uniform colour distribution of the liver. The online learnt tracker performs well on this sequence not only because it is learning the most discriminative information but also because the encoding of the information is robust to changes in illumination. The encoded information is the relative value of pixels, which makes the learning method well-suited to handling changes in illumination because these relative values will remain consistent under linear changes in illumination.

4.4.2.2 Occlusion

The use of tools in surgery leads to occlusion of the operative field and full or partial occlusion of regions on the surface of tissue. **Figure 4.7 (a,d,g)** show deformation of the liver and abdomen wall resulting from respiration. The surgeon introduces tools in these sequences, which leads to occlusion of the surgical field-of-view. Quantitative analysis is provided in **Figure 4.7** and **Table 4.2.** In **Figure 4.7 (a)**, only a small number of regions are occluded starting at frame 200. At this point, the decreased performance of the LK and mean-shift trackers is clear. These trackers require temporal information and have no explicit mechanism for dealing with recovery from full occlusion. SIFT and the online learning tracker remains robust in the presence of occlusion. In **Figure 4.7 (d)**, the number of occluded regions increases. The trackers' performances are similar to that of the first sequence. The online learning tracker outperforms SIFT on these sequences, however, both have a low Λ values, thus demonstrating the correct identification of occluded regions.

In the last sequence, **Figure 4.7 (g)**, the surgeon uses an irrigation tool. This tool occludes the majority of the regions after 300 frames. At this point, there is a significant reduction in the performance of LK and mean-shift (almost zero *sensitivity*). At around frame 700, the suction tool interacts with tissue to remove blood from the surface. This leads to a reduction in *sensitivity* of the online learning tracker and causes an increase in the value of Λ. This is attributed to the change in visual appearance of the scene due to the removal of blood. This scene is different to the data used to train the classifier. It is proven that the proposed approach to tracking deals well with tracking failure recovery resulting from full regional occlusion. This capability is further illustrated in the 3D spatio-temporal plot, **Figure 4.8 (b)**, where the online tracker, shown in green, is capable of recovering from occlusion and continuous tracking, unlike SIFT, which is not continuous.

Table 4.2 In vivo data. Summary of the tracking performance of five algorithms with respect to occlusion.

	LK	**MS1**	**MS2**	**SIFT**	**Learning**
Sensitivity	0.557	0.307	0.322	0.458	0.728
Λ	0.356	0.500	0.494	0.017	0.144

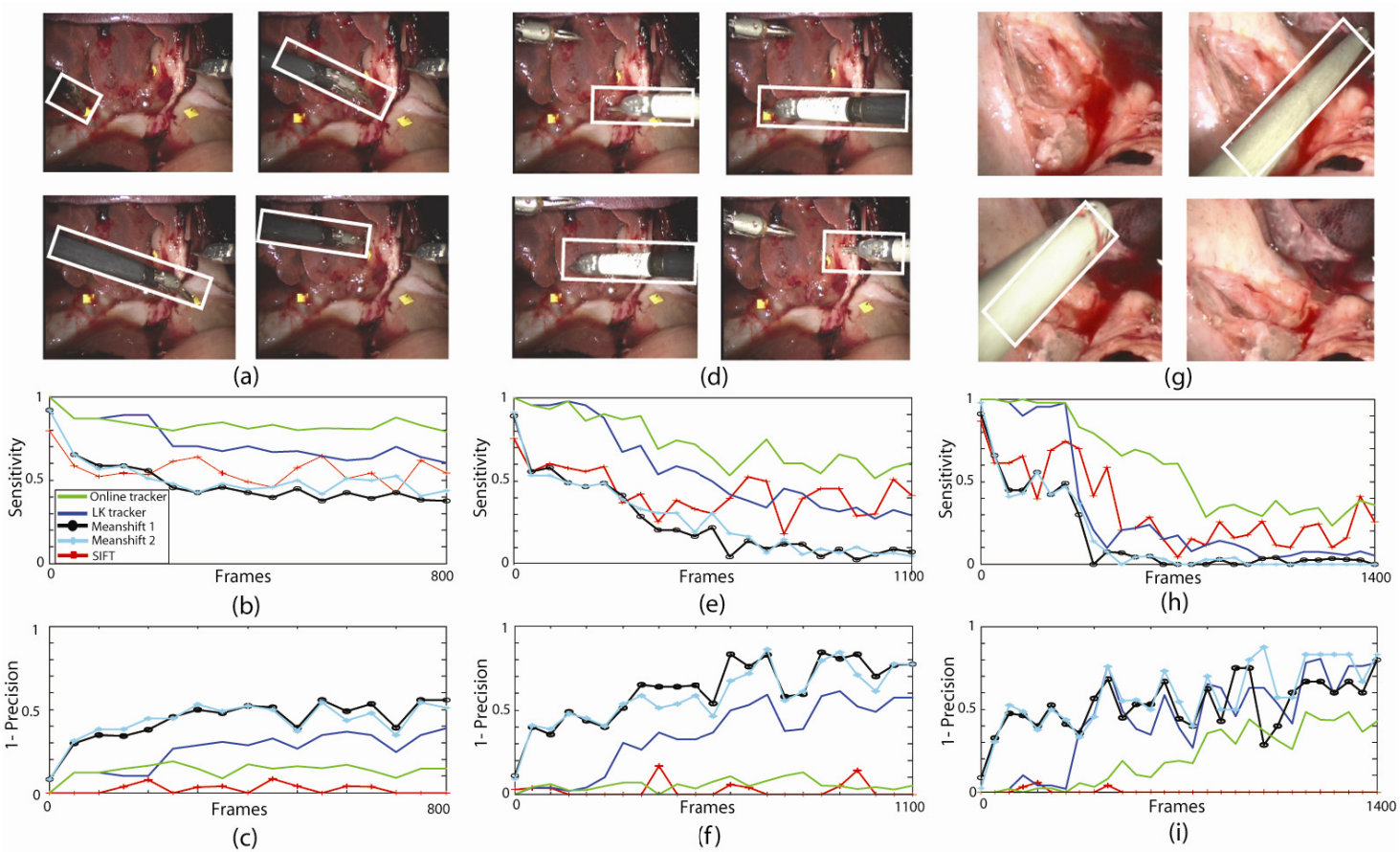
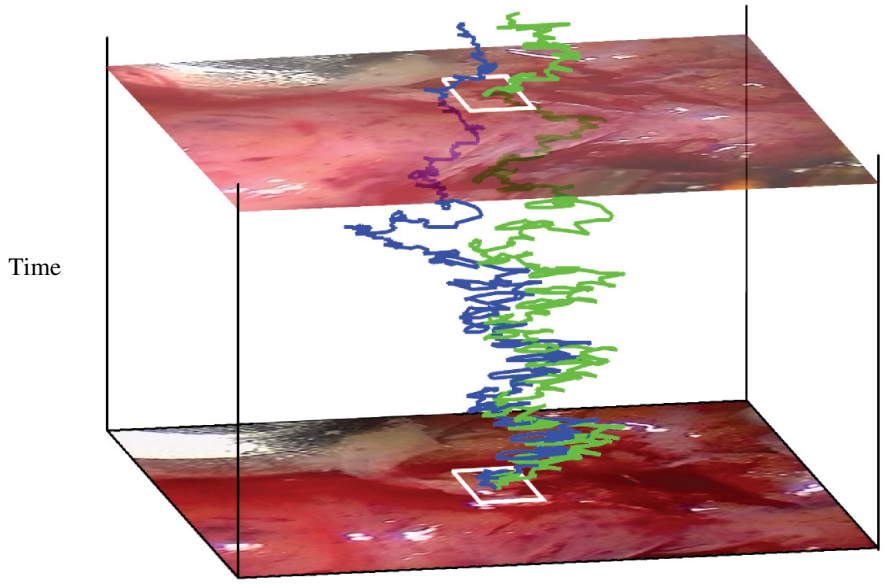
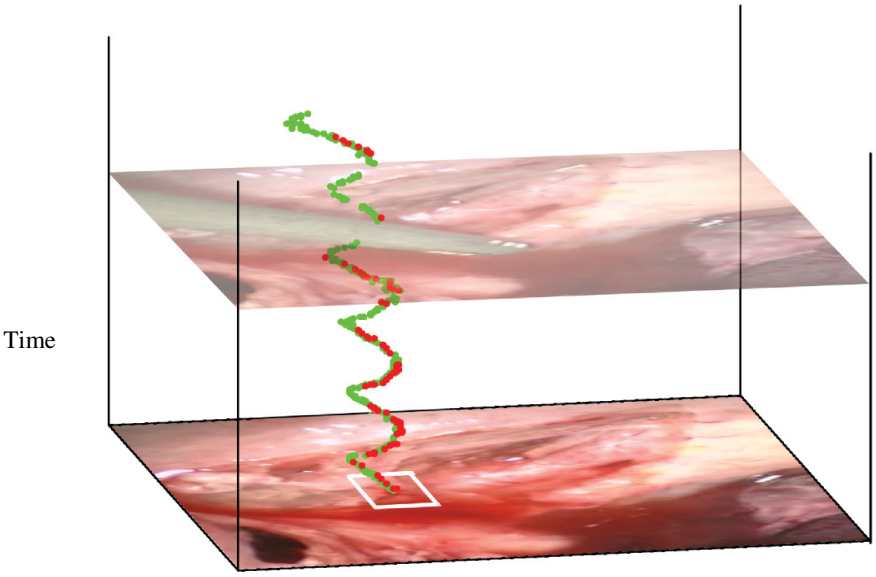


Figure 4.7 Quantitative tracking performance for *in vivo* occlusion sequences. **(a-c)** Occlusion sequence one with tracking analysis. **(d-f)** Occlusion sequence two with tracking analysis. **(g-i)** Occlusion sequence three with tracking analysis. Five trackers are compared; green – the online learnt tracker, red – SIFT, dark blue – Lucas Kanade, black – mean-shift 1, and light blue – mean-shift 2.



(a)



(b)

Figure 4.8 (a) A single region tracked over time showing drift with LK tracking in blue and the robustness of proposed approach in green. **(b)** Illustrates the problem of occlusion by a tool. Green – the proposed online learnt tracker, red – SIFT. SIFT tracking is not continuous.

4.4.2.3 Scale and Rotation

During MIS, the surgeon frequently manipulates the laparoscope for the purpose of navigation. This tendency leads to changes in scale and orientation of the images. The rotation of a laparoscope is limited by the fulcrum effect [218] and, consequently, rotation is usually observed around the optical axis, **Figure 4.9 (a)**. In **Figure 4.9 (b-c)**, the effect of rotation is quantitatively evaluated. First, the laparoscope is rotated approximately 50° anticlockwise. It is then returned to its original position and rotated 120° clockwise. The LK tracker and mean-shift trackers do not explicitly encode information about rotation, however, the LK track is able to perform well in this rotating sequence due to the template update. The inter-frame motion of the camera is small, which means the temporal persistence assumption is maintained. The mean-shift trackers do not encode spatial information but they are theoretically invariant to rotation: the poor performance here is attributable to the lack of distinct colour variation and poorly defined regions. SIFT, which models orientation, is shown to be robust to rotation, and density is higher with the online learnt tracker.

In **Figure 4.9 (g)**, the laparoscope is moved along the optical axis. This causes a scale change of approximately 2.8. The laparoscope is moved away from the wall, back to its original starting position, and then further away. This causes a scale change of approximately 0.66. As a result of this significant scale change, only 18 regions detected in the first frame are visible throughout the entire sequence. These regions are used to evaluate overall performance. Quantitative results are shown in **Figure 4.9 (h-i)** and **Table 4.3**. The LK tracker does not explicitly incorporate scale information, but it is able to track the regions due to the template update and temporal persistence assumption. The *sensitivity* reduces towards the end of the sequence due to error propagation and drift. The mean-shift trackers are generally not invariant to scale changes, although the approach has been extended in [121] for scale. The high value of Λ demonstrates, once again, how colour alone is insufficient for dense region tracking. In the SIFT detection phase, multi-scale DOG regions are extracted, which makes it theoretically robust to scale changes. The *sensitivity* of SIFT decreases as scale changes, and the number of incorrectly matched regions increases significantly when the scale is reduced to 0.66. The proposed online learning algorithm explicitly incorporates scale into the learning process, thus enabling it to remain robust to changes in scale and to out-perform conventional trackers.

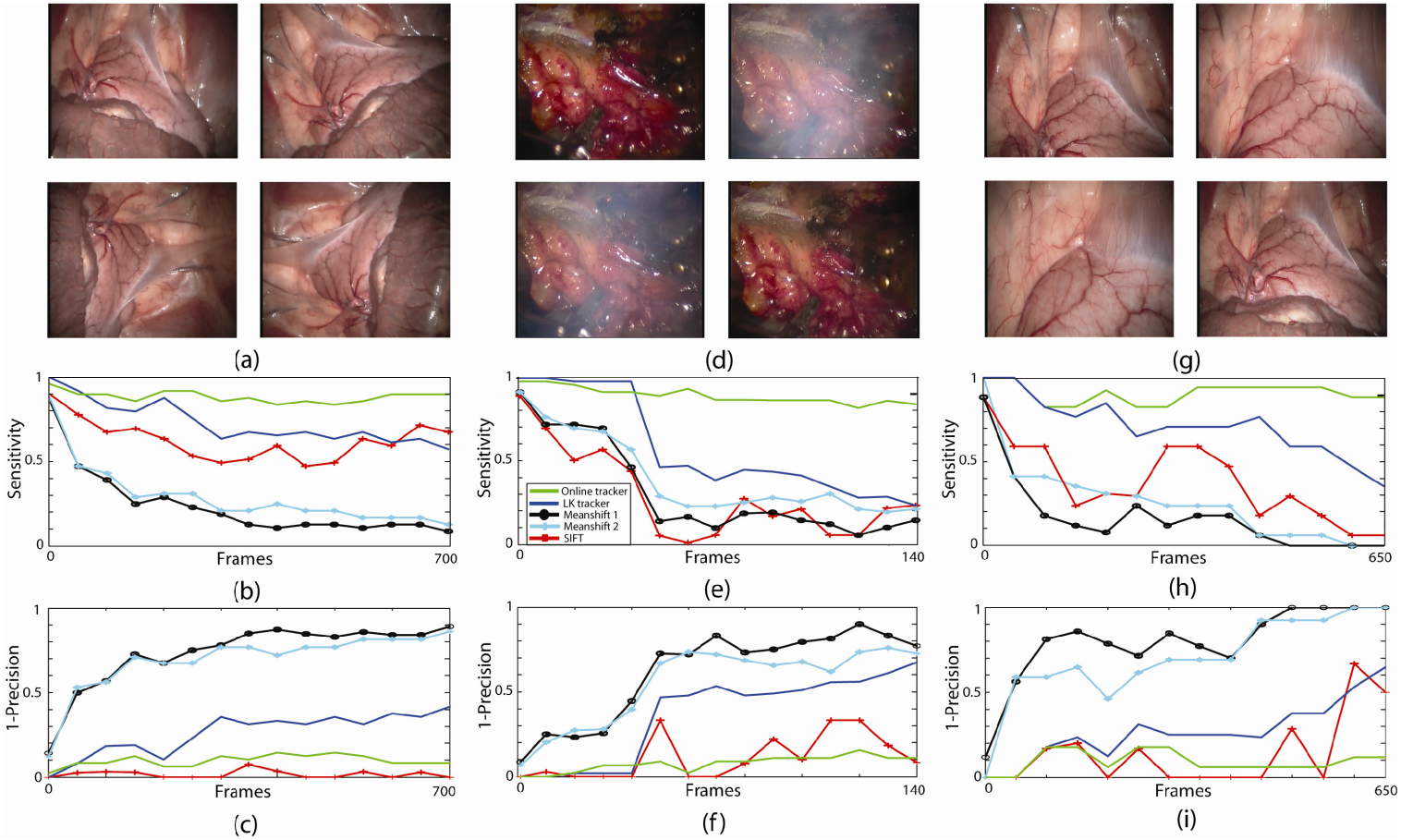


Figure 4.9 Quantitative tracking performance for *in vivo* sequences. **(a-c)** Rotation around the optical axis with tracking analysis. **(d-f)** Surgical smoke resulting from diathermy with tracking analysis. **(g-i)** Scale change with tracking analysis. Five trackers are compared; green – the online

4.4.2.4 Surgical Smoke

Figure 4.9 (d-f) and **Table 4.3** document investigations into the effect of smoke resulting from diathermy. In this sequence, the diathermy is activated twice resulting in the smoke shown in **Figure 4.9 (d)**. The build up of smoke takes place gradually, over a number of frames, and remains visible for a short period before a suction device is used to remove it. Translucent smoke has the effect of greying the pixels, reducing colour distribution, flattening gradients, and making the image appear increasingly homogenous.

The presence of smoke affects the SIFT tracker as the histogram of gradients will be changed, thus complicating the matching process. As demonstrated in **Figure 4.9 (e)**, when the diathermy is activated, causing smoke, the *sensitivity* for SIFT drops to zero. The gradual flattening of gradients during the appearance of smoke means the LK tracker has less structure on which to converge. This results in template drift because the smoke is incorporated into the template. In this sequence, the mean-shift tracker initially performs well and with high *sensitivity*. This is attributed to a number of self-contained blobs of distinctive colour on the surface of the tissue; however, greying of the pixels caused by surgical smoke, changes the colour distribution of the region patches. The model of colour distribution for the mean-shift trackers is no longer valid as a result. This issue could be resolved by incorporating a smoke model into the mean-shift algorithm where the colour distribution model is learnt. The online learnt tracker performs well because the greying of the pixels and flattening of gradients has been simulated and incorporated into the training data. The proposed approach has, therefore, learnt what structures in the scene remain prominent whilst obscured by smoke. The online learnt tracker remains robust to smoke, however, there is a slight drop in performance, which is attributed to the smoke modelling that cannot capture the non-uniform, spatial distribution of smoke. It should be noted that the performance of the online learnt tracking is dependent on the quality of smoke appearance modelling.

Table 4.3 *In vivo* data. Summary of the tracking performance of five algorithms with respect to scale, rotation and surgical smoke.

	LK	**MS1**	**MS2**	**SIFT**	**Learning**
Sensitivity	0.665	0.263	0.332	0.453	0.892
Λ	0.278	0.606	0.571	0.032	0.076

4.4.3 *In Vivo* Tissue Motion Modelling

To demonstrate the potential clinical application of the online learning approach for deformation tracking, the intrinsic global motion of the tissue is extracted from two *in vivo* sequences. The first sequence, shown in **Figure 4.6 (d)**, contains motion resulting from cardiac and respiratory cycles. The recovered 3D motion is decomposed into its independent components of cardiac and respiratory motion using ICA. The results are shown in **Figure 4.10 (a-b)**. The first and second components have two, distinct frequencies. The extracted components contain noise, which is a combination of small tracking inaccuracies, and the small baseline of the stereo camera (5 mm), which causes errors in 3D depth recovery. As a result, the cardiac and respiratory motions are not flawlessly isolated, however, models can be fitted to the separated, noisy signals using the LM algorithm, as shown in **Figure 4.10 (a-b)**, and the residual error, shown in **Figure 4.10 (d-e)**.

Example frames taken from the second sequence are shown in **Figure 4.6 (g)**. The sequence shows liver motion resulting from respiration. PCA is used to identify the principal axis of motion along which the region on the surface of the liver moves. It is along this axis, or principal component, that **Equation (4.9)** is used to fit the respiration model. The results are shown in **Figure 4.10 (c)** and **Figure 4.10 (f)**. It is evident that the model is well-fitted to the data. The extracted global motion of the tissue, along with the ICA mixing matrix and PCA transformation matrix, can be used to model the motion of the surface of the tissue. These motion models can be used to predict the future position of the regions in 3D.

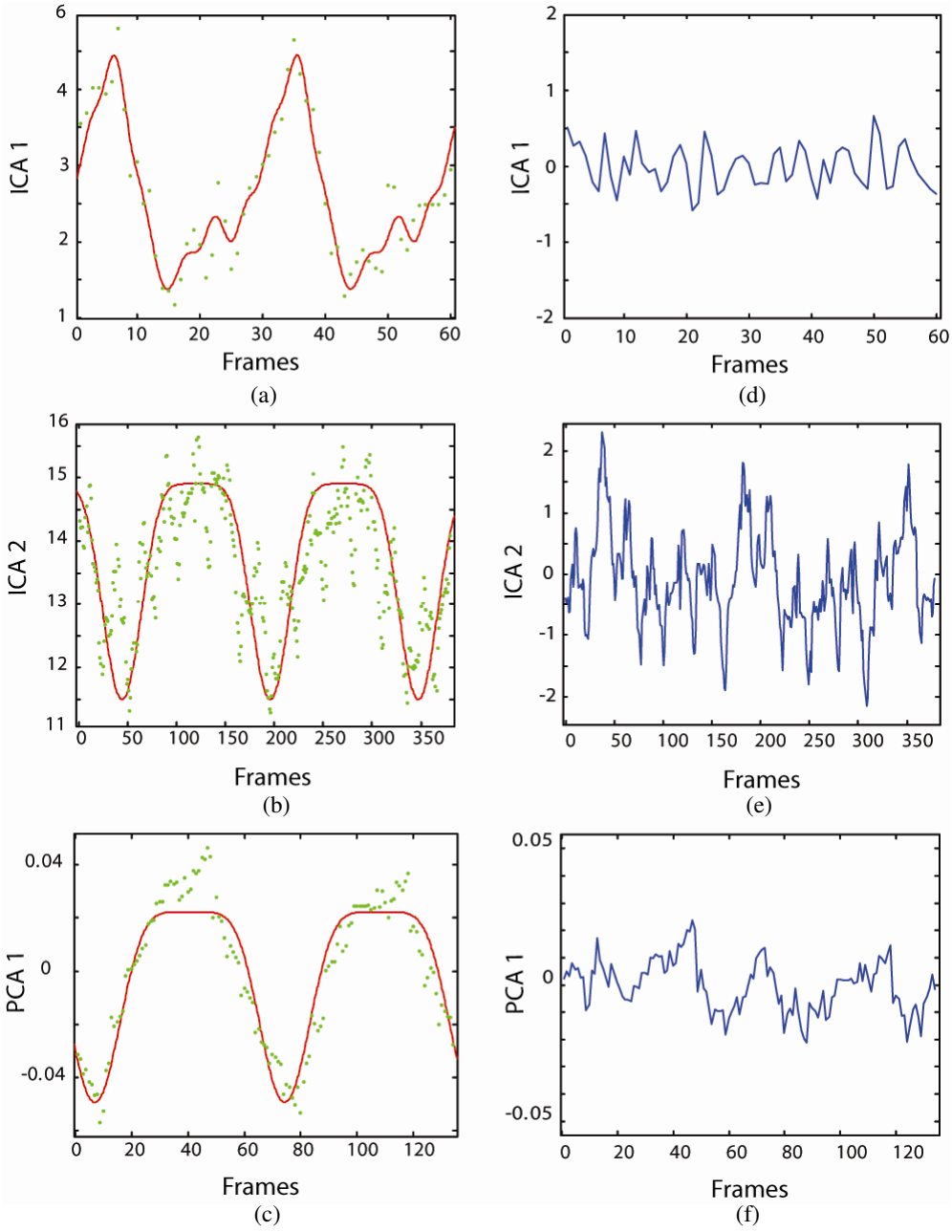
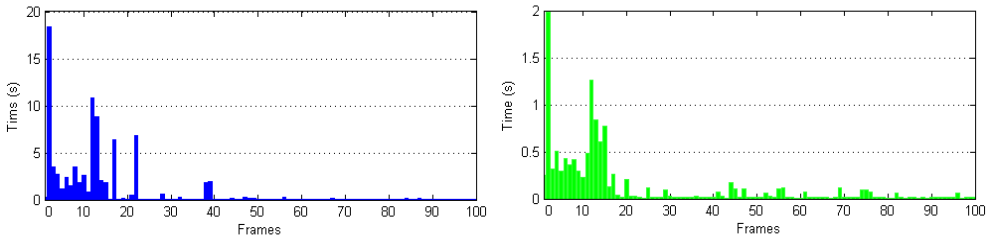


Figure 4.10 Modelling tissue deformation. **(a-c)** The extracted components from global motion (green) and models (red) and **(d-f)** corresponding error plots (blue). **(a)** The first ICA component extracted from footage of the heart representing the cardiac motion. **(b)** The second ICA component extracted from footage of the heart representing the respiratory motion. **(c) T**he first PCA component extracted from footage of the liver representing the respiratory motion.

4.5 Computational Performance Analysis

The proposed framework was implemented on a desktop PC with an Intel Pentium 3 GHz processor and 2GB of RAM. The software was written in C++ on a Microsoft Windows environment. The most computationally demanding component of the framework is learning the feature representation. In **Figure 11**, the average computational requirements for learning a single feature is shown. The figure compares computation timings for exhaustive search and the optimised search in **Figure 4.11 (a)** and **Figure 4.11 (b)** respectively. The optimised search is approximately 10 times faster.

The graphs in **Figure 4.11** share a similar profile. The computation is initially high; this corresponds to building the initial classifier. Subsequent peaks in the graph are the result of retraining the classifier and adaptively updating it. After 20 frames the requirement to retrain the classifier is reduced. This implies that a context specific descriptor has been learnt and it is well-suited to the current endoscopic images. If the images significantly change or a new object is introduced into the scene, the classifier may retrained to take this into account. Tracking the feature with the context specific descriptor is computationally efficient. The average time required to track a feature is 1.9ms with a standard deviation of 0.62ms. As discussed earlier, feature tracking and learning for the online tracker are independent processes which can be performed in parallel to avoid computational bottlenecks.



(a) (b)

Figure 4.11 The computational requirements of the learning phase of the system shown for **(a)** exhaustive search and **(b)** optimised search.

4.6 Discussions and Conclusions

This chapter provided a detailed explanation of a region tracking approach using online learning of local deformation. The approach is capable of dealing with changes in scale and orientation and proposed the use of synthetic smoke simulation to enable robust region tracking in the presence of smoke resulting from the use of diathermy. The approach has been validated on simulated and *in vivo* data and compared to four conventional tracking techniques. Robustness to drift, the presence of smoke, and changes in scale and orientation has been demonstrated and the capability to recover from occlusion. The potential clinical applications of the proposed technique have been highlighted by learning the global intrinsic motion of tissues and modelling it for applications in motion compensation and active constraints.

It is proven that the LK tracker had high region-density and persistency in the presence of deformation, scale change, and rotation where the algorithm's assumptions are held. However, by performing template update, the approach eventually succumbs to error propagation and drift. The regions often drifted only a short distance to nearby parts in the image. This suggests that these image parts are naturally more robust to deformation and are more suitable to tracking. This implies that tracking could be improved with the development of a deformation invariant region detector. The LK approach lacks an explicit mechanism for dealing with occlusions, and it was found to be sensitive to illumination changes. The window size was set constant for all video sequences and performance could be improved by optimising the window size for individual sequences, as shown in [219]. This would require manual parameterisation.

The results demonstrate that the mean-shift trackers are not-well suited to tracking the majority of regions detected on the surface of tissue leading to low region density. In these video sequences, colour alone is not a substantially unique characteristic for tracking regions. Region tracking may also fail under rapid motion where the region falls outside the basin of attraction. It should be noted; however, that when mean-shift can track a region, it is robust to image transformations including deformation, rotation, and partial occlusion. Both algorithms performed at a similar level - with the second approach performing slightly better. DOG regions are not ideally suited for these methods and better results are achieved with a colour region detector, such as maximally stable colour regions [220].

SIFT is not designed for use as a tracker but rather for wide baseline matching. This tracking-by-detection approach means it recovers well from occlusion and tracking failure, however, the histogram of gradients used to represent the regions does not deal well with tissue deformation or specular highlights. The approach can fail if the repeatability of the detector is low. There is a compromise between detecting too many and too few points using tracking-by-detection strategies,: matching too many points is computationally expensive and exacerbates the matching problem, while not detecting enough points leads to low repeatability and tracking failure. This is a compromise between region density and matching persistency. The DOG detector is not invariant to deformation and may not accurately, or consistently, localise the centre of detected points under observed deformation. This is due to the DOG identifying what structural information is important and using this to guide the localisation. Matching results may be improved with a more sophisticated matching strategy, thus incorporating temporal information, prior knowledge, or global outlier removal. If the histogram of gradients is not invariant to the types of deformation and image transformations observed in MIS, the improved matching strategy will have limited effect.

The results presented in this chapter show the online learning tracker performs consistently well under all image transformations with good region density and persistency. Higher density and persistency can both be attributed to the learning approach. Region density is high because the approach learns what is unique about a region relative to its surrounding, encodes this information, and uses said data to distinguish it from other regions. Region persistency is high as learning is performed directly from the observed data. This enable the approach to learn what information is robust to the observed image transformation. This approach is particularly well-suited to tracking periodically deforming tissue, however, its robustness is limited by the set of training data used to build the classifier.

Chapter 5

Simultaneous Localisation And Mapping (SLAM) For the Minimally Invasive Environment

5.1 Simultaneous Localisation And Mapping (SLAM)

Simultaneous Localisation and Mapping (SLAM) is a technique developed by the robotics community to build a map of an unknown environment while simultaneously estimating the position of the robot. It has received significant attention since the late 1980s due, in part, to the increase in computing processor power. Much of the early work was done with lasers and sonar range finders, however, there is a new body of research emerging involving cameras. This is motivated by the cameras capacity to offer a rich source of information in a compact device at a lower cost than lasers or sonar.

The SLAM problem has been solved, thus far, using a number of theoretical formulations. The seminal paper by Smith, *et al.,* [172] is widely credited for developing the basic framework required to simultaneously solve the localisation and mapping

problem. Theoretically, SLAM is considered to be a well-defined problem, and the majority of current research is focused on problems associated with the practical application of SLAM to the real-world. The following section describes some of the significant and recent contributions to the literature and discuss design choices in developing a SLAM system.

The main disadvantage of using a camera as a sensor is its inability to capture depth information because it is a 2D, passive sensor. A stereo system can be employed to address this issue. This was used in one of the earliest vision SLAM papers [221], and the idea was extended to multiple camera configurations [176, 222, 223]. In [178], Davison, *et. al.,* presented the MonoSLAM framework. The system demonstrates that, with a partial feature initialisation strategy, the SLAM problem can be formulated using a single, moving camera. This work has inspired significant research into monocular SLAM [224-227] including the use of a single, Omni directional camera [228].

Probabilistic methods are crucial to the success of SLAM. A probabilistic method enables the system to model noise in sensor measurement, model the correlation between features in the environment, and predict and estimate the sensor position and map with associated uncertainty. The most common probabilistic frameworks for solving SLAM are the Extended Kalman Filter (EKF) and Particle Filter. The EKF has been extensively used in SLAM [178], despite errors that can be introduced into the system through linear approximation. The basic EKF, with single state vector and full covariance matrix, has a computational complexity of $O(N^2)$, where N is the number of features in the map. This limits the size of map that can be built and updated in real-time. In addition, the a posteriori distributions are represented as uni-modal Gaussians allowing only one hypothesis about the state of the system or the position of the sensor and map. This can lead to system failure in the presence of ambiguous sensor readings.

Particle filters are based on Monte Carlo sampling of particle distributions. This enables the process model to be non-linear and the pose distribution of the sensor to be non-Gaussian and multi-modal. In [229], a FastSLAM algorithm is presented, which, by using the Rao-Blackwellised particle filter, addresses the issue of the high-dimensional state space of particle filters. Through marginalisation, the map is represented as independent Gaussians that enable real-time performance. The capabilities of FastSLAM

have been demonstrated in [229, 230]. This approach can suffer statistically from degeneration and the marginalisation of the map creates dependency of measurement and pose history. Particle filters have been applied to vision-based SLAM for re-localisation and to improve robustness of rapid camera motion [224, 225]. In [226], FastSLAM is used to increase the number of features that can be mapped in real-time.

The map can be represented as an occupancy grid [231, 232] or feature map. Occupancy grids are generally used with range finders and represent the environment by dividing it into a cell-grid. Feature maps are the standard representation in vision approaches and consist of a number of 3D positions in the environment. Each position represents a significant part of the environment or landmark. These maps are generally sparse, and each feature is associated with a spatial uncertainty and usually some image information to enable the performance of data association. In early studies, monoSLAM carefully manages the number of features (around 100) to ensure real-time performance. The desire for larger maps has led to topological and sub-mapping [233] algorithms. The work of Klein [227] uses key frames to increase the number of features that are mapped. This approach uses SLAM and Structure-From-Motion, in parallel, to enable sequential mapping and small batch process, thus reducing the map's uncertainty.

Data association is an essential part of the SLAM framework. In vision SLAM there are two approaches to data association; 1) create a dense 3D reconstruction using stereo cameras [176, 222, 234, 235]; and 2) match regions of interest in the image plane [176, 224-226, 236-241]. The former approach aligns the measured map with the existing map using various techniques to minimise entropy or Iterative Closest Point (ICP). These methods can be computationally expensive. The latter is far more common and matches, or tracks, regions of interest. This is discussed, in detail, in Chapter 2. A number of SLAM systems [176, 236-238] have employed Scale Invariant Feature Transform (SIFT) [124] to exploit invariance to scale and rotation. Other systems opt for computationally faster techniques for extraction and matching such as Harris corners [240] and Shi and Tomasi [239] with cross-correspondence and sum of squared difference. In [241], randomised lists are used to increase matching speed and recovery from failure.

Data association can be improved, both in performance and in computational speed, by using a matching strategy. The matching strategy defines how the system will search for the corresponding regions and remove outliers. In [239], the features are projected onto

the image plane along with the uncertainty in the feature location. This is known as active search. Active search reduces the search area, thus increasing computational speed and reducing the likelihood of mismatching data. If data falls outside the search window, however, the data association will fail. Alternative approaches, such as [242] and [241], can be used to search for correspondences with all features in the map. This requires robust region matching and outlier removal but provides recovery from failure.

A variety of methods can be used for camera motion prediction. The more efficient approach is to use the recent history of the camera's movement and assume that the device will follow a similar path. This is known as a constant velocity motion model [239]. If the camera is attached to a robot, odometry [237, 243] may be used. Alternatively, if the system is used outdoors, Global Positioning System (GPS) can determine longitude, latitude, and altitude data. An Inertia Measurement Unit (IMU) [244, 245] can provide information about position by measuring the accelerations and rotations applied to the unit using accelerometers and gyros. Reliable and accurate IMU's can be large and expensive. The motion of a camera can be estimated from frame to frame using Visual Odometry [236] or Structure-From-Motion. This approach is complementary to visual SLAM, however, it shares the same risks of failure, including image blur. Regardless of the method employed, the new position cannot be known with 100% accuracy because the sensor estimates contain errors. The probabilistic model provides a way of dealing with discrepancies between the motion estimate and the actual motion by modelling uncertainty or an unknown element in the system. This uncertainty or unknown element may take the form of calibration error and wheel slippage for odometry, or unknown acceleration for the constant velocity model.

The success of SLAM has led to a variety of applications in situations dangerous to humans [246], service robots [247], planetary exploration [236], entertainment or toy robots, and underwater exploration [244]. This does not mean SLAM is limited to robot navigation applications. It is also applicable to wearable computing [248], augmented reality [178], and the generation of photo-realistic models [176]. Each of these applications has its own set of specific research challenges. This chapter will investigate the use of SLAM in MIS to identify new areas of research. During MIS, the surgeon, or first assistant, navigates the laparoscopic camera through a cavity inside the patient to see the organs, as shown in **Figure 5.1**. SLAM in MIS is used to localise the laparoscope and build a map of the tissue and organs.

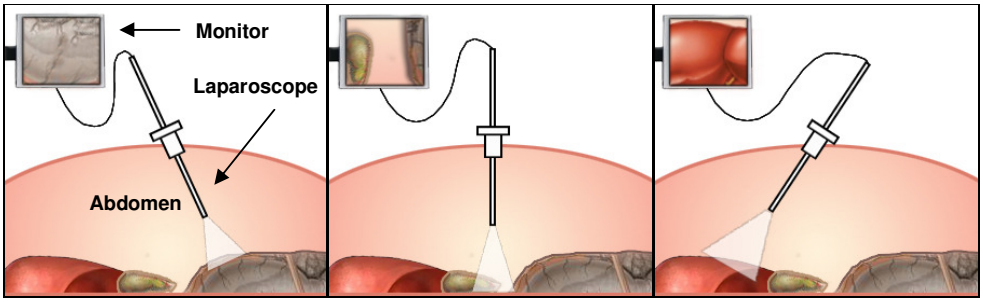


Figure 5.1 An illustration of laparoscopic movement during MIS. The laparoscope is inserted through an incision in the abdomen wall to visualise the internal organs. The surgeon controls the laparoscope and images are displayed on the monitor. The incision-point in the abdomen wall creates a pivot point and the fulcrum effect.

5.2 SLAM for MIS

The main considerations when using SLAM for MIS will be discussed in the following sections. The system employed is based on the MonoSLAM [178] system which is publically available [249]. The system has been extended to function with a stereo system and the region tracking algorithm outlined in **Chapter 4**. In addition, image pre-processing is used to enable tracking with low-quality fibre optic images. The SLAM framework is based on the EKF, and the fundamental steps of SLAM are illustrated in **Figure 5.2**. First a description of the EKF framework and how this is formulated to predict camera motion, measure the state, and to update the state estimation is provided. This is followed by an explanation of feature initialisation and map measurement.

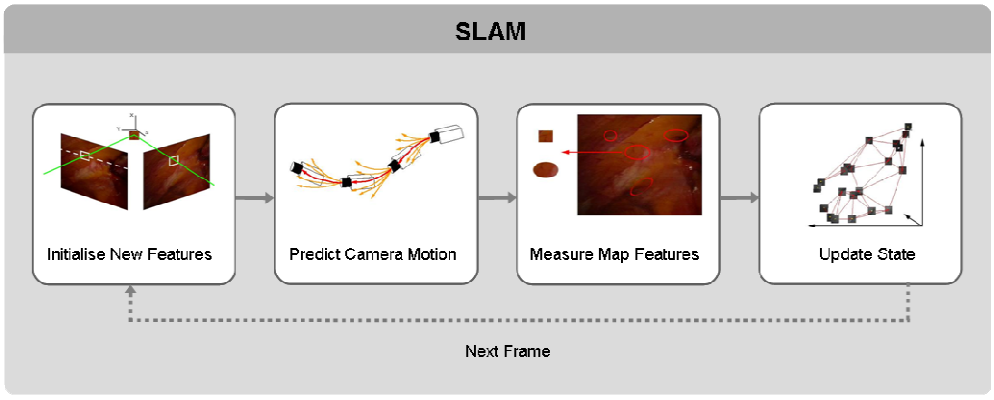


Figure 5.2 A schematic of the SLAM framework including feature initialisation, camera prediction, measurement model, and state (camera and map) update.

5.2.1 Extended Kalman Filter (EKF)

The SLAM system is based on the Extended Kalman Filter (EKF). A Kalman Filter [250] is a recursive Bayesian Filter which estimates the state of a system in the presence of noisy measurements. The filter assumes the system (prediction and measurement models) is linear and that the noise in the system can be modelled as Gaussian. A Gaussian distribution can be fully represented in closed form, by its mean and covariance matrix. The current state can be stored as a multidimensional vector $\hat{\mathbf{x}}$ and covariance matrix $\mathbf{P}$ (size of the state vector squared). The Bayesian framework is used to update the mean and covariance matrix. The computational cost of the update is that of a matrix multiplication $O(N^2)$ where N is the dimension of the state vector. If the assumption is that the system is linear, and the distribution of noise is Gaussian, then the solution is optimal in a least squared sense. Unfortunately, most real-world systems do not hold true to these assumptions and mechanisms are required to cope with non-linearity and non-Gaussian noise.

The EKF [251] is a popular and simple extension of the Kalman Filter when dealing with non-linearity. In the EKF, the prediction and measurement models are differentiable, non-linear functions. The estimated state is linearised around the current state by computing a matrix of partial derivatives, known as the Jacobian.

5.2.1.1 EKF State Prediction

A new state of the system is blindly predicted during the prediction step using the prediction or state transition model. The prediction model $\mathbf{f}$ is comprised of two elements; deterministic and stochastic. The deterministic part predicts the new state of the system at time t using the previous state $\hat{\mathbf{x}}_{t-1}$, the prediction model parameterises to define how the state is expected to evolve over time and, when available, a control input $\mathbf{u}$. The stochastic part of the system models the uncertainty in the prediction and accounts for components in the real world which are hard to model. This is known as the process noise $\mathbf{Q}$ and is used to increase the uncertainty in the covariance matrix $\mathbf{P}$. The predicted state is known as the *a priori* state and does not use any measurement information. The prediction equations are defined as

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{f}(\hat{\mathbf{x}}_{t-1}, \mathbf{u}_t) \tag{5.1}$$

$$\mathbf{P}_{t|t-1} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}_{t-1}} P_{t-1} \frac{\partial \mathbf{f}}{\partial \mathbf{x}_{t-1}}^{T} + \mathbf{Q}_{t-1} \tag{5.2}$$

5.2.1.2 EKF State Update

In the update step, the measurement of the state $\mathbf{z}_t$ is compared to the predicted state $\mathbf{h}(\hat{\mathbf{x}}_{t|t-1})$, where the measurement model $\mathbf{h}$ maps the state prediction $\hat{\mathbf{x}}_{t|t-1}$ into the measurement space. The incorporation of measurement information into the state estimate reduces the uncertainty $\mathbf{P}$ in the system. The new, improved state estimate is known as the *a posteriori* state, and the update equations are defined as

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{W}_t \nu_t \tag{5.3}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{W}_t \mathbf{S}_t \mathbf{W}_t^T \tag{5.4}$$

where v is the innovation that represents the difference between the actual measurement and the predicted measurement calculated from the current state.

$$\nu_t = \mathbf{z}_t - \mathbf{h}(\hat{\mathbf{x}}_{t|t-1}) \tag{5.5}$$

$\mathbf{W}$ represents the Kalman gain

$$\mathbf{W}_t = \mathbf{P}_{t|t-1} \frac{\partial \mathbf{h}}{\partial \mathbf{x}_{t-1|t-1}}^{T} S_t^{-1} \tag{5.6}$$

and $\mathbf{S}$ is the innovation covariance

$$\mathbf{S}_t = \frac{\partial \mathbf{h}}{\partial \mathbf{x}_{t-1|t-1}} \mathbf{P}_{t|t-1} \frac{\partial \mathbf{h}}{\partial \mathbf{x}}^{T} (\hat{x}_{t-1|t-1}) + \mathbf{R}_t \tag{5.7}$$

where $\mathbf{R}$ is the measurement noise.

The EKF has implementation disadvantages because it is computationally more complex than the Kalman Filter. Furthermore, implementing Jacobian calculations can be non-trivial, and it is sensitive to inaccuracies in initialisation. Theoretically, EKF only offers

approximations for non-linear systems. The approximation of a non-linear system can lead to errors in estimations for highly non-linear functions as higher order terms are neglected. This can result in sub-optimal performance or even divergence of the filter.

5.2.2 Extended Kalman Filter for SLAM

5.2.2.1 System initialisation

The goal of laparoscope localisation and soft-tissue mapping in MIS is the recovery of the laparoscopic trajectory and to sequentially build a 3D map of the tissue. With the use of the EKF framework, a vector is used to represent the over-all state of the system $\hat{\mathbf{x}}$. The vector is partitioned into two parts; the camera state $\hat{\mathbf{x}}_{\mathrm{v}}$ and the map state $\hat{\mathbf{y}}_{\mathrm{i}}$, where the map consists of multiple 3D features:

$$\hat{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}}_v \\ \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \end{pmatrix} \tag{5.8}$$

$$\hat{\mathbf{x}}_v = \begin{pmatrix} \mathbf{r}^W \\ \mathbf{q}^{WR} \\ \mathbf{v}^W \\ w^R \end{pmatrix} \qquad \hat{\mathbf{y}}_i = \begin{pmatrix} \hat{X}_1 \\ \hat{Y}_1 \\ \hat{Z}_2 \end{pmatrix} \tag{5.9}$$

where the camera state is represented by the 3D position of the camera in the world coordinate system $\mathbf{r}^W$, the rotation (quaternion) of the camera (pose) in the world coordinate system $\mathbf{q}^{WR}$, the velocity $\mathbf{v}^W$ and angular velocity w^R. The i^{th} feature in the map $\hat{\mathbf{y}}_i$ is represented by its 3D position.

Crucially, a single covariance matrix $\mathbf{P}$ accompanies the state vector. This symmetric matrix represents the uncertainty to first order in all quantities of the state vector and is partitioned such that

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy1} & \mathbf{P}_{xy2} & \cdots \\ \mathbf{P}_{y1x} & \mathbf{P}_{y1y1} & \mathbf{P}_{y1y2} & \cdots \\ \mathbf{P}_{y2x} & \mathbf{P}_{y2y1} & \mathbf{P}_{y2y2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{5.10}$$

$\mathbf{x}$ and $\mathbf{P}$ are able to grow, or shrink, dynamically as feature estimates $\hat{\mathbf{y}}_i$ can be freely added to, or deleted, from the map as required. A full feature covariance matrix $\mathbf{P}$ is maintained. This enables the camera to re-visit and recognise previously visited areas. The filter is initialised with:

$$\hat{\mathbf{x}}_v = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.11}$$

such that the origin of the coordinate system is the camera centre on the first frame.

5.2.2.2 SLAM State Prediction

In the context of SLAM, the EKF prediction step estimates the motion of the camera and its new location. The laparoscope is considered to be under human control, therefore, the prediction model must account for the unknown intentions of the operator. As described above, a two-part prediction model is used to statistically model this unknown entity. The first part is a deterministic element $\mathbf{f}$, which estimates the camera's motion based on an assumption or input. In this work, a *constant-velocity, constant-angular-velocity* motion model is employed. This does not assume that the camera moves at a constant velocity and constant angular velocity. Rather, it assumes that the expected velocity and angular velocity remain constant between frames. The most recent measured camera motion can be used to predict the next camera motion. The second part of the prediction model is stochastic. This stochastic element $\mathbf{Q}$ models the uncertainty in the surgeon's movement of the laparoscope. The uncertainty in this system is the unknown acceleration value. This is modelled with a Gaussian profile. The prediction model assumes that the laparoscope moves smoothly and large accelerations are unlikely.

The prediction model of the state is therefore:

$$\hat{\mathbf{x}}_{t|t-1} = \begin{pmatrix} \mathbf{f}_v(\hat{\mathbf{x}}_{t-1|t-1}) \\ \hat{\mathbf{y}}_{1(t-1|t-1)} \\ \hat{\mathbf{y}}_{2(t-1|t-1)} \\ \vdots \end{pmatrix} \tag{5.12}$$

where

$$\mathbf{f}_v = \begin{pmatrix} \mathbf{r}^W_{new} \\ \mathbf{q}^{WR}_{new} \\ \mathbf{v}^W_{new} \\ w^R_{new} \end{pmatrix} = \begin{pmatrix} \mathbf{r}^W + (\mathbf{v}^W + \mathbf{V}^W)\Delta t \\ \mathbf{q}^{WR} + \mathbf{q}(w^W + \Omega^R)\Delta t \\ \mathbf{v}^W + \mathbf{V}^W \\ w^R + \Omega^R \end{pmatrix} \tag{5.13}$$

The velocity and acceleration in the system are modelled such that:

$$\mathbf{n} = \begin{pmatrix} \mathbf{V}^W \\ \Omega^R \end{pmatrix} = \begin{pmatrix} a^W \Delta t \\ \alpha^r \Delta t \end{pmatrix} \tag{5.14}$$

Where a^W is the unknown acceleration and α^r is the unknown angular acceleration with the process noise covariance for the camera motion $\mathbf{Q}_v$ computed using the Jacobian calculation :

$$\mathbf{Q}_v = \frac{\partial \mathbf{f}_v}{\partial \mathbf{n}} \mathbf{P}_n \frac{\partial \mathbf{f}_v}{\partial \mathbf{n}}^T \tag{5.15}$$

where $\mathbf{P}_n$ is the covariance of noise vector $\mathbf{n}$.

The size of parameter $\mathbf{P}_n$ determines the growth rate of uncertainty in this motion model, and the smoothness of the expected motion is defined by setting this parameter as small or large. A motion model for very smooth motion with small accelerations is created by setting $\mathbf{P}_n$ to be small. This system would be unable to cope with sudden rapid motion or directional change. To cater to rapid accelerations $\mathbf{P}_n$ can be set to a high value. This means the uncertainty in the system increases significantly with each time step. In order

to maintain accurate state estimates with high $\mathbf{P}_n$, each step requires accurate measurements of the features in the map.

The uncertainty in the covariance matrix $\mathbf{P}$ is updated to reflect the increase in uncertainty in the predicted state such that

$$\mathbf{P}_{t|t-1} = \begin{bmatrix} \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v} \mathbf{P}_{xx(t-1|t-1)} \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v}^T + Q_{t-1} & \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v} \mathbf{P}_{xy_1(t-1|t-1)} & \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v} \mathbf{P}_{xy_2(t-1|t-1)} & \cdots \\ \mathbf{P}_{y_1x(t-1|t-1)} \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v}^T & \mathbf{P}_{y_1y_1(t-1|t-1)} & \mathbf{P}_{y_1y_2(t-1|t-1)} & \cdots \\ \mathbf{P}_{y_2x(t-1|t-1)} \frac{\partial \mathbf{f}_v}{\partial \mathbf{x}_v}^T & \mathbf{P}_{y_2y_1(t-1|t-1)} & \mathbf{P}_{y_2y_2(t-1|t-1)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \textbf{(5.16)}$$

Where $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ is the full state transition Jacobians

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} & 0 & 0 & \cdots \\ 0 & I & 0 & \cdots \\ 0 & 0 & I & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \textbf{(5.17)}$$

5.2.2.3 SLAM State Update

The measurement $\mathbf{z}$ is compared to the predicted $\mathbf{h}(\hat{\mathbf{x}}_{t|t-1})$ in the update step. In SLAM, measurement $\mathbf{z}$ is the position of features in 3D relative to the camera. The measurement model $\mathbf{h}$ maps the current state into the space where the measurements are taken and equates to translating the map features $\mathbf{y}_i^W$ from their 3D position in the world coordinate system to the camera coordinate system. This is calculated by using the predicted position of the camera and $\mathbf{y}_i^W$

$$\mathbf{h}_i^R = \mathbf{R}^W(\mathbf{y}_i^W - \mathbf{r}_L^W) \quad \textbf{(5.18)}$$

where $\mathbf{R}^W$ is the rotation matrix and $\mathbf{r}_L^W$ is the position vector of the left camera centre. ^{R} is the camera coordinate system and ^{W} is the world coordinate system. This enables the calculation of ν, which represents the difference between the actual measurement and the predicted, calculated measurement from the current state using **Equation (5.5)**. The Kalman $\mathbf{W}$ gain in the system is estimated by

$$\mathbf{W}_t = \mathbf{P}_{t|t-1} \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}}^T (\hat{\mathbf{x}}_{t-1|t-1}) \mathbf{S}_t^{-1} = \mathbf{S}_t^{-1} \begin{pmatrix} \mathbf{P}_{xx} \\ \mathbf{P}_{y_1 x} \\ \mathbf{P}_{y_2 x} \\ \vdots \end{pmatrix} \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}_v}^T + S_t^{-1} \begin{pmatrix} \mathbf{P}_{xy_i} \\ \mathbf{P}_{y_1 y_i} \\ \mathbf{P}_{y_2 y_i} \\ \vdots \end{pmatrix} \frac{\partial \mathbf{h}_i}{\partial \mathbf{y}_i}^T \quad \textbf{(5.19)}$$

and $\mathbf{S}$ is the innovation covariance

$$\mathbf{S}_t = \frac{\partial \mathbf{h}}{\partial \mathbf{x}_{t-1|t-1}} \mathbf{P}_{t|t-1} \frac{\partial \mathbf{h}}{\partial \mathbf{x}}^T (\hat{\mathbf{x}}_{t-1|t-1}) + \mathbf{R}_t \quad \textbf{(5.20)}$$

$$S_t = \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}} \mathbf{P}_{xx} \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}}^T + 2 \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}_v} \mathbf{P}_{xy_i} \frac{\partial \mathbf{h}_i}{\partial \mathbf{y}_i}^T + \frac{\partial \mathbf{h}_i}{\partial \mathbf{y}_i} \mathbf{P}_{y_i y_i} \frac{\partial \mathbf{h}_i}{\partial \mathbf{y}_i}^T + \mathbf{R}_t \quad \textbf{(5.21)}$$

such that $\mathbf{P}_{xx}$, $\mathbf{P}_{xy_i}$ and $\mathbf{P}_{y_i y_i}$ are the sub-matrix blocks from the covariance matrix $\mathbf{P}$.

5.2.3 Feature Measurement

The state update and measurement models require a measurement of the 3D features relative to the predicted camera position. This work uses stereo cameras to estimate the 3D position of features by matching 2D image regions in the left and right images and performing triangulation, as described in **Chapter 2**. Prior to determining the 3D position a feature, the corresponding region must be identified using the image data..

Although the feature is represented as a point in 3D, an image descriptor or template is associated with it when initialised. This information enables the feature to be matched in future images. In principle, any region tracking method outlined in **Chapter 2** can be used to solve data association. The online learning method described in **Chapter 4** is used. In MIS, identifying corresponding regions is difficult and, to further constrain the

problem, an active search method [239] must be used to reduce the probability of mismatches and increase computational efficiency.

Active search, as proposed in [239], uses the spatial and uncertainty information in the map to constrain the area in which a correspondence will be sought. The spatial information in the map is used to predict the location of the i^{th} feature $\mathbf{y}_i$ in the image plane. The position is predicted using the intrinsic parameters and the predicted position of the camera. First, the location of each feature in the map $\mathbf{y}_i^W$ is computed relative to the camera using **Equation (5.18)** and giving $\mathbf{h}_i^R$.

The point $\mathbf{h}_i^R$ is projected into the image plane using a standard pinhole projection model

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} u_0 - fk_u \dfrac{\mathbf{h}_i^R x}{\mathbf{h}_i^R z} \\ v_0 - fk_v \dfrac{\mathbf{h}_i^R y}{\mathbf{h}_i^R z} \end{pmatrix} \tag{5.22}$$

where fk_u and fk_v represent the camera's focal length and u_0 and v_0 are the principal points and radial warp is considered by using [252].

The area to be searched is derived from the uncertainty of the feature's predicted position. This is a 2D Gaussian p.d.f. over the image coordinates in image space. The size of the search area is defined by gating three standard deviations. This creates an elliptic search window centred on the feature's predicted position in the left image space. A similar approach is used to identify the position of the feature in the right image, with the additional constraint of the epipolar line defined by the position of the matched feature in the left image. The left and right feature positions are triangulated to estimate the 3D position of the feature relative to the camera.

5.2.4 Feature Initialisation

New features are initialised and added to the map if the total number of visible features falls below a pre-determined threshold. MIS is a challenging environment for region tracking, and it was found, by tracking 20 features, an adequate trade-off between

computational performance and robust tracking was established. A feature is initialised by detecting a salient region in the image. Difference of Gaussian (DOG) [124] and Shi and Tomasi [105] features detectors were used. Theoretically, the benefits of DOG features are not fully-realised because the active search approach uses temporal information to constrain the matching problem. Consequently, the computationally less expensive Shi and Tomasi algorithm is favoured where specular highlights are identified in the image using a manually defined threshold in the saturation channel. Features detected close to highlights are discarded.

A region detected in the left stereo image is matched in the right stereo image by searching the epipolar line. To ensure the feature is a good representation of the tissue structure, outliers were removed using RANSAC. The detected points in the left and right images are triangulated to estimate the feature's 3D position relative to the camera $\mathbf{y}_i^R$. This position is re-projected onto the image plane, and features with a large re-projection error are discarded. The feature position in the world coordinate system $\mathbf{y}_i^W$ is computed using the current *a posterior* estimate of the camera position.

The new feature $\mathbf{y}_i$ is inserted into the state vector and covariance matrix such that

$$\hat{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}}_v \\ \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \\ \hat{\mathbf{y}}_i \end{pmatrix} \tag{5.23}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy1} & \mathbf{P}_{xy2} & \cdots & \mathbf{P}_{xx}\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}^T \\ \mathbf{P}_{y1x} & \mathbf{P}_{y1y1} & \mathbf{P}_{y1y2} & \cdots & \mathbf{P}_{y_1x}\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}^T \\ \mathbf{P}_{y2x} & \mathbf{P}_{y2y1} & \mathbf{P}_{y2y2} & \cdots & \mathbf{P}_{y_2x}\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}^T \\ \vdots & \vdots & \vdots & \ddots & \\ \frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xx} & \frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xy_1} & \frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xy_2} & & \frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xx}\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_v}^T + \frac{\partial \mathbf{y}_i}{\partial \mathbf{h}_i}\mathbf{R}\frac{\partial \mathbf{y}_{ii}}{\partial \mathbf{h}_i}^T \end{bmatrix} \tag{5.24}$$

5.2.5 Honeycomb Artefact Removal

Flexible endoscopes are commonly used in gastrointestinal surgery to safely visualise anatomy that cannot be accessed with a rigid laparoscope. These instruments have been used more recently in Natural Orifice Transluminal Endoscopic Surgery (NOTES), to access the abdominal cavity via natural orifices. The requirement for flexible instruments constrains the optical set-up. The optical configuration for endoscopes commonly consists of an objective lens that focuses light onto a coherent optical fibre bundle.

The optical fibre bundle is flexible and acts as a light guide, transmitting light from the distal tip to the proximal end of the endoscope. The proximal end of the fibre is aligned with a Charge-Coupled Device (CCD) camera, which digitally images the transmitted light. To reduce cross-talk interference between adjacent fibres in the bundle, each individual fibre is manufactured with a non-transparent coating. This coating is visible in the image captured by the CCD camera and creates a honeycomb effect wherein the light from each individual fibre is surrounded by negative space, see **Figure 5.3 (a)** and **Figure 5.3 (f)**. This can have an adverse effect on SLAM.

The noise introduced into the image by the fibre coatings, or honeycomb effect, can make the region detection and tracking challenging. The noise causes sharp gradients in the image, which can lead to the erroneous detection of regions of interest in homogenous areas. In region tracking and matching without image pre-processing, the noise is incorporated into the representation of the regions. This creates an inaccurate representation of the region and can lead to mismatches and localisation inaccuracies. This problem is further complicated by sub-millimetre movements of the optics relative to the CCD. This can lead to changes in the image and movement of the honeycomb structure on the CCD chip. It is not possible to assume that the structure of the noise is constant.

Several techniques have been proposed for removal of the honeycomb effect. For practical surgical use, it is possible to defocus the proximal imaging optics. This removes the honeycomb structure but can cause blurring and the subsequent loss of information. The optical fibre bundle used in this experiment, and described later, was relatively low resolution (10,000 fibres) and it was not desirable to use this approach. In [253], a band pass Fourier transform is employed to remove the noise. Following this work, in [254],

the authors use the Bayer CCD pattern of the camera and shaped Fourier filters designed to estimate the structure of the noise. It was experimentally observed that sub-millimetre movements of the optics relative to the CCD could result in changes in the image. As such, it can not be assumed the honeycomb structure is static in the image, and a robust approach, that did not require re-calibration after each use, is required.

A band pass filter in the Fourier frequency space is used to remove the honeycomb effect. This demonstrated adequate, quality image restoration and enabled region tracking. The alignment of the fibres creates a regular spatial structure in the image, as shown in **Figure 5.3 (f)**. In the Fourier spectrum, this regular pattern is distinct from underlying image data. The pattern is represented by high frequencies caused by edges in the honeycomb structure, **Figure 5.3 (d)**. Suppressing frequencies associated with noise in the Fourier domain will remove the pattern when the Fourier image is converted back to image space using the inverse Fourier transform. To suppress these frequencies, a circular band pass filter is used, **Figure 5.3 (e)**. The parameters of the filter were empirically defined. The parameters of the filter remain constant once estimated. A processed image is shown in **Figure 5.3 (f)** and **Figure 5.3 (b)**. In the SLAM framework, the images are pre-processed with the band pass filter and smoothed with a Gaussian filter prior to tracking.

5.3 Experimental Results

5.3.1 *In Vivo* Experiments

In order to evaluate the practical value of the SLAM framework, an *in vivo* porcine experiment was performed. The experiment was performed with a da Vinci robot. During the procedure, stereo video data was captured from a stereo laparoscope. Two data sequences were used. Only qualitative results are provided since the ground truth data was not available. The surgeon explores the abdomen in both sequences by navigating the laparoscope. A small amount of deformation is visible, however, the tissue is assumed to be static. Both sequences contain view dependent specular highlights and changes in illumination resulting from alterations in the camera and light source position.

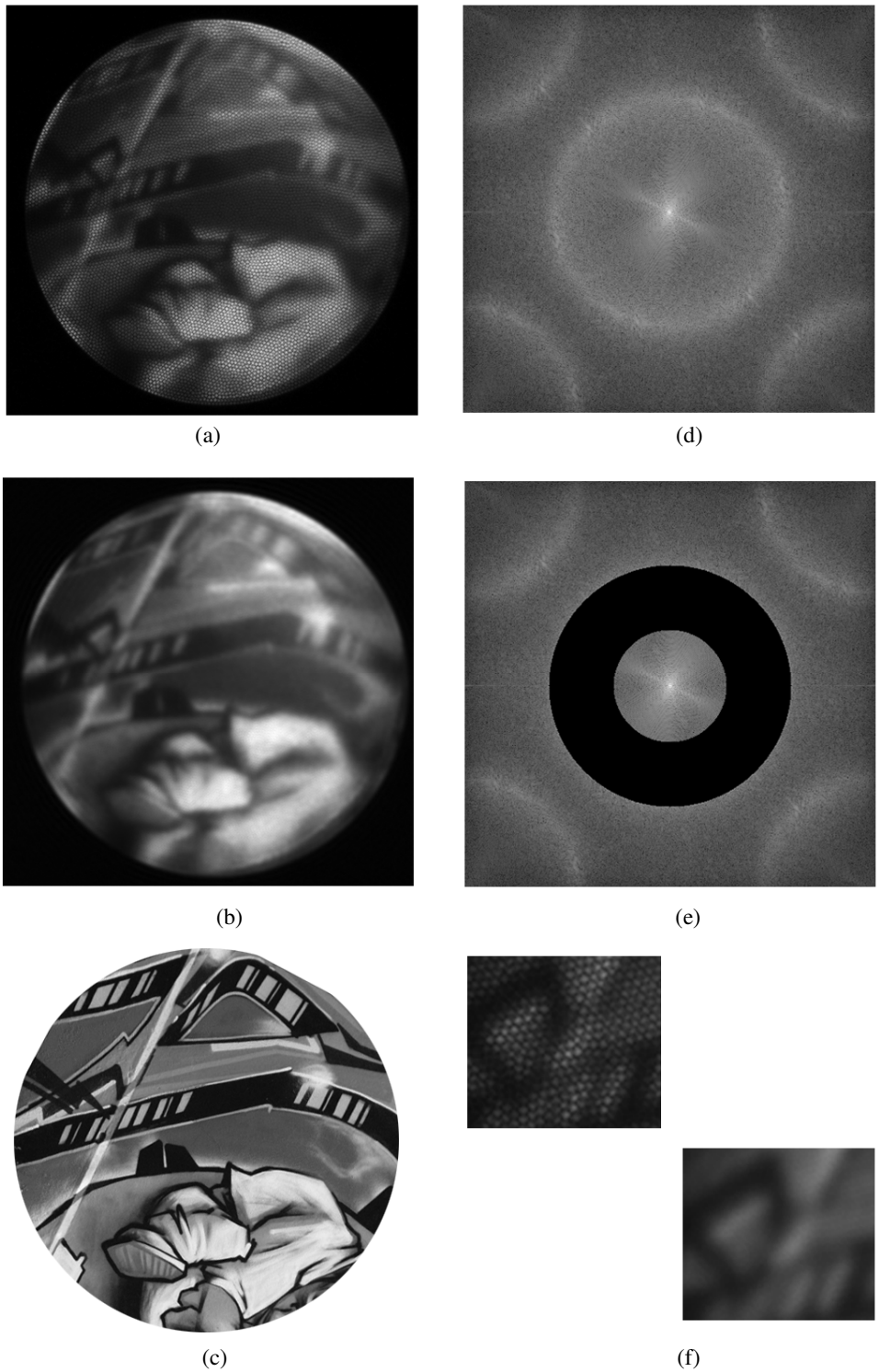


Figure 5.3 Honeycomb noise removal from fibrescopic images. **(a)** Original test image captured by fibre bundle, **(b)** test image after honeycomb removal, **(c)** test image, **(d)** original image in Fourier domain, **(e)** band pass filter applied in Fourier domain **(f)** - top close-up of *(a)* and **(f)** - bottom close-up of *(b)*.

The first sequence is shown in **Figure 5.4**. The surgeon first navigates the laparoscope left and right (along the X axis) before moving it up and down (along the Y axis) and returning to the start position. The manoeuvre requires approximately 40 seconds. The corresponding results are provided in **Figure 5.4** and **Figure 5.5**. **Figure 5.4** illustrates the motion of the laparoscope and the recovered 3D structure of the tissue. It is evident that the map is built incrementally over time. In these figures, the 3D positions of the map features are drawn as a 3D ellipse, thus representing their inherent uncertainty. The recovered motion of the laparoscope visually corresponds to the motion observed in the images. The fulcrum effect is clearly visible from the orientation of the laparoscope, which can be used to indicate the 3D position of the surgical port. The sequence shown in **Figure 5.4** demonstrates the incremental map building and loop closure (*i.e.*, the ability to return to previously visited viewing positions). In **Figure 5.4 (f)**, the derived map is meshed to create a model of the tissue surface.

Figure 5.5 shows a graphic representation of the laparoscope's estimated pose over time. The rotations around the X, Y and Z axes are shown in **Figure 5.5 (a-c)**, and the translation in the X, Y and Z global coordinates system are shown in **Figure 5.5 (d-f)**. The corresponding motion paths along the left, right, up, and down directions are illustrated therein. Translation in the X axis is accompanied by a rotation around the Y axis and translation in the Y axis is accompanied by a rotation around the X axis due to the fulcrum effect.

The second experimental sequence is shown in **Figure 5.6**. In this case, the surgeon freely explores the abdominal cavity, navigating the laparoscope for a period of over 40 seconds. During this exploration, the laparoscope translates and rotates around all axes. The corresponding results are provided in **Figure 5.6** and **Figure 5.7**. In **Figure 5.6**, the estimated laparoscope position and the 3D SLAM map are shown, thus demonstrating consistent, incremental mapping and localisation. The rotation and translation of the laparoscope are graphically represented around the X, Y and Z axes in **Figure 5.7**.

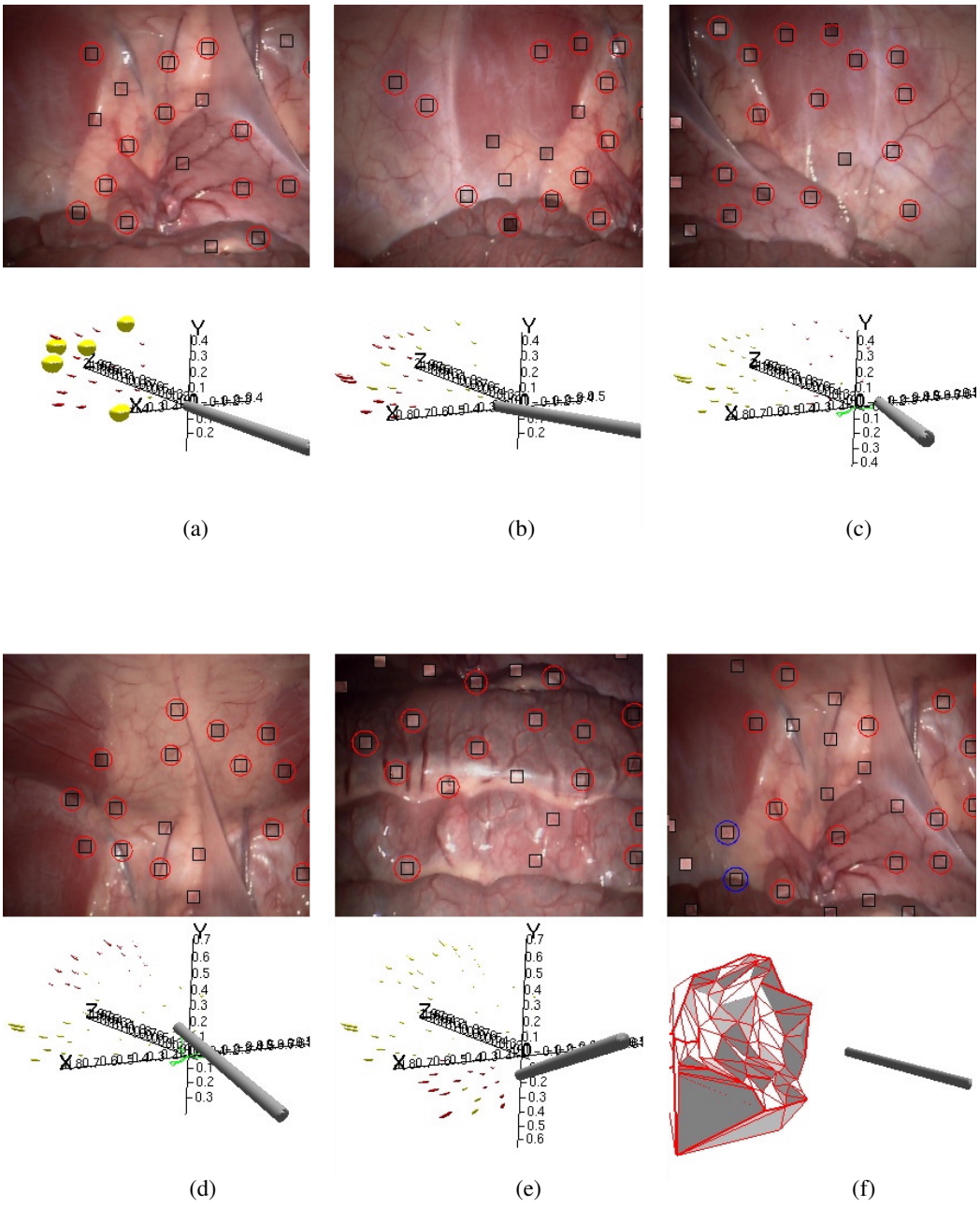


Figure 5.4 (a-f) Results from an *in vivo* experiment with the SLAM framework showing the laparoscopic images and the SLAM coordinate system. The grey cylinder indicates the current position and pose of the laparoscope in the SLAM coordinate system. The position of the map features are represented by their elliptical uncertainty. In the laparoscopic images, the black boxes indicate the position of features and the red ellipses show the uncertainty in the features position. **(a)** System initialisation, laparoscope moves **(b)** left, **(c)** right, **(d)** up, and **(e)** down. **(f)** Shows a surface model.

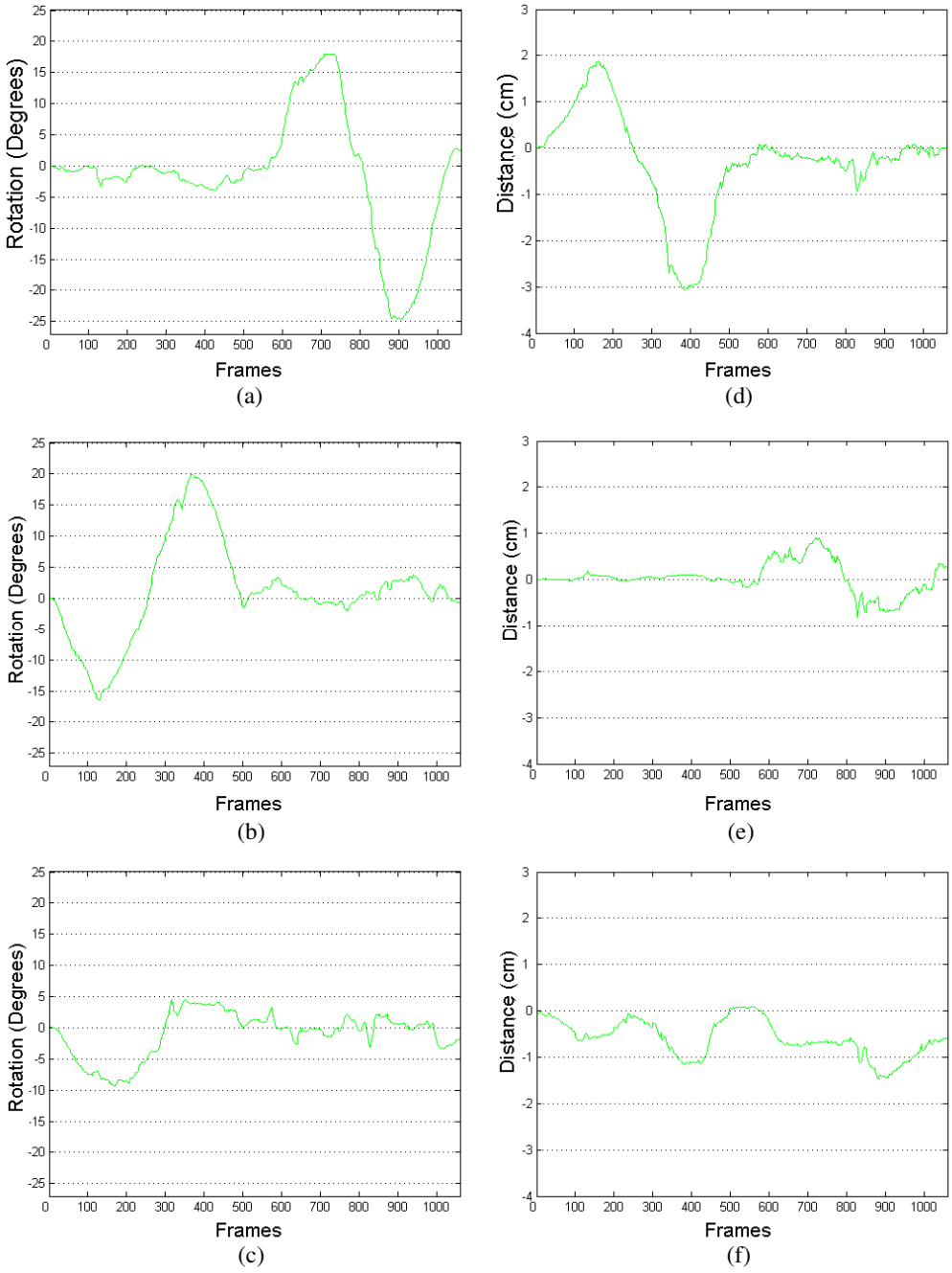


Figure 5.5 Results from *in vivo* experiments with SLAM framework for **(a-c)** rotation around the X, Y, and Z axes and **(d-f)** translation along the X, Y, and Z axis.

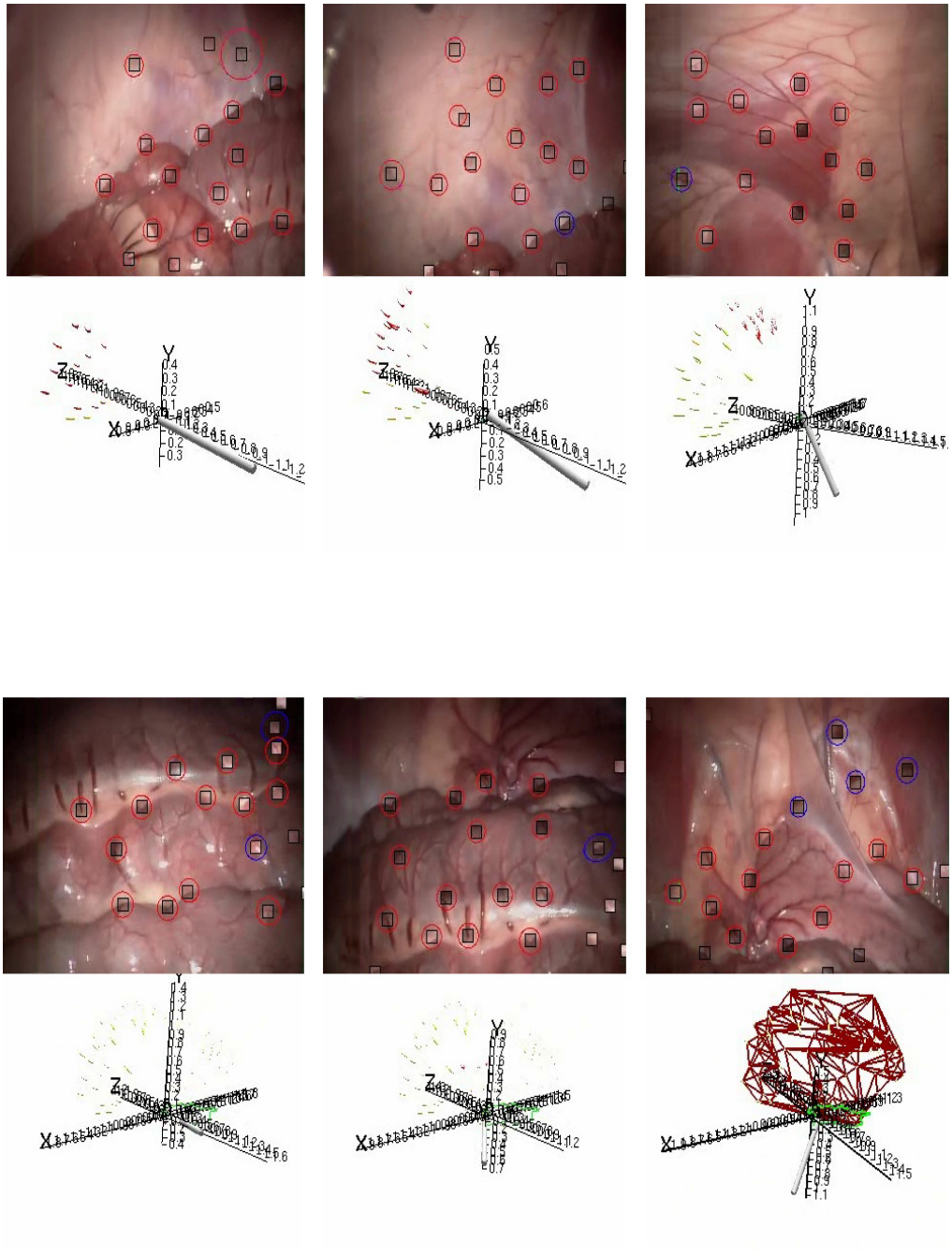


Figure 5.6 Results from a second *in vivo* experiment with the SLAM framework showing the laparoscopic images and the SLAM coordinate system. The current position and pose of the laparoscope in the SLAM coordinate system is shown by the grey cylinder. The position of the map features are represented by their elliptical uncertainty. In the laparoscopic images, the black boxes indicate the position of features and the red ellipses show the uncertainty in the features position. Features shown in blue are not being tracked.

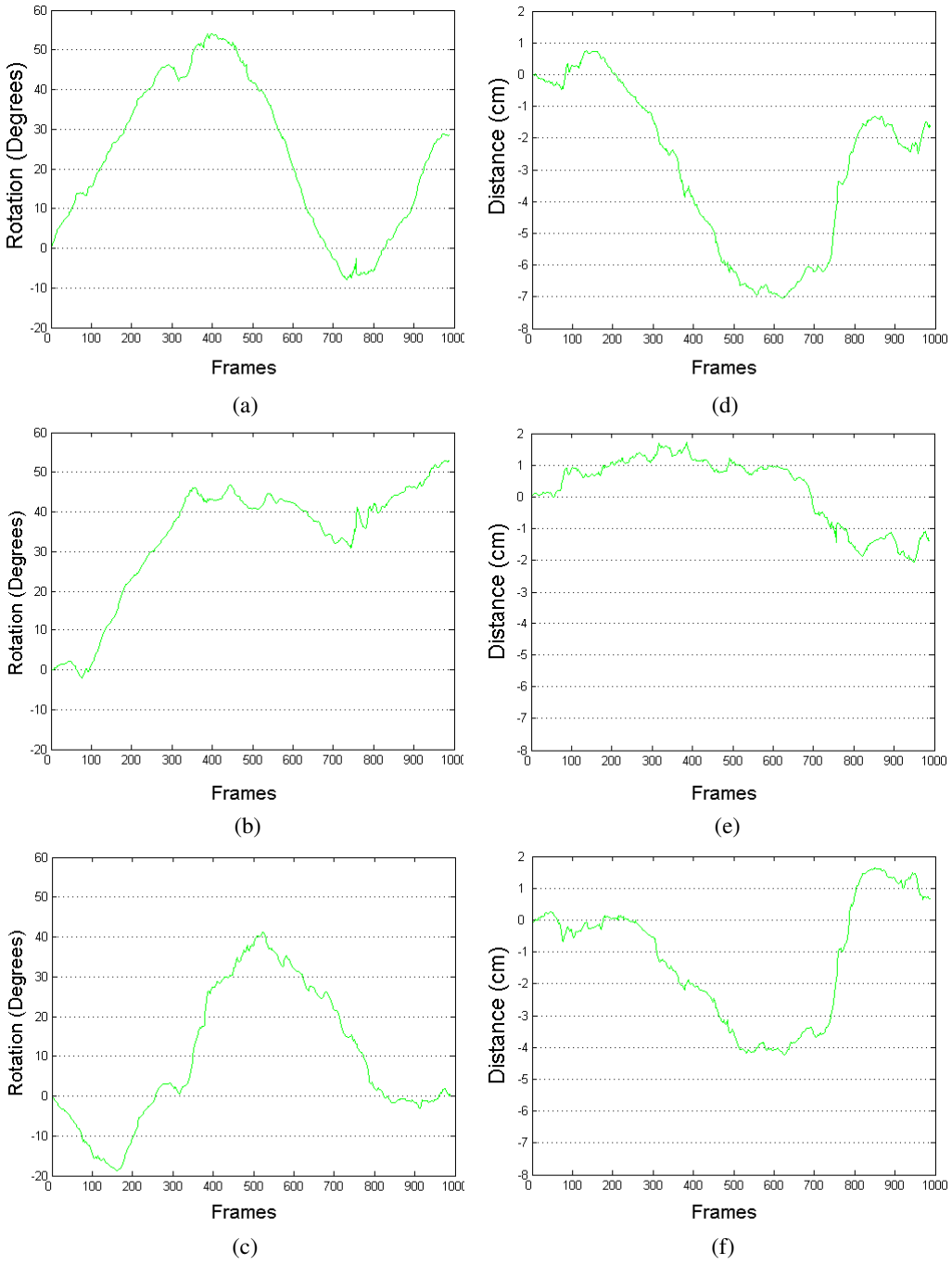


Figure 5.7 Results for the second *in vivo* experiments with the SLAM framework for (**a-c**) rotation around the X, Y, and Z axes and (**d-f**) translation along the X, Y, and Z axes.

5.3.2 Quantitative Validation

In order to provide quantitative validation of the proposed SLAM framework, further experiments were performed with both simulated and phantom data.

5.3.2.1 Simulated Experiments

A simulation, with a virtual stereo camera moving through a texture-mapped 3D world, was rendered and used to validate the SLAM framework. The simulator provides known ground truth data of camera motion within the virtual environment, thus allowing for detailed quantitative evaluation. In this experiment, the motion of the camera was constrained such that the inter-frame pixel motion did not exceed 20 pixels. This is consistent with observations from *in vivo* data during navigation, however, rapid camera motion can occur.

The virtual stereo camera was parameterised to replicate a stereo-laparoscope with similar intrinsic and extrinsic properties and a baseline of 5mm. The virtual environment contains a plane, which is textured with an MIS image. The image was acquired from a robot-assisted procedure involving the liver. This provides a realistic image rendering upon which to perform region tracking and SLAM, as shown in **Figure 5.8**. A planar model of the environment is used for simplicity, however, the proposed method is not restricted to this environment and is capable of mapping complex geometric structures. It is important to note: the simulation does not fully replicate the MIS environment because it does not model specular highlights, changes in illumination, image noise, or calibration error.

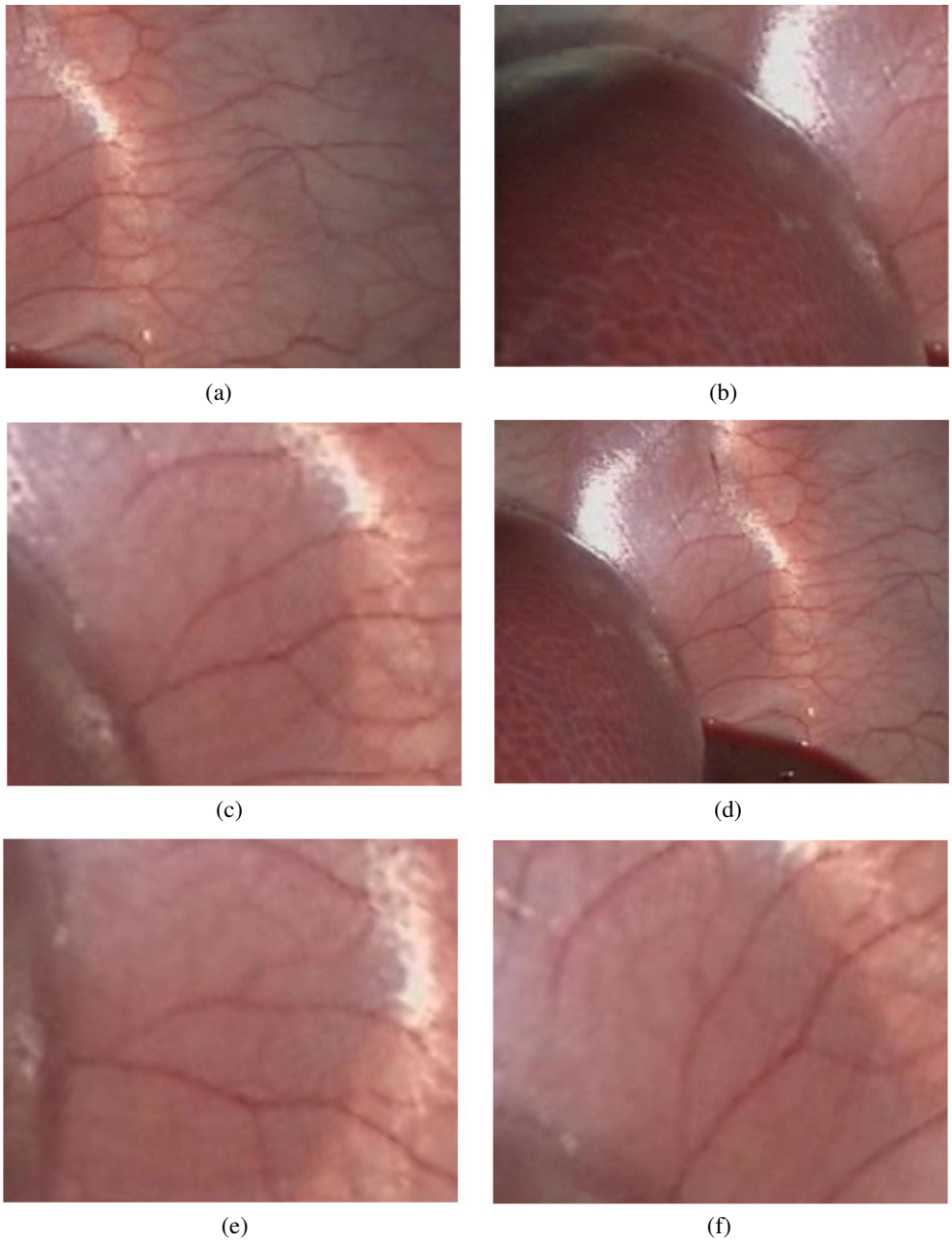


Figure 5.8 Images from simulated data illustrating translation along the X axis **(a-b)**, translation along the Z axis **(c-d)** leading to a change in scale. **(e-f)** Shows rotation around the Z axis.

Quantitative validation of the SLAM framework on the simulated data is shown in **Figure 5.9**. **Figure 5.9 (d-f)** demonstrates the results for camera translation with the ground truth data shown in red and the estimated position by SLAM in green. The virtual camera is first navigated left and right, then up and down and finally along the optical axis. The average error was 0.4, 0.22, and 0.1 cm in the X, Y, and Z axes, which corresponds to standard deviations of 0.28, 0.22, and 0.09 cm, respectively. With respect to the total translation in the X, Y, and Z axes, this error represents average errors of 2.3%, 1.5%, and 0.5% of the total movement.

These results demonstrate the accuracy of the SLAM framework on simulated data. The method can accurately localise the camera's position. The element of error involved is relatively small. Sharp changes in direction of translation can cause larger errors. This is not surprising as the constant velocity, constant motion model does not model acute changes in direction. It is proven that navigating along the Z axis (away and towards the tissue model) introduces small errors into the X and Y estimation. These errors are a result of higher uncertainty along the Z position of the features caused by the small baseline of the stereo cameras.

Figure 5.9 (a-c) displays the virtual camera rotating in pitch, yaw, and roll around the X, Y and Z axes, with an average error of 1.34°, 0.8°, and 0.295°, respectively. The associated standard deviation is 1.57°, 0.75°, and 0.32° respectively. These errors represent 2.23%, 1.33%, and 0.49% of the total rotation observed. Rotating around the Z axis introduces small errors in localisation accuracy along the X axis. This is caused small localisation inaccuracies of features in the image space.

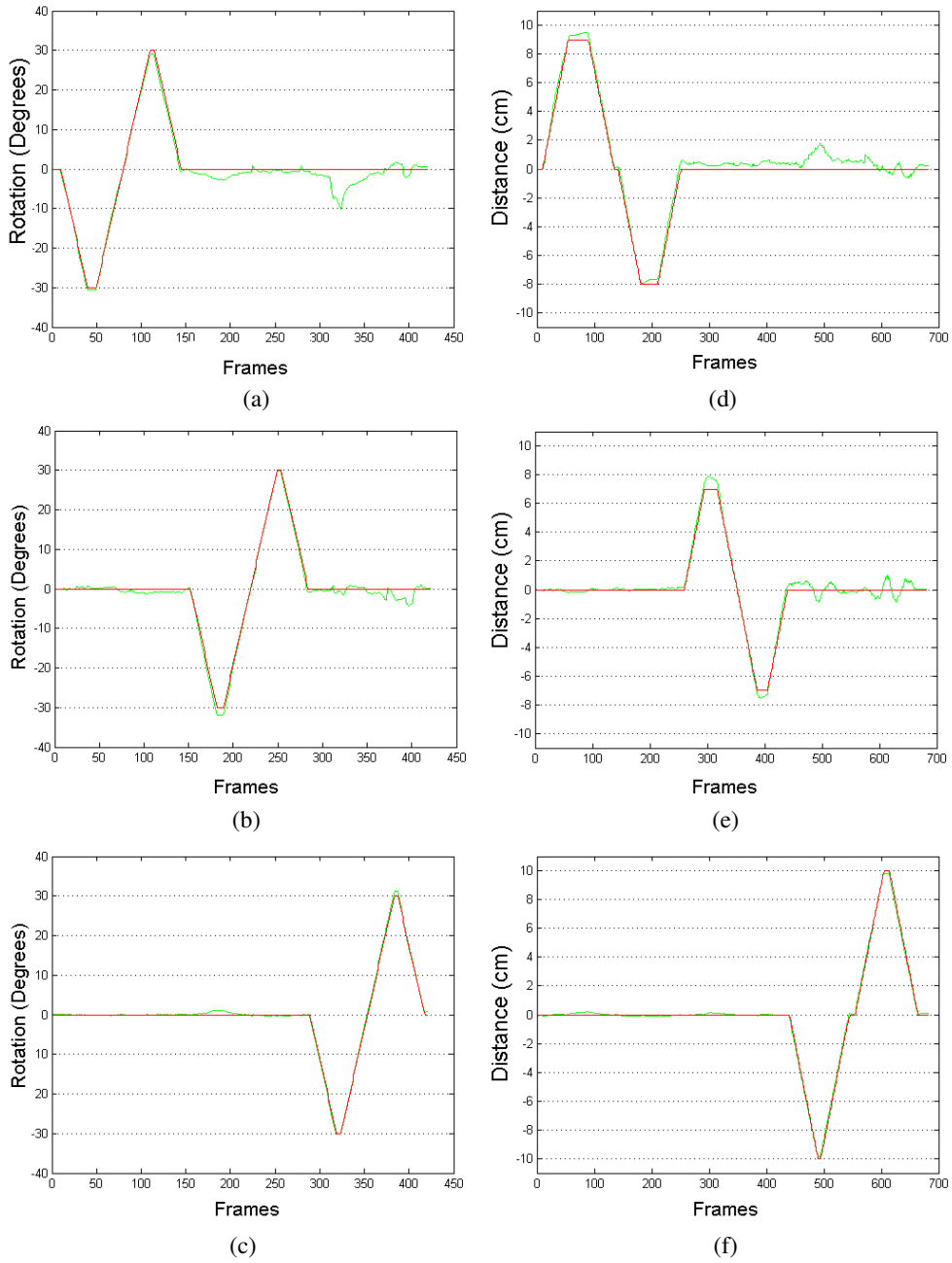


Figure 5.9 Quantitative analysis of the laparoscopic camera motion for simulated data. The SLAM estimated position is shown in green, and the ground truth is shown in red for **(a-c)** rotation around the X, Y, and Z axes and **(d-f)** translation along the X, Y, and Z axes.

5.3.2.2 Phantom Experimental Set-up

The proposed SLAM system was validated on phantom data using a custom-made stereo fibrescope, as illustrated in **Figure 5.10**. It is a bespoke, laparoscopic imaging system with twin, 10,000 pixel coherent fibre bundles (590µm diameter, length 1.5 m, minimum bend radius 25mm) [255]. On the end of each fibre image guide, a graded index (GRIN) lens (Grintech GmbH) is cemented. The GRIN lens has a 0.5mm diameter and is capable of imaging an area of $35\times35mm^2$ at a working distance of 20mm. At the distal tip, the fibre image guides are clamped into place with a baseline of 3.8mm. A micrometre precision stage is used to mount the fibres, both of which are imaged onto a single CCD camera (UEye, UI-2250-C/CM) with 100mm focal length using an achromatic ×10 microscope objective lens.

The stereo fibrescope presented a number of challenges in the application of SLAM. The system has a small working distance, small field-of-view, and is low-resolution (only 10,000 pixels). The captured images exhibit a honeycomb structure, which is removed using the method described above. The small baseline makes stereo reconstruction difficult, and the fibre image guides are sensitive to changes in lighting.

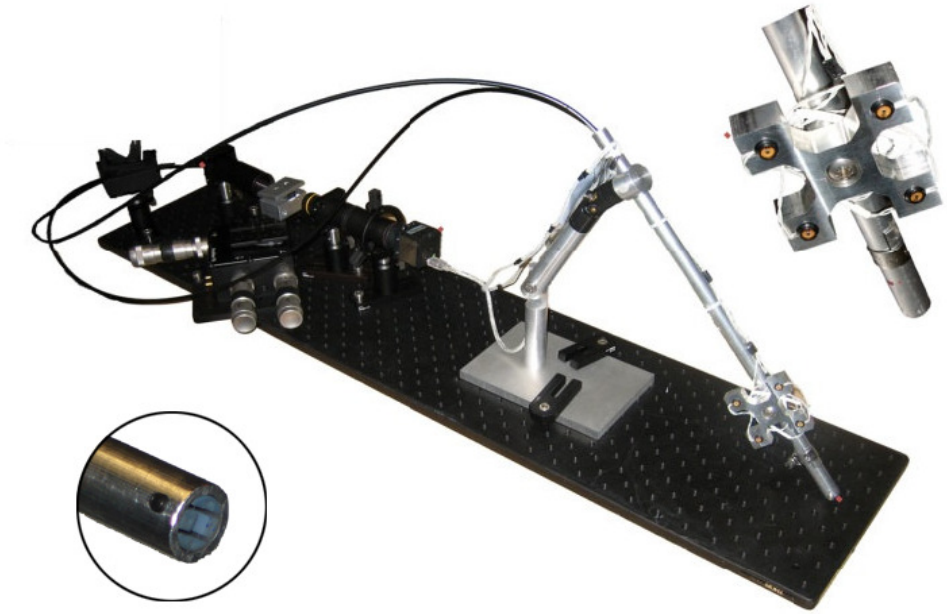


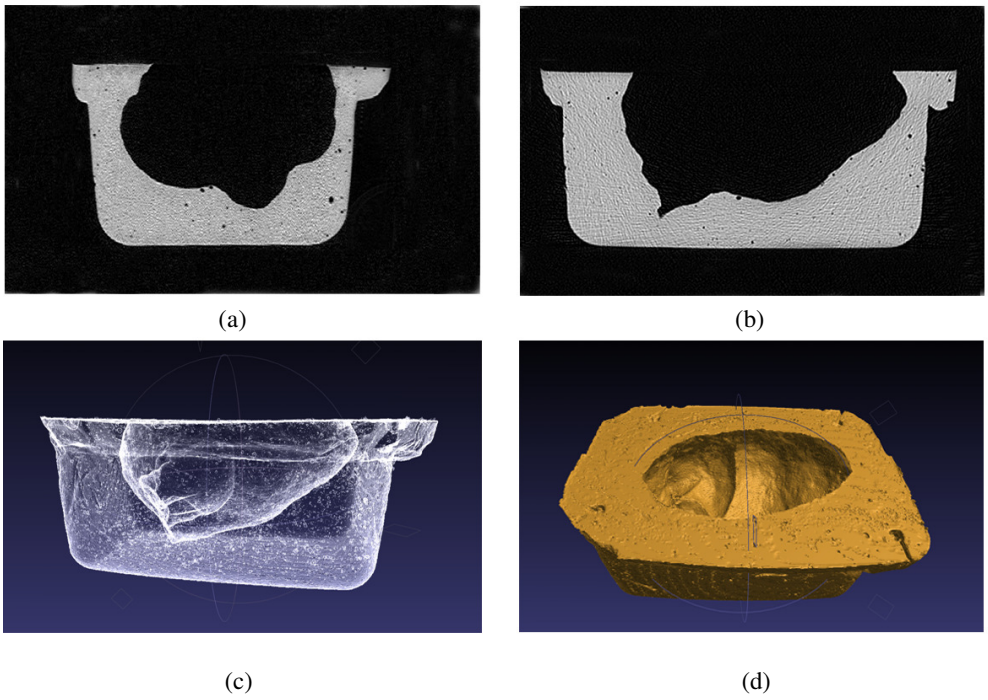
Figure 5.10 Image showing the custom-made optical configuration of the stereo fibrescopic system. The optical set-up includes the fibre mount, objective lens, and camera. The rigid body, embedded with optical markers used for validation, is shown in the top right. A close-up of the laparoscope tip is shown in the bottom left.

In order to validate the accuracy of the camera motion estimated by the SLAM algorithm, a rigid body with four active optical tracking markers (Northern Digital Inc, Ontario, Canada) was created. The rigid body, shown in **Figure 5.10**, is comprised of four optical markers on a plane with an attachment enabling it to be rigidly fitted to either a da Vinci laparoscope or the custom made fibrescope. The four markers define the Rigid Body co-ordinate system at the origin of one of the markers. The external optical tracking system is capable of measuring the position $\mathbf{Tr}^{\mathbf{W}}$ and orientation $\mathbf{TR}^{\mathbf{W}}$ of this rigid body with respect to the world co-ordinate system constantly. Hand-eye calibration was performed using a technique similar to [26]. This provides the transformation $\mathbf{Hr}^{\mathbf{tl}}$ and rotation $\mathbf{HR}^{\mathbf{tl}}$ between the left camera centre and the rigid body coordinate system. The measured position $\mathbf{Cr}^{W}$ and orientation $\mathbf{CR}^{W}$ of the camera centre with respect to the world co-ordinate system can be computed using the following transformation:

$$\mathbf{Cr}^{W} = \mathbf{Tr}^{\mathbf{W}} + \mathbf{TR}^{\mathbf{W}}\mathbf{Hr}^{\mathbf{tl}} \tag{5.25}$$

$$\mathbf{CR}^{W} = \mathbf{TR}^{\mathbf{W}}\mathbf{HR}^{\mathbf{tl}} \tag{5.26}$$

This process provides the rotation and position of the left camera centre for validation.



(a) (b)

(c) (d)

Figure 5.11 Ground truth map data. **(a-b)** A CT of the silicon phantom. **(c-d)** The CT is segmented and meshed to create surface models.

The SLAM algorithm was validated on a phantom data-set with known ground truth using the custom fibrescope. The ground truth data for the position and orientation of the laparoscope was provided using the optical tracking configuration described earlier. The 3D geometry of the phantom was obtained from a Computed Tomography (CT) scan as illustrated in **Figure 5.11 (a-b)**. The phantom was constructed using silicone and coated with latex paint to simulate specular reflections and tissue texture. The phantom was embedded with CT visible markers, which were easily identified and segmented in the CT scan. The location of each marker was identified during the data acquisition phase using a stylus. Attached to the stylus was a second rigid body which enabled the CT visible markers to be registered to the world coordinate system. This allows the CT visible and the CT scan to be registered to the camera coordinate system, thus providing the ground truth data for the geometry of the phantom.

The CT scan was meshed to create a surface shown in **Figure 5.11 (c-d)**, and this surface is compared to the point map generated by the SLAM algorithm. This requires that for each 3D point in the SLAM map, a corresponding point is identified on the surface of the CT model. Regions of interest, detected in the laparoscopic images and tracked in the SLAM framework, were projected into the registered CT model from the camera's ground truth position (provided by the optical tracking system). The projected ray was traced through the 3D CT model until it intersected a surface. This point of intersection was taken as the corresponding point in the CT surface.

5.3.2.3 Phantom Results

Quantitative analysis of SLAM on phantom data is provided in **Figure 5.12**. This graph shows the translational motion of the laparoscope decomposed into motions along the X, Y, and Z axes for 1,400 frames. The ground truth data is shown in red, and the position of the laparoscope estimated by SLAM is shown in green. The absolute error in the X, Y, and Z axes was 0.19 cm, 0.07 cm, and 0.17 cm, respectively. The graph demonstrates the accurate recovery of laparoscopic motion using SLAM. To further illustrate the motion accuracy derived, **Figure 5.13** displays the trajectories of the SLAM estimate and the ground truth in 3D. The SLAM estimate is shown in green, and the ground truth is shown in blue. **Figure 5.13 (e-h)**, illustrating the trajectory over time. The position of the laparoscope is indicated using a green cube (SLAM) and blue cube (ground truth). The generated SLAM map is shown in **Figure 5.13** as a surface.

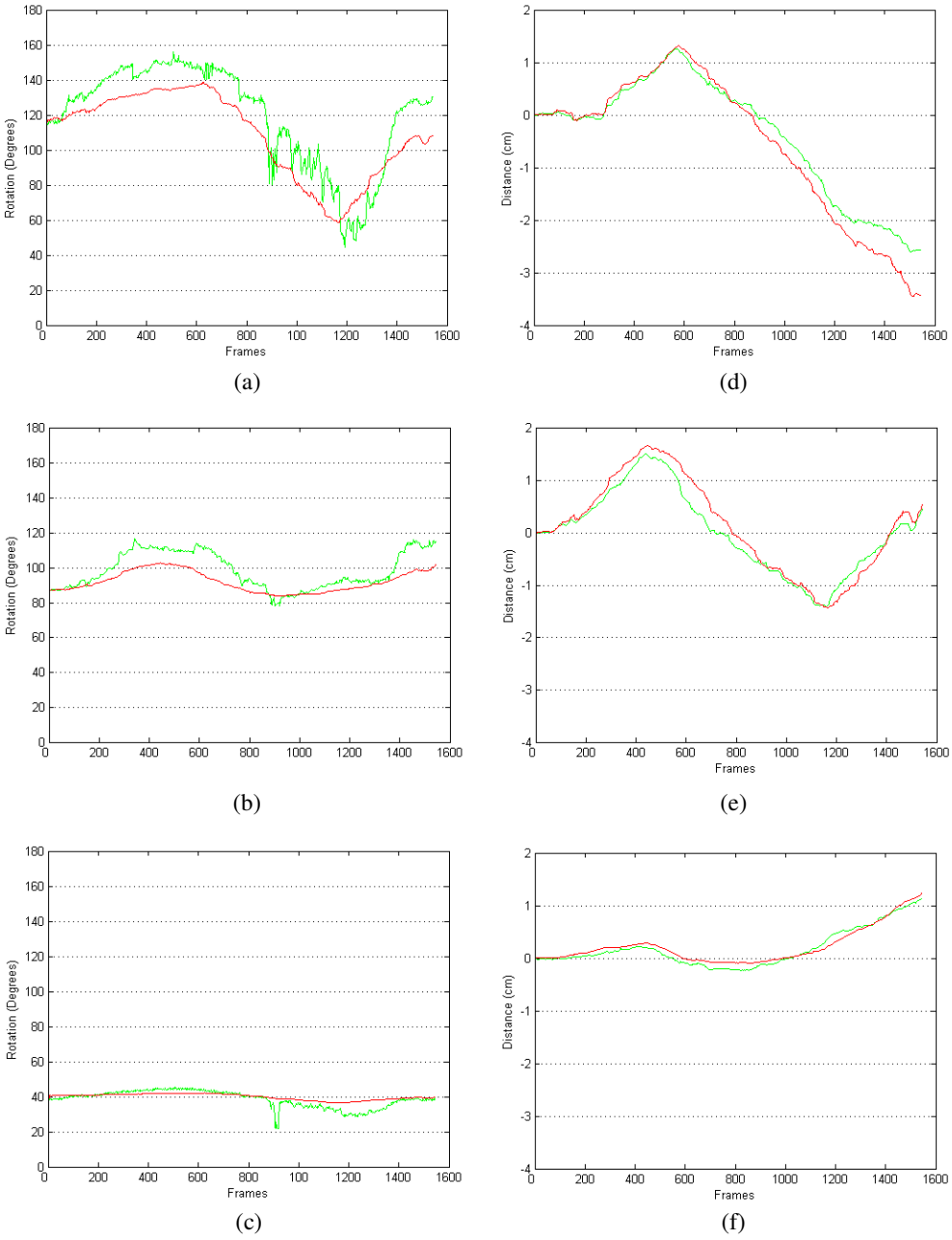


Figure 5.12 Phantom data. Quantitative analysis of the camera trajectories decomposed into individual rotations and X, Y and Z translations. The ground truth is shown in red and the SLAM recovered camera position is shown in green for the **(a-c)** rotation around the X, Y, and Z axes and **(d-f)** translation along the X, Y, and Z axes.

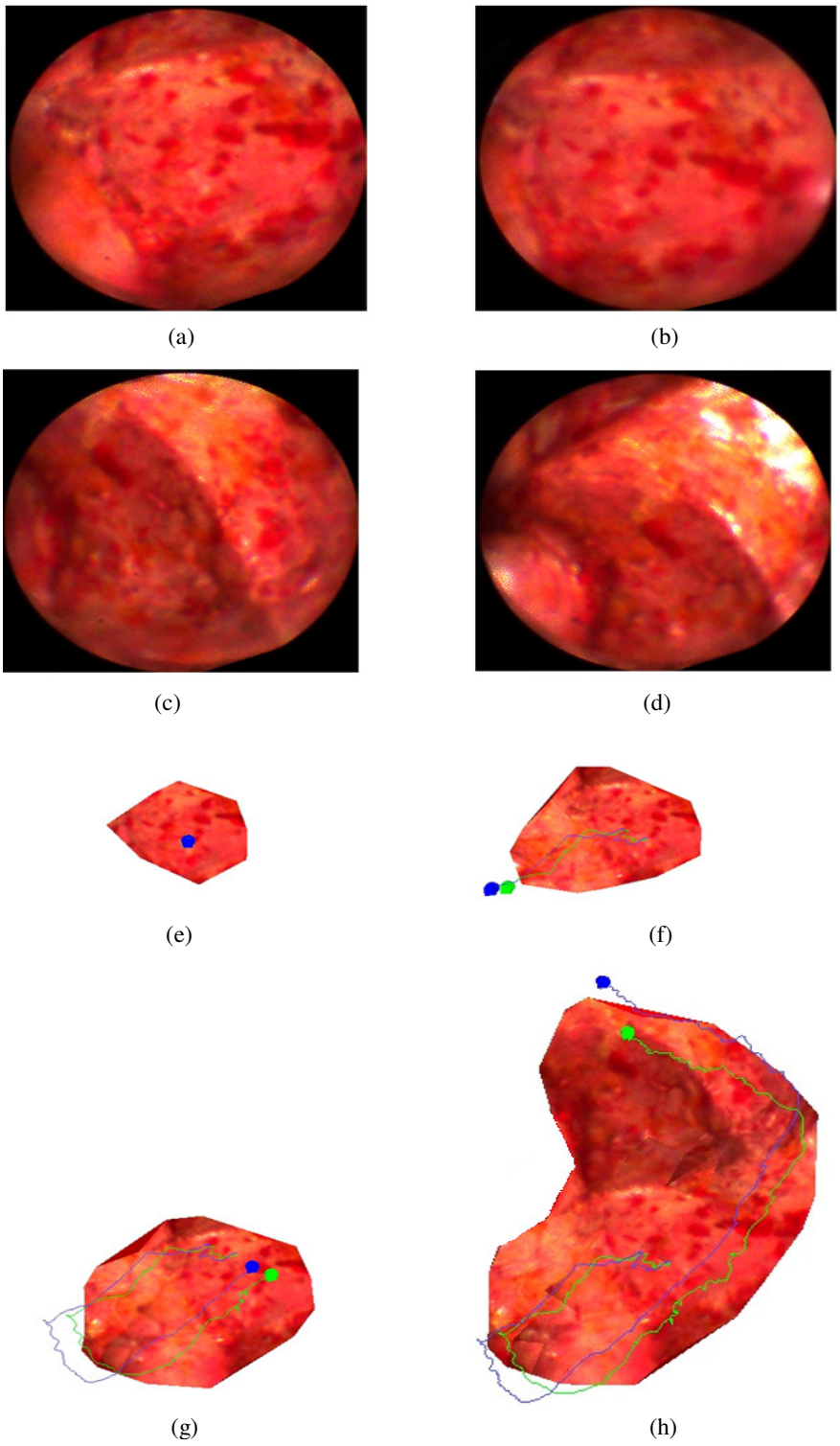


Figure 5.13 Phantom Data. **(a-d)** Example images from the fibrescope. **(e-h)** The SLAM recovered 3D textured surface model and camera position, ground truth trajectory (blue) and SLAM estimated camera trajectory (green).

The 3D surface reconstruction of the phantom data is evaluated in Figure 5.14. The figure shows the 3D ground truth surface data of the phantom on the left hand side and the 3D surface generated by the SLAM algorithm on the right. The 3D surface reconstructions are shown from three different views. The recovered SLAM surface is visually similar in shape to the ground truth CT surface in scale and orientation indicating accurate reconstruction. Local inaccuracies exist in the SLAM surface reconstruction. These inaccuracies are attributed to the use of a sparse map and surface meshing. The use of a sparse SLAM map means only a small number of data points are used to represent the surface. The surface is interpolated between these points by performing a Delaunay triangulation. This is a simple and relatively crude method of interpolation. More accurate results can be obtained by incorporating dense feature matching into the map. Quantitatively, the average reconstruction errors for all points in the SLAM map was 0.2 cm, 0.13 cm, and 0.29 cm in the X, Y, and Z axes respectively. During data acquisition, the surface was approximately 3.5 cm from the camera. The larger reconstruction error in the Z axis is a result of the small base-line of the stereo camera (3.8mm) and low image resolution.

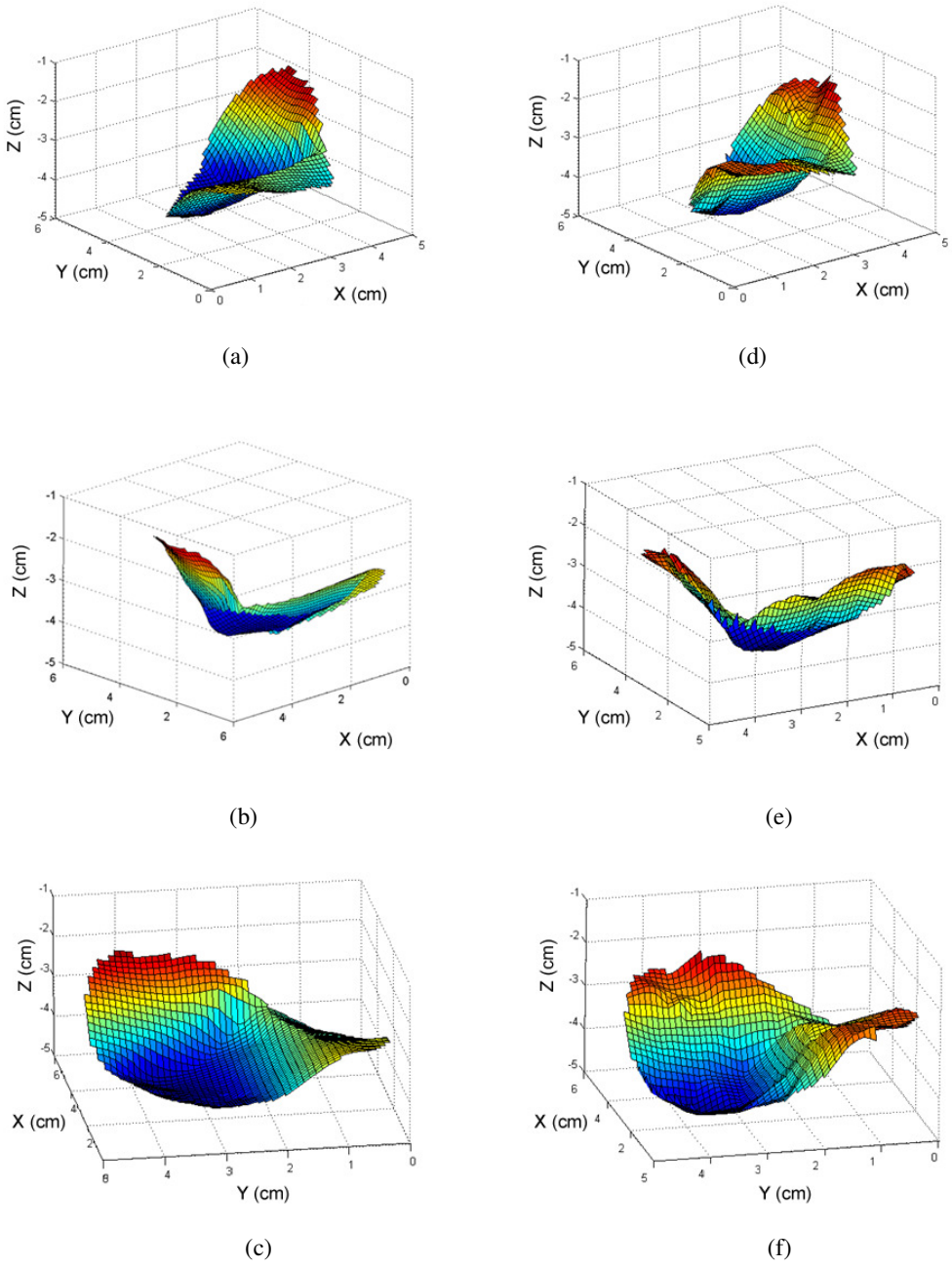


Figure 5.14 A comparison of reconstructed 3D surface generated from CT ground truth data **(a-c)** and by SLAM **(e-f)**.

5.4 Discussions and Conclusions

In this chapter, a solution is described to the 3D mapping of soft-tissue using a moving laparoscopic camera. This is achieved while simultaneously estimating the laparoscopic camera's position and pose using a SLAM framework. The SLAM system is based on an EKF implementation. The experiments show that the EKF formulation is sufficient for creating localised 3D maps. The framework utilises the region tracking algorithm outlined in **Chapter 4** to ensure robust tracking in a challenging environment. The proposed method has been demonstrated with both standard, commercially available laparoscopes and a custom built fibrescope with low image resolution and small field-of-view. Pre-processing is used to remove honeycomb artefacts caused by the fibrescope. It is shown that a *constant velocity, constant angular velocity* motion model is suitable for both smooth hand-held or robotically-controlled laparoscopes.

The method described in this chapter was quantitatively validated using a simulated data-set with real MIS textures. Additional quantitative validation was performed using a silicon phantom. It is shown that the SLAM approach can incrementally build long-term maps, and it is capable of loop-closing (*i.e.*, returning to previously visited regions without drift). This was demonstrated on soft-tissue with sparse features under a point light source illumination condition involving specular highlights. The main assumption of this work is that tissue under consideration is relatively static. This is not a realistic assumption for MIS. Soft-tissue is highly deformable, and dynamic scene motion must be considered. In the following chapters, the application of SLAM in situations involving dynamic tissue motion will be investigated. **Chapter 7** consider some of the practical applications of the proposed SLAM framework to selected MIS settings.

Chapter 6

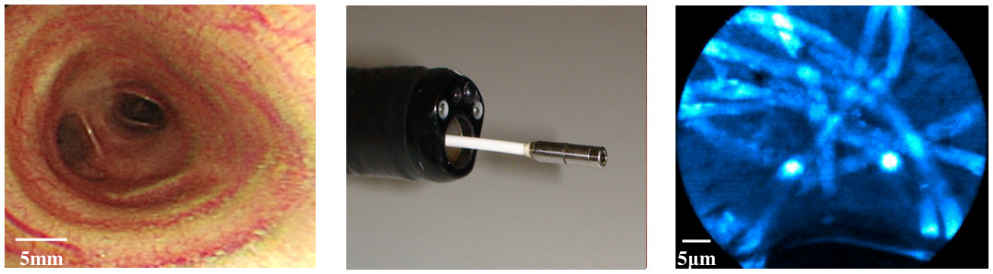
Applications of SLAM to MIS

In **Chapter 5**, the use of a SLAM framework was proposed for soft-tissue mapping and laparoscope localisation. Its performance was quantitatively evaluated on simulated and phantom data, and demonstrated on *in vivo* sequences. The system is based on the static environment assumption. Although this assumption is difficult to satisfy in MIS, there are selected anatomies where the assumption holds. This chapter considers the practical application of SLAM using two clinical examples - Optical Biopsy Mapping and Dynamic View Expansion.

6.1 Optical Biopsy Mapping

The development of new biophotonics and surgical instrumentation has been motivated by the quest to provide *in vivo*, real-time tissue characterisation and functional mapping during MIS. Biophotonic probes – miniaturised to fit down the instrument channel of standard endoscopes, are capable of revealing cellular and sub-cellular tissue microstructures, thus allowing excision-free *optical biopsy*, as shown in **Figure 6.1**, Technologies, such as miniaturised confocal laser scanning microscopes, have been used in conjunction with the application of contrast agents for the detection of colorectal adenomas, disruption in the pit pattern of the colon, angiogenesis, and neoplasia in Barrett's oesophagus [256]. Such technologies have also been used to detect malignant disruption of the bronchial basement membrane using elastin auto fluorescence [257] without a contrast agent. Optical Coherence Tomography (OCT), two photon excited

fluorescence, and high magnification endoscopy [258] are some of the other techniques enabling microscopic detection and characterisation of tissue. Successful clinical trials using techniques that acquire detailed spectroscopic information have been carried out for cancer detection (e.g., using the time- or wavelength-resolved fluorescence or Raman properties).



(a) (b) (c)

Figure 6.1 (a) A typical endoscopic white light image of the bronchus used for navigation, **(b)** the relative configuration of a confocal fluorescence probe when inserted through the instrument channel of a standard endoscope, and **(c)** a typical microconfocal fluorescence image showing the microstructure of a sample.

Optical biopsy has the potential to facilitate paradigm shift in current clinical practices. **Figure 6.2** schematically illustrates the work-flow for traditional biopsy analysis and potential work-flow for optical biopsy analysis. The time frame for current work-flow is significant and requires the patient to visit the hospital multiple times. The biopsy sample sent away for diagnosis also requires the interaction of several healthcare system stakeholders. Optical biopsy technology offers the potential for *in situ, in vivo* diagnosis. This enables the surgeon to make immediate decisions regarding patient management, thus dramatically reducing the time between biopsy and intervention. This is not only beneficial to patient health, but it is efficient, and potentially less expensive for health care providers, as it reduces the number of stakeholders involved in the diagnosis.

The practical *in vivo* applications of these techniques are limited by the size of the region the probe can biopsy. This prevents large area surveillance and integrated functional mapping. These techniques provide only a small, localised region, whilst the organs of interest may require a large surface area to be surveyed. Unlike traditional biopsy, which may be marked with a scar or with ink, optical biopsies leave the tissue unmarked. This makes tracking the biopsy sites, for the purposes of retargeting and mapping, a challenging task. In addition, the endoscope, controlled by the surgeon, is mobile, and

the biopsy site can move in and out of the camera's field-of-view. A system is presented in [259], which has been developed in parallel with the work in this thesis. The authors' use the epipolar geometry between monocular endoscopic images to estimate the position of a biopsy site relative to the camera.

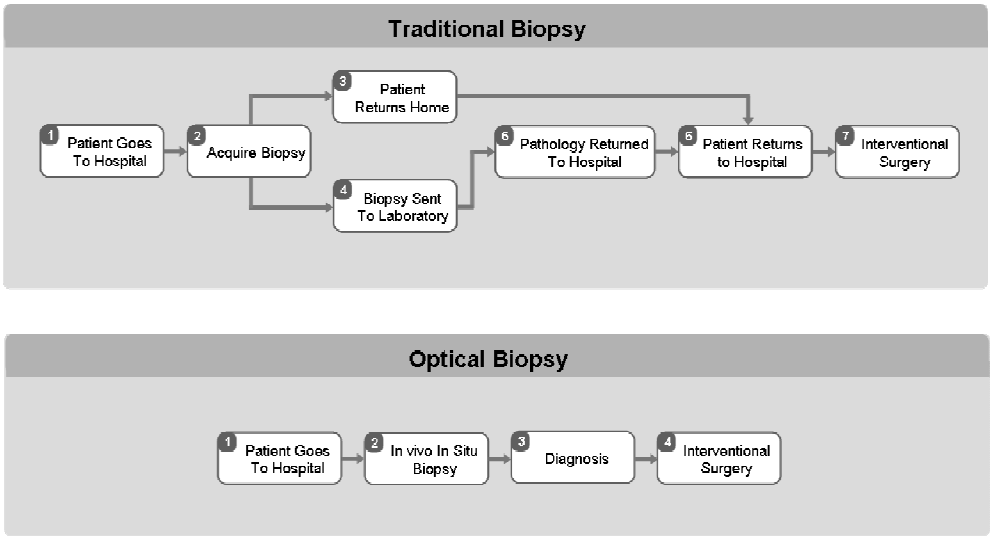


Figure 6.2 Top - the clinical work-flow of traditional biopsy. Bottom – a potential new clinical work-flow that may be facilitated by optical biopsy.

Tool tracking and registration between image modalities are a prerequisite to facilitating intra-operative guidance and the augmentation of the endoscopic image with functional imaging data. As discussed in **Chapter 2**, current approaches to instrument tracking assume the use of rigid instruments and availability of optical markers [260], which are inappropriate for flexible instruments such as endoscopes or biopsy probes. Electromagnetic tracking systems may be employed however these are susceptible to interference and two systems will be required to track the probe and endoscope.

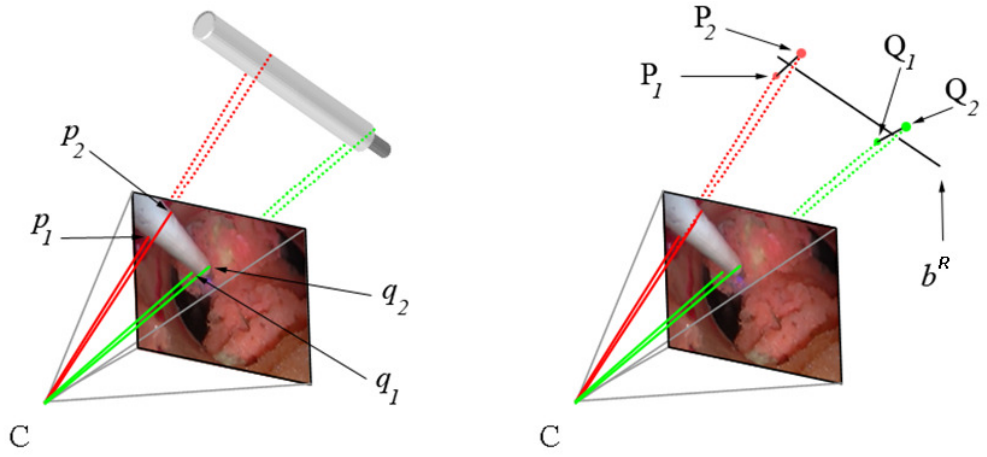
This work proposes the use of SLAM to track the position of the scope and to estimate the biopsy site relative to the scope by tracking the biopsy tool in the endoscopic image. When the biopsy occurs, the optical probe is typically stationary: the probe must be in contact with the tissue. The biopsy site can be estimated by tracking the tip of the probe. The position of the biopsy site is integrated explicitly into the SLAM probabilistic map, thus creating a 3D model of the tissue surface and spatio-temporally tracked biopsy sites. The endoscopic image is augmented with the position of the biopsy sites by re-projecting

the 3D position back into the image plane. Validation was performed on phantom data with known ground truth.

6.1.1 Probe Tracking and Biopsy Site Estimation

Estimating the position of the biopsy site is difficult because the probe obscures the site during the biopsy procedure and leaves no visual mark on the tissue surface. Since the probe must be placed in contact with the tissue during optical biopsy acquisition, tracking the tip of the probe enables the location of the biopsy site to be inferred. The flexible, optical probe is typically introduced via the instrument channel while holding the endoscope stationary. The current approach to surgical instrument tracking discussed in **Chapter 2** may be suited to estimating the position of the probe, however, the physical constraints (flexible probe, small working area) make such estimation challenging. An alternative approach is to track the probe in the endoscopic image and estimate its position using image based techniques. The additional benefit of this approach is: tracking and visualisation are performed in a shared co-ordinate space, thus removing the requirement for additional equipment and the registration of images across multiple data streams.

This chapter uses an approach to tool tracking outlined in [261] in order to track the white shaft of the tool. No changes are made to the colour of the imaging probe in this technique, and no markers are attached. The approach is based on background subtraction and colour segmentation. The position of the shaft of the probe is estimated using the Hough transform and the eigenvectors and eigenvalues from the moment of inertia. This method is used to estimate four points on the shaft; two at the join between the shaft and the metal tip of the probe (q_1 and q_2) and two at an arbitrary distance along the shaft (p_1 and p_2) as shown in **Figure 6.3**. These points are used in conjunction with prior knowledge of the probe's geometry in order to estimate the 3D position of the probe tip relative to the camera.



(a) (b)

Figure 6.3 Estimation of the biopsy site via model-based instrument tracking. **(a)** The points on the shaft of the tool are estimated in 3D relative to the camera centre C. **(b)** The orientation and 3D position of the tool are estimated. A geometric model is used to extrapolate the position of the tip and infer the biopsy site in 3D.

The 3D position of the probe tip is estimated using a semi-model based approach. It is assumed the short section of the probe extending out the instrument channel can be modelled as rigid and that prior knowledge of the probe's geometry is available. Given these assumptions, a vector is computed that describes the orientation and position of the shaft in 3D space. The vector is defined by two points in 3D space, which are the mid point of P_1, P_2 and the mid point of Q_1, Q_2 and are computed using the image points p_1, p_2, q_1, q_2 and the known physical width of the probe such that a 3D point is represented as

$$P_1 = (X, Y, Z) = \left(\frac{u_0 - p_{1u}}{fk_u} P_{1Z}, \frac{v_0 - p_{1v}}{fk_v} P_{1Z}, P_{1Z} \right) \tag{6.1}$$

where fk_u and fk_v represent the camera's focal length, and u_0 and v_0 are the principal points. Assuming $P_{1Z} = P_{2Z}$ the P_{1Z} can be computed using the known width of the probe W where

$$W^2 = \left({p_{1u}}^2 - {p_{2u}}^2 \right)^2 \left(\frac{P_{1Z}}{fk_u} \right) + \left({p_{1v}}^2 - {p_{2v}}^2 \right)^2 \left(\frac{P_{1Z}}{fk_v} \right) \tag{6.2}$$

and Q_1, Q_2 can be computed using the same equations. Given a vector describing the orientation and position of the shaft of the probe, the geometric model of the tip of the probe is simply added to the vector to estimate the tip of the probe in 3D $\mathbf{b}^R$ as illustrated in **Figure 6.3 (b)**.

6.1.2 Global Biopsy Mapping with SLAM

The SLAM system described in the previous chapter is used to track the position of the laparoscopic camera and build a map of the environment. Each time a new biopsy is taken, the biopsy site $\mathbf{b}^R$ is incorporated into the SLAM map. It is first estimated in the camera coordinate system using the approach described in the previous section. It is then transformed into the SLAM world coordinate system using

$$\mathbf{b}^W = \mathbf{R}^W \mathbf{b}^R + \mathbf{r}^W \tag{6.3}$$

where $\mathbf{b}^W$ is the biopsy site in the world coordinate system and $\mathbf{R}^W$ and $\mathbf{r}^W$ are the orientation and position of the camera in the global SLAM coordinate system.

In this study, the 3D position of the biopsy site is defined when the surgeon activates the foot pedal controlling the optical biopsy probe. This 3D position is not directly measured or observed again for two reasons; 1) the actual biopsy site on tissue surface is usually occluded by the probe when the biopsy is taken; and 2) there may not be any salient regions of interest at or around the biopsy site to track. In the case of the latter, a simple 2D tracking approach would fail. The strength of the proposed SLAM based approach lies in the position of the biopsy site, which can be updated without direct measurement. This is facilitated by the co-variance matrix, which models the uncertainty of all the map and biopsy positions. The i^{th} biopsy site $\mathbf{b}_i^W$ is inserted into the standard SLAM state vector, and the co-variance matrix P is updated. The co-variance matrix is updated in **Equation (6.4)** with the partial derivatives $\partial \mathbf{b}_i / \partial \mathbf{x}_v$ of the biopsy site with respect to the camera position, as well as the measurement model $\partial \mathbf{b}_i / \partial \mathbf{h}_i$ and measurement noise $\mathbf{R}$.

$$\mathbf{b}_i^W = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{6.4}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy_1} & \mathbf{P}_{xx}\frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}^T \\ \mathbf{P}_{y_1x} & \mathbf{P}_{y_1y_1} & \mathbf{P}_{y_1x}\frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}^T \\ \frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xx} & \frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xy_1} & \frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}\mathbf{P}_{xx}\frac{\partial \mathbf{b}_i}{\partial \mathbf{x}_v}^T + \frac{\partial \mathbf{b}_i}{\partial \mathbf{h}_i}R\frac{\partial \mathbf{b}_i}{\partial \mathbf{h}_i}^T \end{bmatrix} \tag{6.5}$$

where $\mathbf{x}_v$ is the position and pose of the endoscope, and y_i is the i^{th} feature in the map.

The position and uncertainty of the biopsy site are then correlated to the rest of the features in the map and the camera position. Sequential map building is illustrated in **Figure 6.4**. This demonstrates how the biopsy sites, camera, and 3D features are correlated and how this information is temporally updated. At the moment when the biopsy site is observed, the relative position of the biopsy site to surrounding features is well-defined, see **Figure 6.4 (b)**, but the uncertainty of the camera's position may be high.

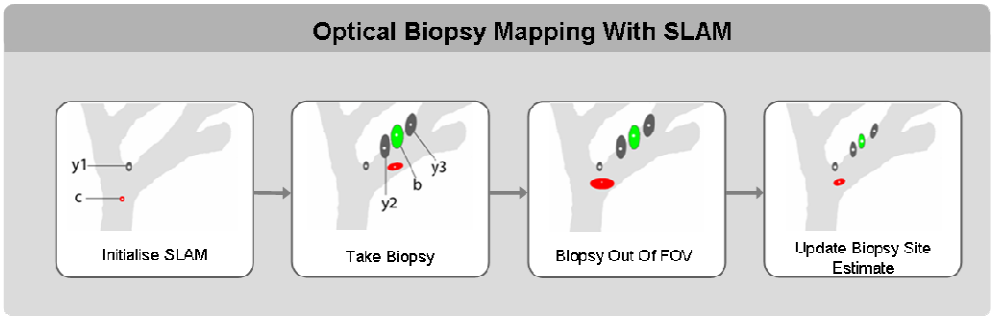


Figure 6.4 (a-d) Schematic representation of SLAM's sequential probabilistic mapping updates. The laparoscopic camera's position c is shown in red. An ellipse represents its spatial uncertainty. The tissue is shown in light grey. Map features y1, y2, and y3 are represented in dark grey, and the biopsy site b is shown in green. **(a-d)** shows the sequential progression where **(a)** c measures y1 with low uncertainty, **(b)** c is navigated to a new position with growing uncertainty. Features y2 and y3 are measured and biopsy b is taken. **(c)** c is navigated close to y1 and positional uncertainty increases. **(d)** Feature y1 is measured and the estimated position of c is improved which results in improved estimate of b as it is correlated to c.

During subsequent camera movement during exploration, the camera re-measures these surrounding features in the SLAM map, see **Figure 6.4 (d)**. As the position estimation of the camera improves, (*i.e.* with reduced uncertainty), the 3D position estimation of the biopsy site also improves due to its correlation with the estimated position of the camera. In order to facilitate intra-operative guidance and augmented reality, the biopsy sites $\{\mathbf{b}_1^W ... \mathbf{b}_i^W\}$ are re-projected from the 3D points onto the camera plane. This uses the estimated camera position from SLAM and the intrinsic camera parameters.

6.1.3 Experimental Set-up

The application was validated on a silicon phantom of the airway using a stereo laparoscope. The phantom was coated with acrylic paint to provide inter-reflection and realistic texture. Sponge cell structures were attached to the internal surface of the phantom in order to enable optical biopsies taken using a confocal fluorescence endoscope system (Cellvizio, Mauna Kea Technologies, Paris). The accuracy of the re-projected biopsy sites in the image space was measured over time to validate the system. Ground truth data was collected using an optical tracking device (Northern Digital Inc, Ontario, Canada) and an experienced observer. The laparoscope's position $\mathbf{r}_{gt}^W$ and orientation $\mathbf{R}_{gt}^W$ was obtained using the optical tracking approach described above. An experienced observer who manually identified the biopsy sites on the stereo images obtained the ground truth of the 3D position of the biopsy sites. The camera's intrinsic and extrinsic parameters were used to compute the 3D position $\mathbf{b}_{gt}^R$ of the biopsy site relative to the camera. This enables its position in the world coordinates system $\mathbf{b}_{gt}^W$ to be determined as $\mathbf{b}_{gt}^W = \mathbf{R}_{gt}^W * \mathbf{b}_{gt}^R + \mathbf{r}_{gt}^W$. The ground truth was obtained at each subsequent frame by projecting the biopsy site $\mathbf{b}_{gt}^W$ into the ground truth camera position $u = u_o - fk_x(\mathbf{b}_{gt}^R x / \mathbf{b}_{gt}^R z)$ and $u = u_o - fk_y(\mathbf{b}_{gt}^R y / \mathbf{b}_{gt}^R z)$, where fk_x and fk_y are the focal length, and u_o and v_o are the principal point.

6.1.4 Results

The approach to optical biopsy mapping was validated on phantom data with known ground truth. During a two-minute video sequence, the laparoscope was used to explore the airway. The laparoscope was navigated to four separate areas where biopsies were taken. During navigation the biopsy probe was not in the field-of-view. Once an area for biopsy was identified, the probe was introduced into the field-of-view in order to acquire a biopsy image. Six biopsies were collected in total. Biopsy sites 1-3 were collected in separate areas, and biopsy sites 4-6 were collected from the same area. Biopsies were re-targeted during the exploration.

The optical biopsy map is presented to the surgeon as an augmented reality visualisation. The accuracy of this visualisation in the image plane is also quantitatively analysed. Once the optical biopsy has been acquired, it is constantly re-projected onto the image plane - assuming it is visible in the current field-of-view. The re-projected point in the 2D image is then compared to the ground truth position in the image. This provides a quantitative metric for assessing the accuracy of biopsy mapping. Results for the six biopsy sites are presented in **Table 6.1**. The augmented position of the biopsy sites has an average visual angle error ranging from 1.18° to 3.86°. This corresponds to 2.99% and 10.09% of the field-of-view.

There are two sources of error in the system that can affect the accuracy of the optical biopsy mapping; 1) the accuracy of the SLAM algorithm; and 2) the accuracy of the biopsy probe estimation. Inaccuracies in the localisation of the laparoscope, with respect to the map, will result in the biopsy site being visualised incorrectly in the image plane. It should be noted that the SLAM algorithm uses the laparoscopic image to perform localisation within an EKF framework, and it finds a solution based on the observed data. By using visually observed data, the localisation is likely to appear visually accurate even if it contains absolute errors in global localisation. The second source of error in the system is due to tool localisation. Accurate 3D estimation of the probe position, from 2D images alone, is difficult due to orientation of the tool relative to the laparoscopic camera. The probe is introduced to the surgical scene parallel to the optical axis of the camera. As a result, the 2D visual appearance of the tool can vary greatly and is affected by perspective visualisation. Small inaccuracies in the estimation of the probe in 2D are magnified when the probe position is estimated by projecting the probe model

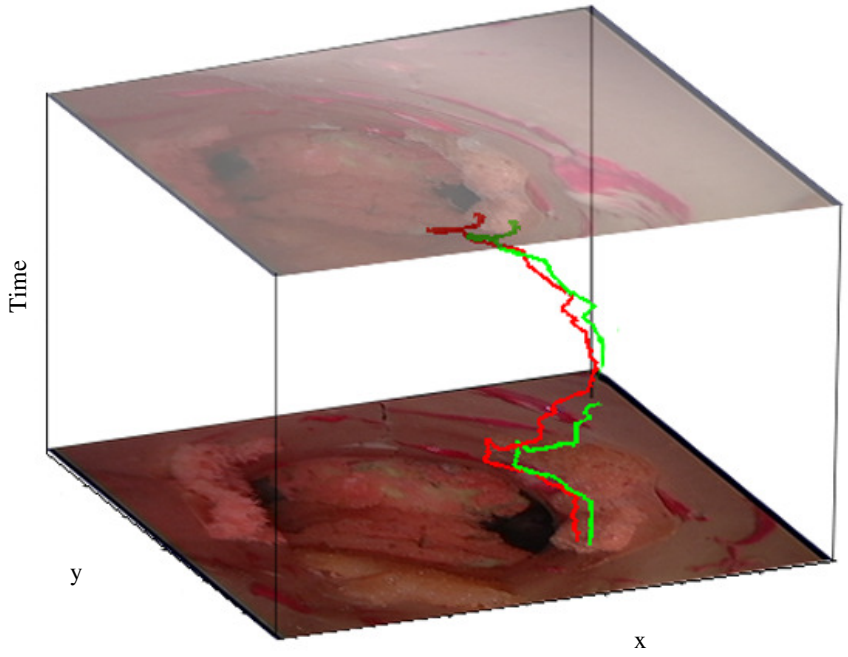
into 3D. Inaccuracies in the 3D location of biopsy sites are disproportionately represented in 2D and can result in a large visual angle error. The proximity of the camera to the biopsy site affects the accuracy of biopsy position estimation where close proximity causes a magnification of the visual angle error.

In order to provide an overall assessment of the performance of the proposed method, the re-projected position of biopsy site three in the image plane is visualised with respect to time, as shown in **Figure 6.5 (a-c)**. In **Figure 6.5 (a)**, the position is plotted in a spatio-temporal visualisation, where the estimated biopsy site is shown in green, and the ground truth is shown in red. **Figure 6.5 (c-d)** directly compares the position of the augmented biopsy site in the X and Y image axes. It is shown that, given an accurate estimation of the position of the biopsy site in 3D, an accurate augmented reality visualisation can be created using the proposed SLAM framework.

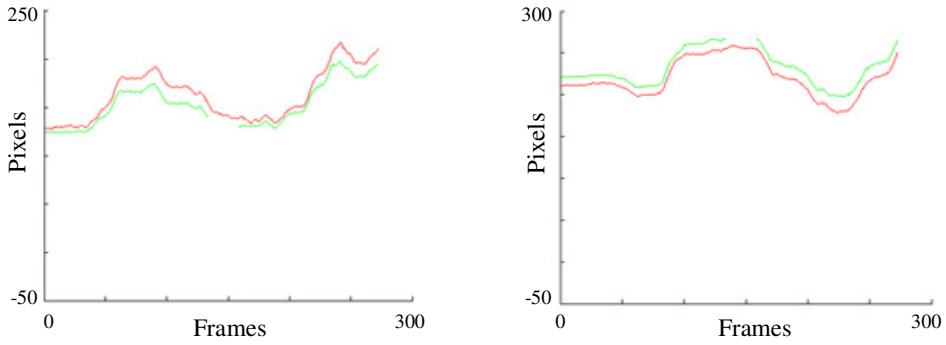
To qualitatively illustrate the benefit of optical biopsy mapping, an augmented reality visualisation is provided in **Figure 6.6**. The figure summarises the laparoscopic sequence showing the areas on the airway where the biopsies were taken. **Figure 6.6 (a-d)** shows the augmented reality visualisation at different stages of the procedure; as undergoing changes in illumination, scale, and view point. This demonstrates the clinical relevance and potential practical value of the proposed method. In **Figure 6.6 (e)**, the entire procedure is represented, including the six biopsies sites in the global map with the associated biopsy images of the sponge cell structures. This demonstrates the feasibility of combining probe tracking and SLAM to co-register multi-modality, intra-operative images for enhanced navigation.

Table 6.1 Average error of biopsy site estimation for the phantom experiment.

Biopsy site number	**Augmented biopsy sites**	
	Visual angle error	**Percent of FOV**
1	2.34°	5.37%
2	3.06°	7.58%
3	2.22°	5.59%
4	1.18°	2.99%
5	2.06°	4.61%
6	3.86°	10.09%



(a)



(b) (c)

Figure 6.5 Analysis of biopsy site number three. **(a)** The ground truth projected position in red, and the estimated position in green for a short section of the procedure. **(b-c)** The ground truth projected position (red) and the SLAM estimated position (green) compared in the X and Y axes of the images plane.

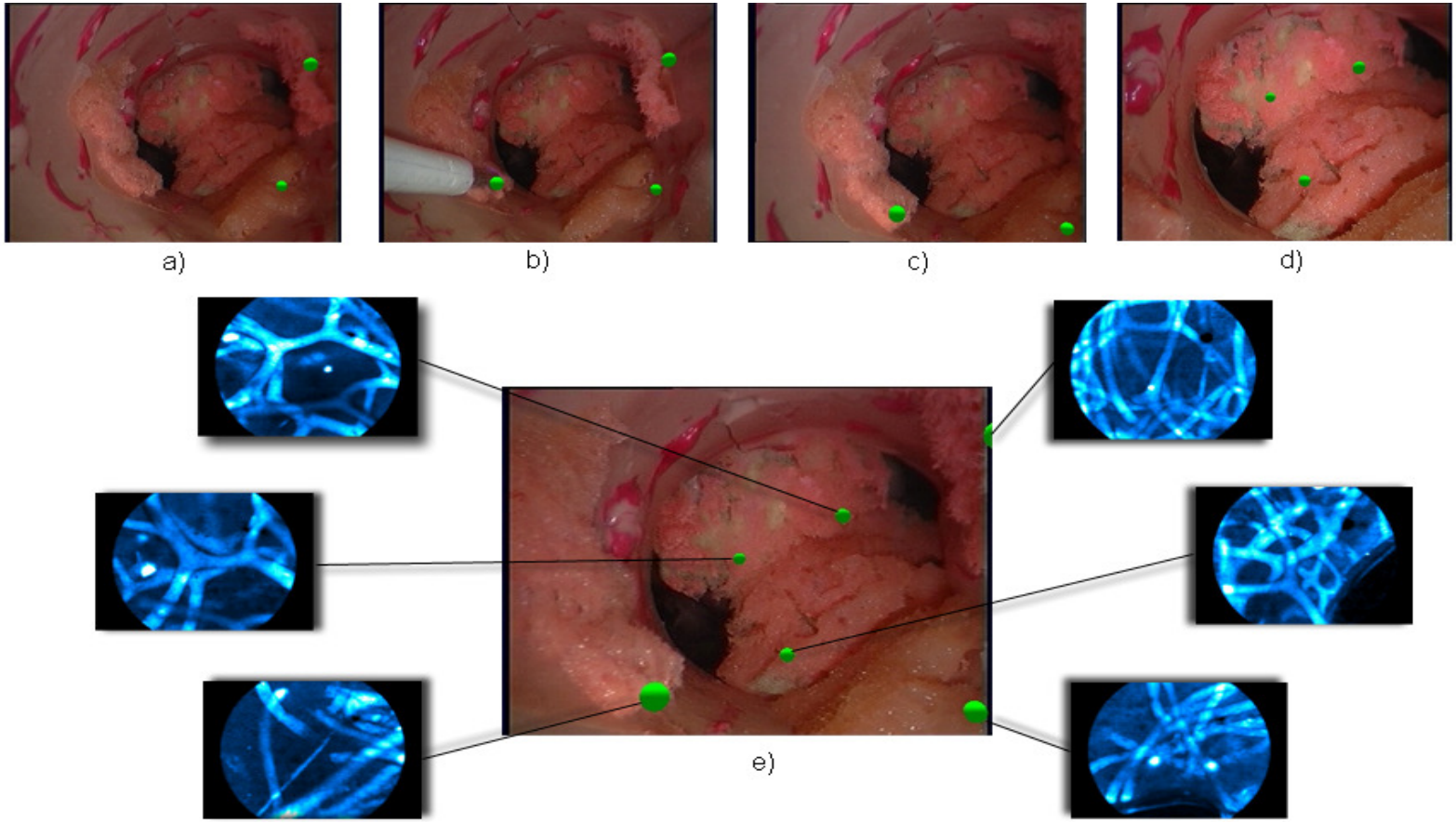


Figure 6.6 (a-d) Position of biopsy sites (green spheres) at different times of the procedure. The spheres are 0.2 cm in diameter and appear in different sizes when they are projected onto the image from different depths; **(e)** shows the six biopsy sites with corresponding micro-confocal fluorescence endoscope images.

6.1.5 Discussion and Conclusions

This section proposed an intra-operative navigation system, which registers two intra-operative imaging modalities onto a common coordinate system. AR is used to visualise the position of optical biopsy sites (acquired using a microconfocal probe) in laparoscopic images. The white light laparoscopic images and microconfocal images cannot be directly aligned using the standard registration techniques. The proposed approach to registration uses the laparoscopic images to track the microconfocal probe, thus enabling the location of biopsy sites to be inferred relative to the camera. Probe tracking is combined with a SLAM framework to enable the biopsy sites to be mapped onto a global coordinate space, which is consistently updated as the camera moves. The system is capable of tracking biopsy sites in 3D and re-projecting them into the camera plane. This allows the retargeting of previously examined tissue regions. Validation of the method was performed on phantom data, which demonstrates the practical use of the proposed SLAM framework for accurate biopsy re-projection. It is proven that the proposed system is capable of operating in a sparse feature environment without prior information regarding tissue geometry.

The system makes a number of assumptions when estimating the location of the biopsy site. The probe tracking approach assumes the instrument can be segmented from the tissue based on colour and background subtraction. The colour of the probe may change when brought into contact with bodily fluids, such as blood, and the approach may be adversely affected by specular highlights. The probe is introduced parallel to the endoscopic imaging device, which leads to large changes in scale and orientation in the image. Small inaccuracies in probe estimation can be magnified, thus creating errors in 3D estimation. A rigid model based approach is used to estimate the 3D position of the tip of the probe. This assumption holds for short sections of the shaft, however, for imaging tissue far away from the endoscope, a more sophisticated, flexible model, or a generalised model, is required.

6.2 Dynamic View Expansion

This section proposes a second application of SLAM for MIS. Navigation during MIS has acknowledged difficulties due to the physical constraints of the endoscope or laparoscope. Off axis visualisation, a loss of direct 3D vision, and a limited field-of-view are all factors contributing to said difficulties. These cause visual-spatial disorientation when exploring complex anatomical structures. The problem is particularly acute in Natural Orifice Transluminal Endoscopic Surgery (NOTES) where a flexible endoscope is used to access the abdominal cavity. One of the main difficulties of navigating during MIS is the limited view of the surgical site provided by the imaging device.

Restricted vision, caused by the above interference, effects the surgeon's awareness of peripheral events and decreases visual-spatial orientation. One solution to this problem is an increase in the camera's field-of-view, specifically a simple fisheye lens. This increases the spatial range projected onto the imaging device but it does not increase the resolution of the imaging device This affects the quality of the image captured. A fisheye lens can also distort the image and change the appearance of *in vivo* structures. Rectilinear lenses can be used to limit the affect of distortion, however, the capabilities of such a hardware solution are limited by the physical confines of the workspace. Replacing existing hardware is not ideal for hospital administrators.

Dynamic view expansion, as proposed in [63], offers a potential solution to expanding the surgeon's field-of-view. This approach is based in image space, does not require additional hardware. The field-of-view of a monocular endoscope is expanded using optical flow. The use of optical flow forces this approach to rely on the brightness constraint, which is not generally held in MIS due to the conjoined light source and endoscope. Large, homogeneous regions introduce additional problems.

The problem of dynamic view expansion can be framed as a temporal registration problem where endoscopic images, captured at different time intervals, are registered to a common coordinate system. Image based approaches, such as mosaicing, can use multiple images to create a single, large image. Difficulties with mosaicing are overcome under constrained conditions where the environment is be assumed to be planar [262].

Mosaicing has been applied a variety of anatomy (a detailed review is provided in **Chapter 2**).

It is important to note that the planar assumption does not hold for general MIS applications. Alternatively, the proposed SLAM framework can be used because it does not require prior knowledge of the environment nor does it make assumptions about scene geometry. SLAM has a distinct theoretical advantage for temporal registration: it is sequential and maintains a long-term estimate of the scene's structure, which is updated at each new video frame.

This work demonstrated the use of SLAM for dynamic view expansion. An overview of the main technical components is shown in **Figure 6.7** and the general concept of the method is illustrated in **Figure 6.8**. The method first generates a sparse probabilistic 3D map or model of the unknown surgical site and estimates the position of the laparoscope relative to the map. This enables the model to be augmented on the current video feed provided by the endoscope. This work addresses the issue of creating visually accurate, textured models via texture selection and blending. Visual inconsistencies created by augmenting the model to the laparoscopic video are also addressed with texture blending. Results are presented for *in-vivo* porcine data in order to demonstrate the potential clinical value.

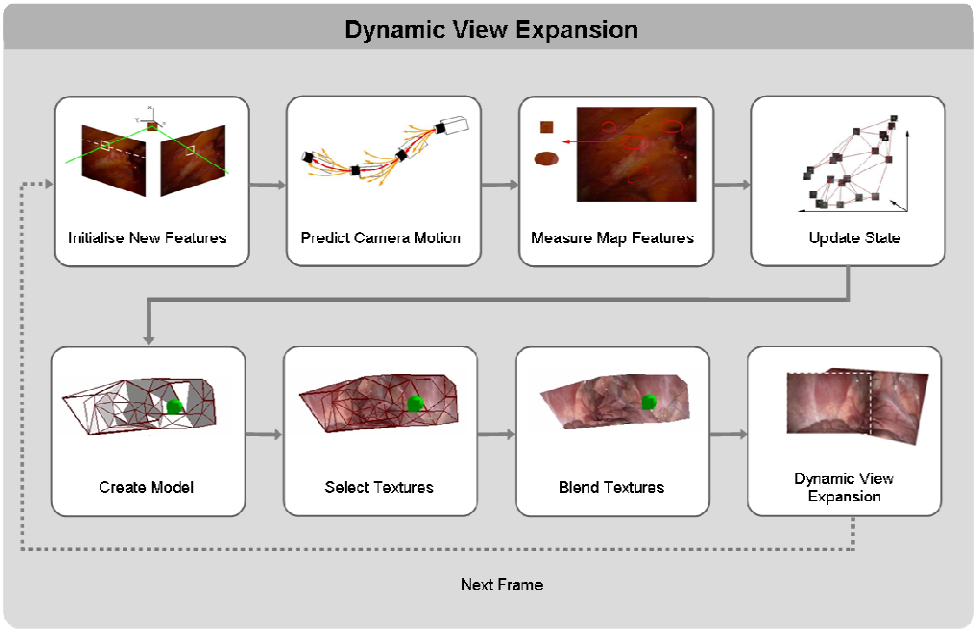


Figure 6.7 A schematic illustration of the Dynamic View Expansion system implementation based on the SLAM framework described in the previous chapter.

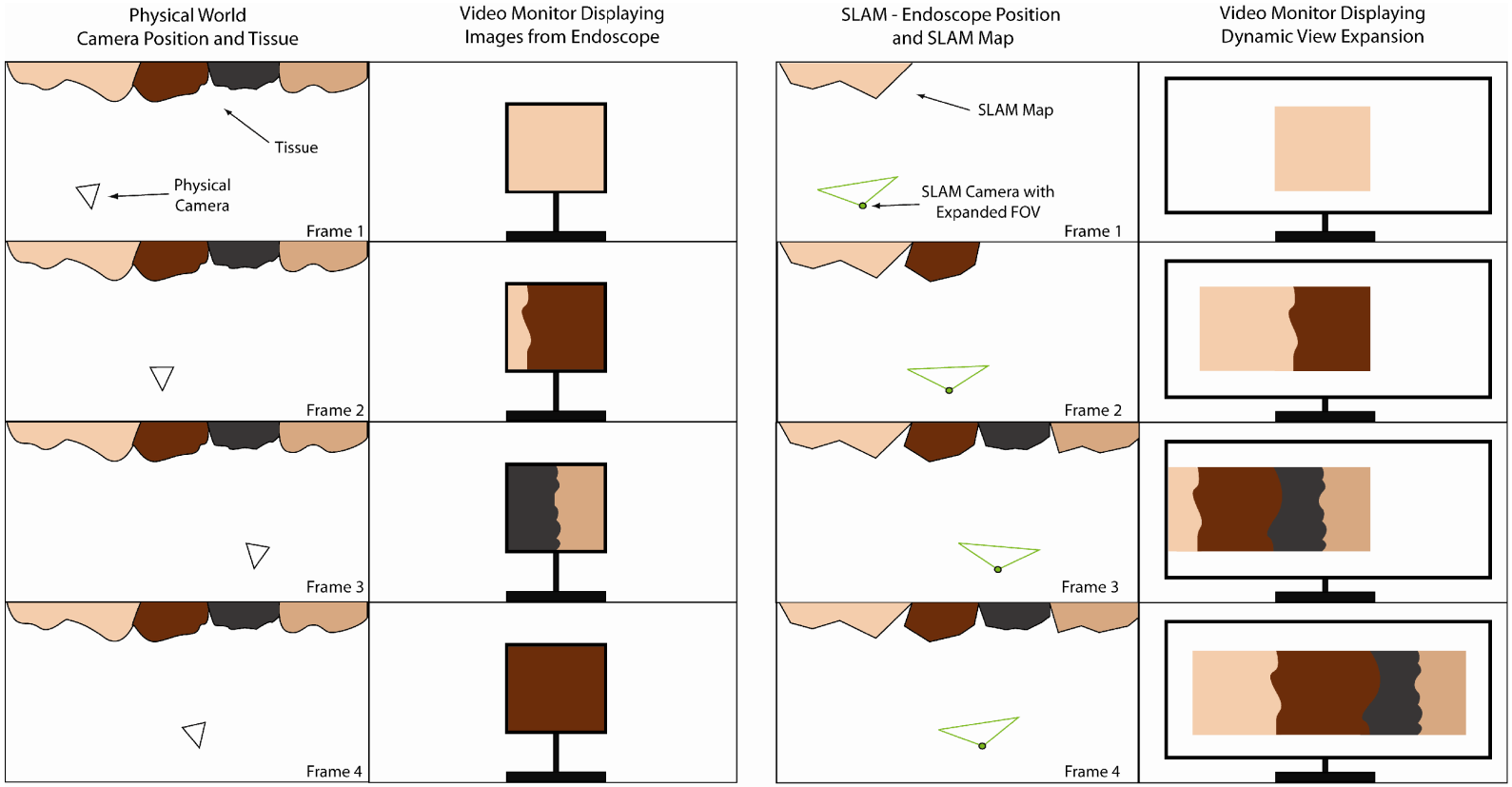


Figure 6.8 Left – The physical world coordinate system showing the position of the camera relative to the tissue as it is navigated by the surgeon and accompanying images from the endoscope. Right – The SLAM coordinate system showing the estimated position of the camera and the incrementally built SLAM map. The camera has an enlarged field-of-view enabling dynamic view expansion shown to the right.

6.2.1 Dynamic View Expansion with SLAM

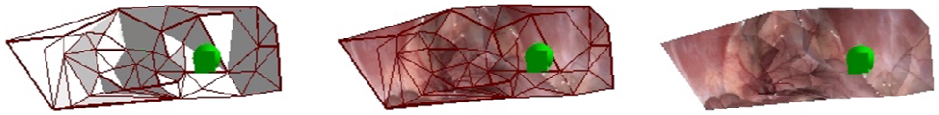
The proposed method expands the effective field-of-view by incrementally creating a 3D model of the tissue as the laparoscope navigates the surgical scene. The approach, illustrated in **Figure 6.8**, is compared to normal visualisation with a laparoscope. When the system is initialised, a 3D textured model of the tissue is created and a global coordinate system defined. The origin of the global coordinate system is the position of the camera at the time of initialisation. The SLAM system proposed in the previous section is used to temporally register images captured by the laparoscope from different locations to the global coordinate system. Temporal image registration with SLAM creates a 3D model of the tissue and estimates the position of the laparoscope in the global coordinate system. This information is used to dynamically expand the camera's field-of-view.

AR is used to visualise the SLAM generated tissue model in the context of the intra-operative images. The registration step is performed by SLAM, and the 6 DOF transformation between the model and the imaging device is the current estimate of the laparoscope pose in the SLAM state vector. This transformation is applied to the model in order to align it with the current laparoscopic field-of-view. The model is then re-projected onto a virtual camera plane with an enlarged field-of-view. The re-projected image of the model is augmented to the current image captured by the laparoscope, thus creating an expanded view. The image captured by the laparoscope is not altered under this scheme while the accuracy of the augmented field-of-view is directly affected by the 3D model's spatial representation of the tissue and texture composition.

6.2.1.1 Tissue Model

The map generated by the SLAM system provides a model of the tissue surface and represents the environment with a sparse set of 3D points. This sparse representation is an approximation of the tissue surface and does not contain sufficient information to perform dynamic view expansion. A solid surface representation of the tissue is generated by meshing the sparse 3D points; a form of interpolation that create an approximated representation of the surface. To perform the meshing Delaunay triangulation [263] is used. This simple approach to meshing creates good surface representation by maximising the minimum angle of all angles in the triangulation. A 2D

method is used that disregards information in the Z axis. This axis corresponds to the initial optical axis of the camera in the global coordinate system. An example of the triangulation is shown in **Figure 6.9 (a)**. The meshed model is constrained to use the features in the SLAM map, which should be distributed across the image to represent as much of the observed environment as possible. Features detected close to the perimeter of the image are given priority when adding new features to the SLAM map.



(a) (b) (c)

Figure 6.9 (a) Delaunay triangulation of the points in a SLAM map with current camera position shown in green. **(b)** Selected textures for each triangle **(c)** the textured 3D tissue model before seam removal.

6.2.1.2 Texture Selection

The visual fidelity of the 3D model is fundamentally important for acceptable aesthetic dynamic view expansion. To this end, the 3D surface model is textured with images taken from the laparoscopic camera in order to recreate a realistic representation of the environment. A single texture is selected for each triangular face on the model. Visual inconsistencies emerge from using multiple images during the model texturing process. Artefacts can be introduced by inconsistent illumination caused by the moving light source, variation in brightness caused by changes in gain, registration errors in the SLAM framework, and interpolation of the surface between sparse points.

A requirement of combined spatial and temporal information makes the formation of these artefacts inevitable. Visually inconsistent texturing and artefacts are reduced by using a small set of images to texture the model. This set is chosen by searching for video frames that can texture the largest number of faces in the model. Areas close to the edge of the image are ignored because the point light source attached to the laparoscope can cause poor visual quality. Image rectification is performed before the textures are applied to the mesh in order to remove possible distortions. This texture selection process effectively reduces visual artefacts in the model appearance, however, the resulting seams are visible where adjacent faces in the model are textured with different images, **Figure 6.9 (c)**. Seams in the model are removed by blending adjacent textures.

6.2.1.3 Seam Removal

The removal of seams in composite images, in this study, is based on Poisson image editing [63, 264]. This approach requires the new image to be mapped onto the existing image by formulating it as a partial differential equation. The border on the new image is constrained to equal the intensities on the existing image by enforcing the Dirichlet boundary conditions [264]. The new texture is added to the existing texture with an overlap of one pixel $\delta\Omega$. A large, sparse positive definite system of linear equations is solved iteratively in the Red, Green and Blue (RGB) channels using a conjugate gradient method with a pre-condition of successive over relaxation

$$\left|N_i\right| f_i^{'} - \sum_{j \in N_i} f_j^{'} = \sum_{j \in \delta\Omega \cap N_i} g_j^{'} + \left|N_i\right| f - \sum_{j \in N_i - \delta\Omega} f_j \qquad \textbf{(6.6)}$$

where N_i is the set of pixels neighbouring pixel i, f_i is the pixel values of the mosaic before updating, $f_i^{'}$ is the unknown pixel values of the updated mosaic, and $g_i^{'}$ are the pixel values of the new texture. The blending is performed in 2D image space, however, the model of the tissue is a 3D structure with arbitrary topology.

Blending is applied to the arbitrary topology of the surface by mapping the textures into a common space. Blending is performed using a pair wise approach whilst considering two textures from adjacent faces. The faces share a common edge in the model space. Different video frames were used to texture each face captured from different camera poses. The resulting faces represented in image space may have varying scale and orientation. The triangular textures in image space must be mapped to a common space of consistent scale and re-orientated before blending can be applied. This is achieved by registering the faces in the image space using the scale and orientation of the shared edge. The difference in scale and orientation of the shared edge in each image defines an affine transformation, which can be used to register the textures. With the application of Poisson blending, the existing image is presumed to be the image that covers the largest number of faces. Before blending is performed, an affine transformation is applied to the new image, thus aligning and scaling it to the same coordinate system as the existing image. This generates a seamless, textured model of the tissue.

6.2.1.4 Augmented Visualisation Seam Removal

The final step in dynamic view expansion is to augment the current 2D laparoscopic image with the back-projected 2D image of the model. This introduces artefacts that result from inconsistencies between the illumination in the current image and the model. The point light source attached to the endoscope causes spatial variation in illumination, which is concentrated in the centre of the image. Unfortunately, illumination is typically poor around the perimeter of the image where the laparoscopic image and back projected model meet. This results in a significant seam. To reduce this effect, the Possion approach described above was adopted. In this approach the brightness of the back-projected image is locally adjusted to match the perimeter of the laparoscopic image.

6.2.2 Experiments and Results

The application of dynamic view expansion is demonstrated for MIS on an *in vivo* porcine experiment. **Figure 6.9** shows an example of the 3D model created by SLAM with the camera position shown in green. In **Figure 6.9 (a)**, the points in the map have been meshed to create a surface. The edges of the triangles are shown in red and the surface in white. **Figure 6.9 (b)** shows the surface mesh and the texture-mapped surface. **Figure 6.9 (a-b)** demonstrates a solid model of the surface that can be approximated using the sparse SLAM map. The use of sparse points creates a coarse representation of the surface. Although the surface model is not an accurate 3D geometric representation of the tissue surface, it will be demonstrated that its visual appearance can be sufficiently improved to enable dynamic view expansion. **Figure 6.9 (c)** shows the texture mapped 3D surface before texture selection. Seams corresponding to the triangles of the mesh are clearly visible on the model. This 3D textured surface forms the basis of dynamic view expansion and augments the current view from the laparoscope using the current estimated position and pose of the camera, indicated in green.

Figure 6.10 (a) shows a textured 3D surface model before texture selection and blending. Seams on the model surface are visible between textured facets from different images. In **Figure 6.10 (b)**, the same surface model is shown after texture selection and blending. In this case, the visual appearance of the seams is greatly reduced, and the coarse surface model is no longer visible, thus making the conception appropriate for visual augmentation. Although illumination variation is visible in the model, the Possion

blending creates a visually smooth transition between the textures. This yields a visually acceptable result.

As a comparison, **Figure 6.10 (c)** shows a textured model augmented onto the current laparoscopic video stream without texture selection and blending. This figure also highlights the seams caused by augmentation between the laparoscopic image and the 3D textured model. In **Figure 6.10 (d)**, the modified surface model is augmented to the laparoscopic image, and the seams between the model and image are removed with Possion blending. The brightness values of the model are locally adjusted to reflect the brightness values of the pixels in the perimeter of the laparoscopic image.

Dynamic view expansion for the sequence is presented in **Figure 6.11**, which demonstrates incremental mapping and the dynamic view, growing over time, as the abdomen is explored. The white box indicates the current laparoscopic image. **Figure 6.11 (e)** indicates where the surgeon returns the laparoscope to its initial position. The entire textured surface model is visualised relative to this point and demonstrates the capability of the system to close navigation loops without incurring error accumulation and drift. **Figure 6.11** further demonstrates the potential clinical value of dynamic view expansion for *in-vivo* abdominal exploration via temporal registration and the enhanced visualisation of intra-operative images.

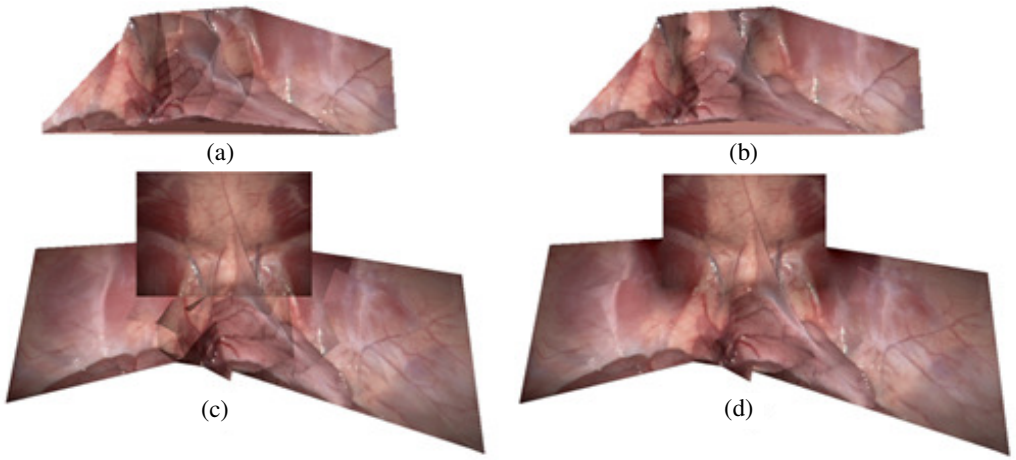


Figure 6.10 A visual comparison of the effect of Poisson texture blending on *in vivo* data. **(a)** Model without blending. **(b)** Model with blending. **(c)** Current view augmented with model without blending. **(d)** Current view augmented with model with blending.

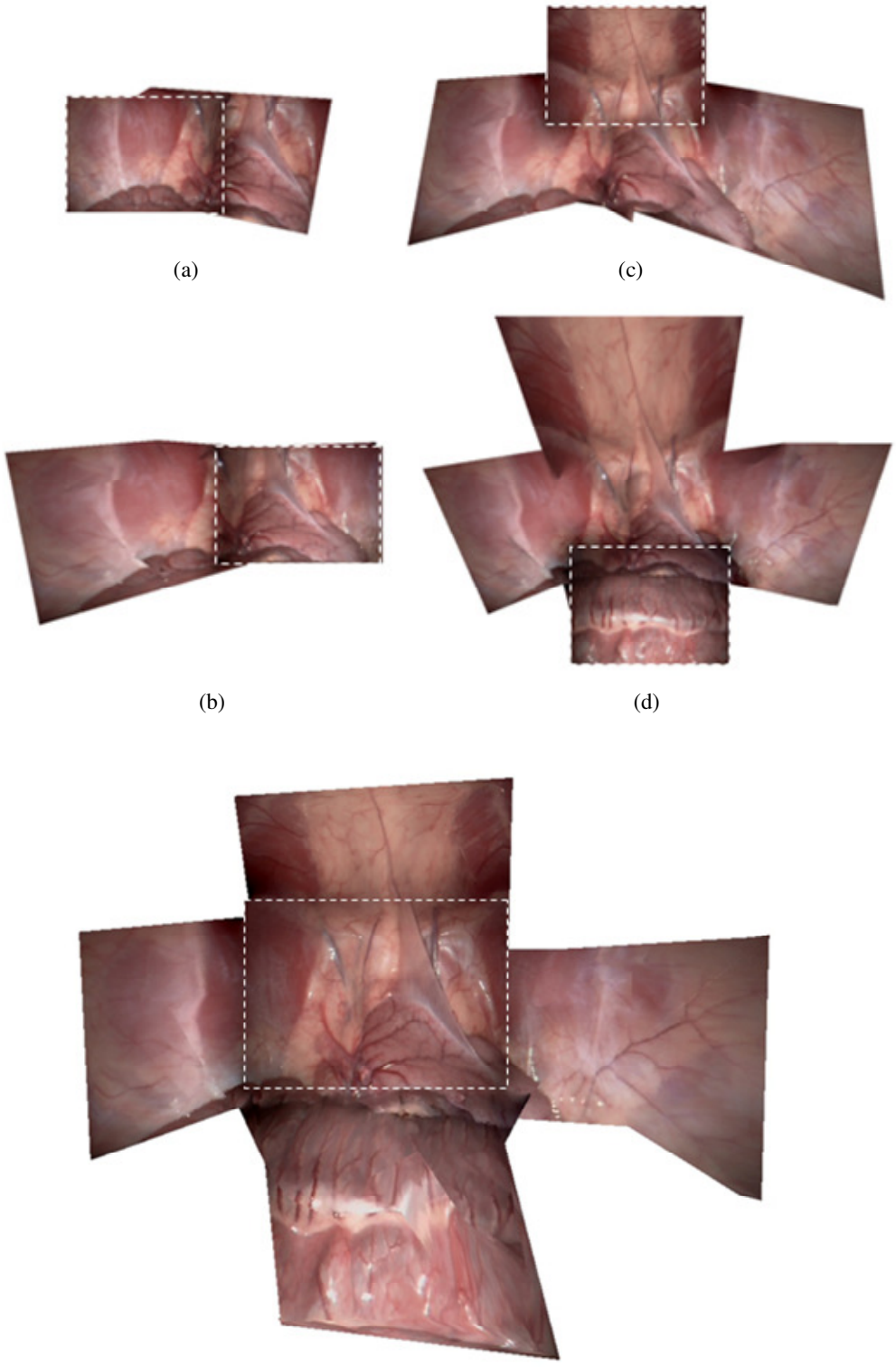


Figure 6.11 Five *in vivo* examples of dynamic view expansion performed during an exploration of the abdomen. The current image from the laparoscope is highlighted with a white, dashed border.

6.2.3 Discussions and Conclusions

This section documents the feasibility of using SLAM for dynamic view expansion. The method demonstrates the capability of temporally registering intra-operative images to a common, global coordinate system for intra-operative AR and enhanced visualisation. It is shown that a visually consistent model of the tissue surface can be generated using texture selection and Possion blending. It is also demonstrated that artefacts caused by augmenting the laparoscopic image with the model may be reduced.

The proposed techniques enable the expansion of the camera's field-of-view by using a model of the tissue. It should be noted that this model is a structural approximation of the tissue surface. Texture selection, blending, and back-projection can introduce artefacts into the visualisation, therefore, dynamic view expansion should not be considered a true representation of the scene and should be used only as a navigational aid to help the surgeon localise the laparoscope and navigate between target anatomies.

The intended use of the system is to provide assistance to navigation and laparoscopic manoeuvres. It is not intended to directly guide tissue-tool interaction. As a result, the system can tolerate higher levels of inaccuracy in the temporal registration of the SLAM system. Absolute global inaccuracies registered to a world coordinate system are tolerable. Unlike image guided surgery, which registers two data-sets to a world coordinate system, dynamic view expansion using SLAM builds the augmented view from a single data-set, which is registered to the camera's initial position.

In summary, two practical scenarios for the use of the proposed SLAM framework have been demonstrated. As previously mentioned, it is important to consider realistic tissue deformation of the surgical site for *in vivo* applications of SLAM. This topic will be addressed in the following chapter.

Chapter 7

Motion Compensated SLAM for Image Guided Surgery

Organs of interest such as the liver and heart undergo constant deformation during Minimally Invasive Surgery (MIS) and the laparoscopic cameras used to observe these deforming organs are rarely static. To cater to both laparoscope and tissue motion, the SLAM framework discussed must explicitly incorporate the deformation. The challenge of recovering the motion of a camera in a non-rigid, dynamic environment is significant. In this case, the image motion, as observed by the camera, contains two coupled components. The first is caused by camera motion, and the second is caused by tissue deformation.

This problem receives increasing interest from the computer vision community. Structure-from-Motion has been extended to non-rigid contexts such as face [167, 168], clothing [169], and heart [136] tracking. This approach is based on the factorisation method and shape basis representation where motion is modelled by rigid rotation, translation components, and non-rigid deformation. This approach requires batch processing and is not conceptually ideal for real-time applications, such as those encountered during MIS.

Simultaneous Localisation And Mapping (SLAM) requires a fixed reference against which error in the map and the pose of the camera is bound. Such a fixed reference is usually a landmark in the static world. Therefore, in dynamic environment the problem is ill-posed [265] as a fixed reference may not be available. SLAM has been adapted to deal with environments involving dynamic motion caused by cars and people. When SLAM is applied to such environments, it exploits the static part of the environment (fixed reference). This simplifies problem, thus leaving the classification of parts of the environment as either static or dynamic. Dynamic motions are generally treated as outliers. Incorporating dynamic motion into the tracking framework [266], however, provides a more accurate representation of the environment and enables more sophisticated interaction with the environment.

This chapter will present a new method for simultaneous estimation of camera motion and dynamic structure. It extends the static SLAM framework to not simply to cope with dynamic motion but to learn a high-level model capable of accommodating organ motion. The learnt motion model is explicitly incorporated into the probabilistic SLAM framework, thus enabling estimation of dynamic tissue motion, including tissue outside the camera's field-of-view. The basic steps of the proposed algorithm are illustrated schematically in **Figure 7.1**. This process is referred to as Motion Compensated SLAM (MC-SLAM). This is the first known work that estimates camera motion and completely dynamic structures online. The proposed method is validated with both synthetic and *ex vivo* data-sets. An *in vivo* application is also demonstrated.

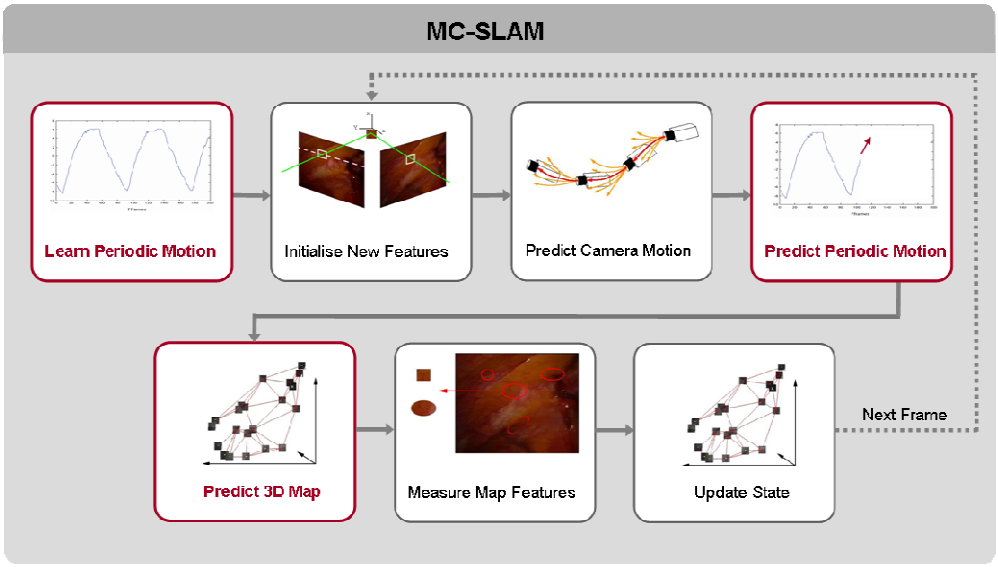


Figure 7.1 Schematic of MC-SLAM system. Additional steps for dealing with dynamic map motion are highlighted in red including; learning the periodic motion model, predicting the motion model and predicting dynamic motion in the map.

7.1 Modelling Dynamic Tissue Motion

As mentioned earlier, dynamic tissue motion is mainly caused by respiration, cardiac motion, or tissue instrument interaction. The respiratory and cardiac cycles are periodic resulting in periodic tissue deformation during MIS. The periodic nature of tissue motion is exploited in this research in order to create a SLAM system capable of working in a totally dynamic environment. This work focuses on hepatic surgery; however, the proposed method can be adapted to any organ or environment where periodic motion can be modelled as shown in **Chapter 4**.

It has been shown in [129] that the motion of the liver is correlated to the periodic motion of the diaphragm and, therefore, to respiration. During MIS, the patient's breathing is often controlled by a ventilator, which regulates the frequency of respiration. The goal is to learn a periodic motion model for respiration and use said model to predict the dynamic motion of the liver. A description of how to estimate and incorporate the motion model into the MC-SLAM framework is provided in the following sections.

7.1.1 Learning the Periodic Motion Model

A calibrated stereo laparoscopic camera with a 5 mm baseline is used to extract the 3D motion of the liver. Features are detected on the liver surface, matched in the stereo images, and tracked temporally. The respiration model is estimated using the stereo laparoscope to measure the 3D motion of points on the liver, see **Figure 7.2**. The 3D positions of points on the liver are estimated by matching regions of interest in the stereo images and performing triangulation. The temporal motion of the 3D points is estimated by tracking the regions over time using the approach outlined in **Chapter 4**.

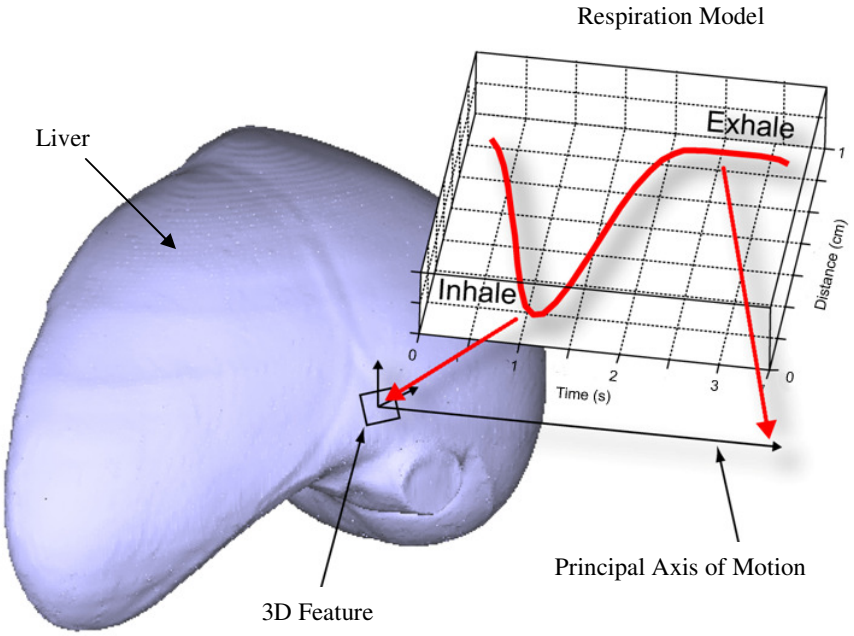


Figure 7.2 Graphical illustration of respiratory modelling from organ motion. This involves: 1) the motion of a region or feature point (of a liver) is tracked with respect to time in 3D, 2) the principal axis of motion (a vector representing the dominant direction of organ motion) is estimated, 3) the periodic motion along this axis is examined, and a respiration model is estimated.

During MIS, the abdomen is insufflated with carbon dioxide, which reduces organ contact and allows the liver to move more freely. The 3D motion of the liver relative to a static camera is shown in **Figure 7.3 (a-c)**. The signal is periodic in all three axes. The data, and thus the modelling of respiration in the coordinate space, depends upon the position of the camera relative to the tissue. The observed data is in 3D space, however, respiration can be modelled as a 1D signal [132].

Previous work has shown that the motion of the liver takes place predominantly along a single axis that corresponds to the superior-inferior direction [132]. Each feature on the liver surface has a principal axis along which the feature will move, see **Figure 7.2**. By examining the motion of the feature along this axis, it is possible to infer a respiration model in 1D. A transformation is required between the 3D coordinate space and the principal axis of motion. This transformation is determined using Principal Component Analysis (PCA) as described in **Chapter 4**. The result of PCA on the data in **Figure 7.3 (a-c)** is shown in **Figure 7.3 (e-f)**. The first component of PCA, indicated in blue, is periodic and corresponds to respiration. The second component contains a small variance caused by hysteresis, and the third component contains negligible variance.

Modelling the motion of organs due to respiration is well considered in medical imaging. A typical respiratory cycle is asymmetrically periodic with a longer dwell time at exhalation [132]. In this case, the following model can be used

$$z(t) = z_0 - b\cos^{2n}(\frac{\pi t}{\tau} - \phi) \qquad \textbf{(7.1)}$$

where z_0 is the position of the liver at the exhale, b is the amplitude, τ is the respiration frequency, ϕ is the phase and n describes the shape or gradient of the model. **Equation (7.1)** is used to model the data in the first component of PCA (shown in **Figure 7.3 (d)**). The parameters of **Equation (7.1)** are estimated using Levenberg-Marquardt. This minimisation algorithm poses the problem as a least squares curve fitting.

It is possible to estimate the respiration cycle using a feature at any position of the liver surface, assuming it can be tracked throughout the cycle. The transformation from the XYZ coordinate system to the respiration coordinate system is unique to each feature. This means features on the surface of the liver can move in independent directions while sharing the same respiration model.

It is shown that the 1D respiratory cycle can be estimated from the observed 3D data. It is possible to estimate the dynamic motion within the environment using a model of respiration. The dynamic motion of the point on the tissue is estimated by multiplying the current point in the respiratory cycle $z(t)$ and the inverse PCA transformation

matrix. The PCA transformation matrix determines a vector, or principal axis of motion, in 3D along which the point on the surface of the tissue moves.

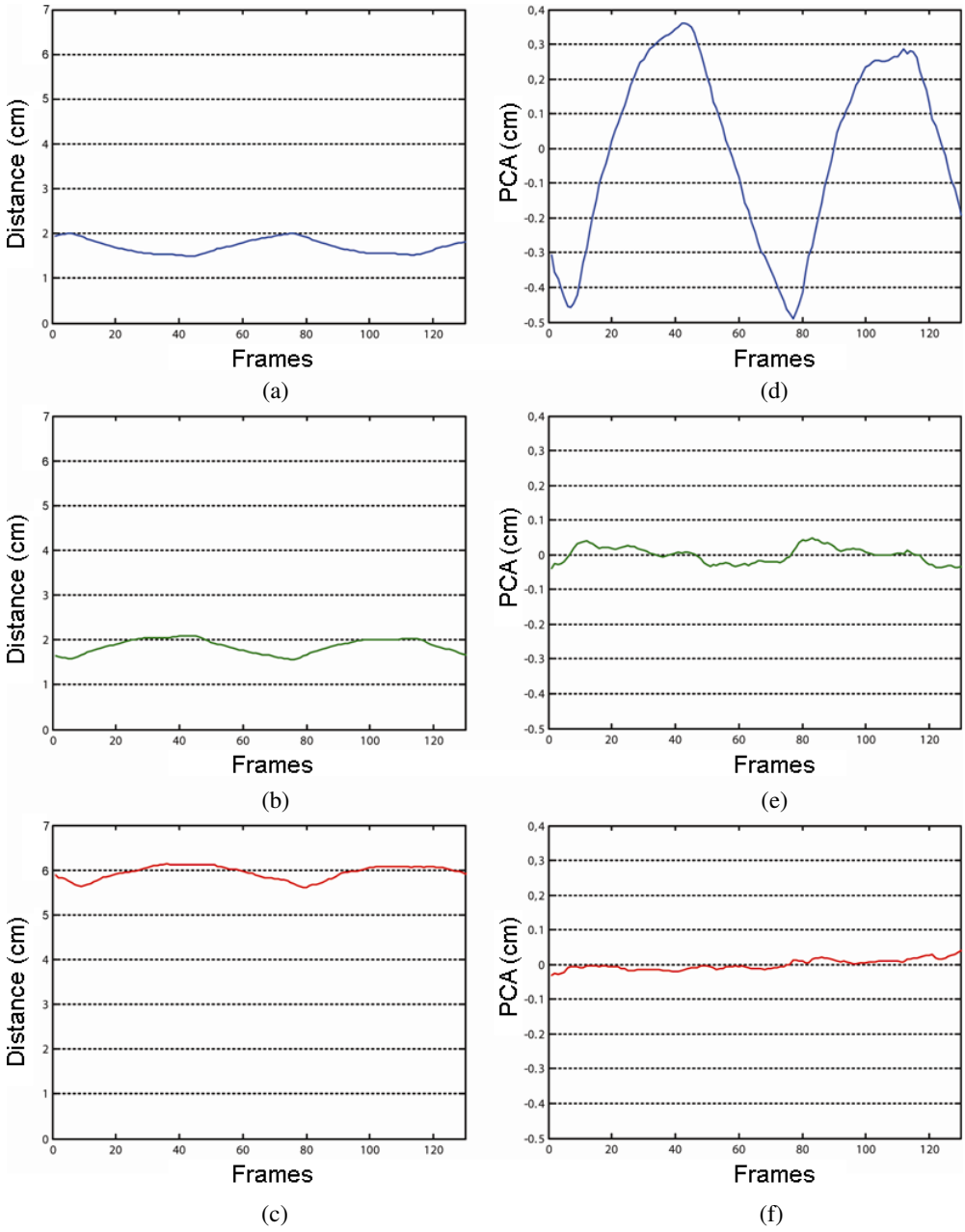


Figure 7.3 **(a)** The X, **(b)** Y, and **(c)** Z coordinates of a tracked feature on the surface of an *in vivo* liver, **(d)** the first, **(e)** second, and **(f)** third components from PCA.

7.1.2 MC-SLAM Formulation

The core of the proposed framework is an extended SLAM framework. It is generally assumed that the map is static and constant over time in SLAM. In MC-SLAM, a periodic motion model is introduced to compensate for the dynamic motion in the map, thus enabling the accurate localisation of the camera. This introduces three additional steps into SLAM, as shown in **Figure 7.1**. The first step requires an initial estimate of the periodic respiration model, using the method described in the previous section, to be learnt. The second and third steps are the prediction of the respiration motion model and the dynamic motion in the map. A new state vector, prediction model, and measurement model are introduced in conjunction with these steps.

7.1.2.1 Probabilistic Framework

MC-SLAM is implemented using an Extended Kalman Filter (EKF). The state vector $\hat{\mathbf{x}}$ is composed of three elements representing the camera, the periodic respiration model, and the map. $\mathbf{P}$ is the square covariance matrix.

$$\hat{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}}_v \\ \hat{\mathbf{m}} \\ \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \end{pmatrix} \quad \mathbf{P} = \begin{pmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xm} & \mathbf{P}_{xy_1} & \mathbf{P}_{xy_2} & \cdots \\ \mathbf{P}_{mx} & \mathbf{P}_{mm} & \mathbf{P}_{my_1} & \mathbf{P}_{my_2} & \cdots \\ \mathbf{P}_{y_1x} & \mathbf{P}_{my_1} & \mathbf{P}_{y_1y_1} & \mathbf{P}_{y_1y_2} & \cdots \\ \mathbf{P}_{y_2x} & \mathbf{P}_{my_2} & \mathbf{P}_{y_2y_1} & \mathbf{P}_{y_2y_2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{7.2}$$

The camera's state vector $\hat{\mathbf{x}}_v$ contains the position $\mathbf{r}^W$, orientation $\mathbf{q}^{WR}$, translational velocity v^W, and angular velocity w^R of the camera. The periodic respiration model $\hat{\mathbf{m}}$ is represented in the state by the parameters derived from **Equation (7.1)** such that

$$z(t) = z_0 - b\cos^{2n}(\alpha) \tag{7.3}$$

where $\alpha = \pi t / \tau$, t is the time step, z_0 is the exhale position of the liver, b is the amplitude, τ is the frequency, and $n = 3$, in accordance with [132]. The phase ϕ is ignored since the system is initialised at $\phi = 0$.

$$\mathbf{m} = \begin{pmatrix} \alpha \\ \tau \\ \mathbf{b} \\ \mathbf{z_0} \end{pmatrix} \quad \hat{\mathbf{y}}_i = \begin{pmatrix} \overline{\mathbf{y}} \\ \mathbf{eig} \end{pmatrix} \tag{7.4}$$

The features in the map $(\hat{\mathbf{y}}_1 \cdots \hat{\mathbf{y}}_i)$ are represented in the state by two components, each of three elements are taken from the PCA transformation. In **Equation (7.4)**, $\overline{\mathbf{y}} = (X, Y, Z)$ represents the mean position of the feature in 3D space during a respiration cycle, and $\mathbf{eig} = (\mathbf{eig_x}, \mathbf{eig_y}, \mathbf{eig_z})$ are the eigenvectors which describe the transformation from 3D space to the periodic respiration model or the first component of PCA. These eigenvectors can be used to define a vector along which the feature will move in 3D space. The feature's position along the vector is determined by the phase of the respiration cycle.

The system is initialised once the periodic respiration model has been learnt. During this learning phase, it is assumed that the camera is static or an accurate positional estimate is available. A full cycle is detected by normalising the data in the principal component, smoothing using a moving average and detecting points where the signal changes from positive to negative or negative to positive. This learning phase can be reduced by combining the frequency and phase data from the ventilator with the 3D amplitude and exhale position acquired from the image data.

7.1.2.2 State Prediction Model

The camera motion is predicted using a standard *constant velocity, constant angular velocity* model. The state prediction model requires the additional step of predicting the periodic respiration and, subsequently, the motion in the map. The prediction model $\mathbf{f}_v^m$ and process noise covariance $\mathbf{Q}_v^m$ for the periodic respiration $\hat{\mathbf{m}}$ are defined as

$$\mathbf{f}_v^m = \begin{bmatrix} 1 & t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q}_v^m = \begin{bmatrix} \frac{\Phi_\tau t^3}{3} & \frac{\Phi_\tau t^2}{2} & 0 & 0 \\ \frac{\Phi_\tau t^2}{2} & \Phi_\tau t & 0 & 0 \\ 0 & 0 & \Phi_b t & 0 \\ 0 & 0 & 0 & \Phi_{z_0} t \end{bmatrix} \tag{7.5}$$

where Φ_τ is the noise in the frequency, Φ_b is the noise in the amplitude, and Φ_{z_0} is the noise in the exhale position.

7.1.2.3 Measurement Model

The measurement model is used to transform the state space into the measurement space. In MC-SLAM the features in the map are measured relative to the predicted position of the camera. As previously described, a feature's position in the camera coordinate system is computed using the predicted camera's position $\mathbf{r}^W$, rotation $\mathbf{R}^{RW}$, and the position of the feature in 3D $\mathbf{y}_i^W$, such that $\mathbf{y}_i^R = \mathbf{R}^{RW}(\mathbf{y}_i^W - \mathbf{r}^W)$. The predicted position of the feature in the world coordinate system is calculated using the predicted point in the respiration cycle and $\hat{\mathbf{y}}_i$, such that $\mathbf{y}_i^W = \mathbf{eig}(z_0 - b\cos^{2n}(\alpha)) + \overline{\mathbf{y}}$. The measurement model is therefore

$$\mathbf{h}_i^R = \mathbf{R}^{RW}(\mathbf{eig}(z_0 - b\cos^{2n}(\alpha)) + \overline{\mathbf{y}} - \mathbf{r}^W) \tag{7.6}$$

and the partial derivatives of the measurement model for the periodic respiration model $\hat{\mathbf{m}}$ are

$$\frac{d\hat{\mathbf{m}}}{d\alpha} = \mathbf{R}^{RW}\mathbf{eig}(nb\sin(\alpha)\cos(\alpha)^{n-1}) \tag{7.7}$$

$$\frac{d\hat{\mathbf{m}}}{d\tau} = 0 \tag{7.8}$$

$$\frac{d\hat{\mathbf{m}}}{db} = -\mathbf{R}^{RW}\mathbf{eig}(\cos(\alpha)^n) \tag{7.9}$$

$$\frac{d\hat{\mathbf{m}}}{dz_0} = \mathbf{R}^{RW}\mathbf{eig} \tag{7.10}$$

The features are projected into the image plane and matched using active search.

7.1.2.4 Feature Initialisation

Once the system is initialised, new features can easily be added to the state when the camera is moving. To obtain accurate estimates of $\bar{\mathbf{y}}$ and $\mathbf{eig}$, features are added after observing their motion in the 3D world for one respiratory cycle. The motion of a partially initialised feature is determined by tracking the feature relative to the camera position. The feature position is transformed into the world coordinate system using the camera's position in the state vector.

7.2 Experiments and Results

In order to validate the theoretical and practical value of MC-SLAM, the method is quantitatively evaluated on simulated and *ex vivo* data with induced deformation. To demonstrate its practical application, *in vivo* data from a porcine experiment, is used.

7.2.1 Simulated Experiments

For detailed quantitative validation with ground truth, a simulated data-set was created with known camera motion, tissue location, and respiration cycle. An image of the liver and gall bladder was textured onto a curved 3D surface. Periodic motion was applied to the surface using **Equation (7.1)** and the parameters in **Table 7.1**. A virtual stereo camera is navigated through the simulated environment along the X, Y, and Z axes, thus capturing images of the dynamic surface motion. Gaussian noise is added to the images.

Validation of the dynamic map's accuracy and the methodology employed to recover the respiration model from laparoscopic data are provided in **Figure 7.4 (a)** where the model for respiration is shown in green, and the ground truth is shown in red. The graphs are similar, and the extracted parameters for respiration are compared to the ground truth in **Table 7.1**. These parameters provide the initial estimate of the model and are incorporated into the EKF framework.

The recovered position of the camera is quantitatively evaluated in **Figure 7.4 (b-g)**. **Figure 7.4 (b-d)** shows the motion of the camera accurately recovered using MC-SLAM

(green) in all axes of motion as well as a comparison to the ground truth (red). The mean position error is 0.25 cm with a standard deviation of 0.19 cm. The error in the system is attributed to sharp changes in acceleration when the camera begins and ceases to move and is similar to the results in Chapter 5. These changes are not well modelled by the *constant velocity, constant angular velocity motion* model. The recovered motion of the camera using static SLAM is shown in **Figure 7.4 (e-g)** where the ground truth is shown in red and the static SLAM position estimate is shown in blue. It is clear, as demonstrated in **Figure 7.4 (e-g)**, there is a periodic error in the position estimate when static SLAM is used. This is further confirmed by the mean position error, which is 1.31 cm with a standard deviation of 0.6 cm. The motion in the simulated map follows the Z-axis resulting in the error in camera position shown in **Figure 7.4 (g)**. The roll, pitch, yaw, and rotations are accurately recovered by both systems.

Figure 7.5 illustrates the results of MC-SLAM on simulated data when compared to static SLAM. The figure shows the MC-SLAM and SLAM coordinate system from an aerial perspective looking down the Y axis. The camera is navigated along the X axis, and both camera and map motion may be seen in the figure. **Figure 7.5 (a-b)** shows the MC-SLAM coordinate system at frame zero and 500, respectively. **Figure 7.5 (a)** shows the position of features in the map at full inhale, and **Figure 7.5 (b)** shows the position of features in the map at full exhale. The map evidently contains dynamic motion. **Figure 7.5 (a)** and **Figure 7.5 (b)** show both camera motion and the estimation of the dynamic feature positions, including when these features are beyond the field-of-view. Feature positions estimated outside the field-of-view are delineated using a yellow surrounding, and features measured by MC-SLAM are shown using a red surrounding. The motion of the camera is recovered when tissue and camera motion are observed together.

Figure 7.5 (c-d) illustrates the above situation for static SLAM. The map remains static throughout the sequence, and there is a periodic error in the Z axis. This error corresponds to the motion in the map along the Z axis. The MC-SLAM map is denser than the static SLAM map. In MC-SLAM, a feature is only added to the map once it has been observed for one respiration cycle. As a result, some features near the perimeter of the image cannot be tracked continuously for one respiration cycle, thus only allowing features in the centre of the image to be added to the map. Static SLAM, notably, does not have this requirement and adds features that move outside the current field-of-view as a result of motion in the map.

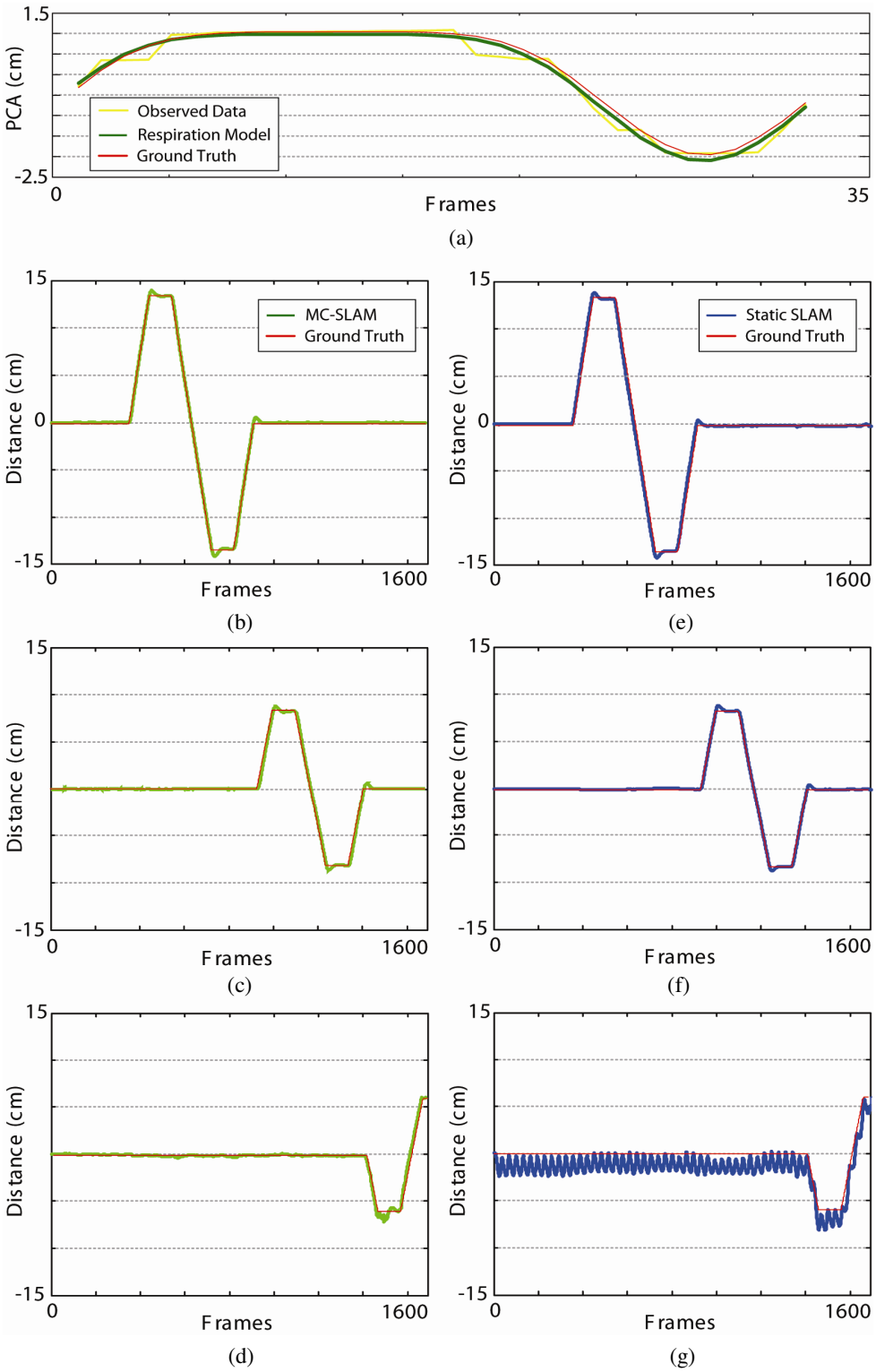


Figure 7.4 Simulated data. **(a)** Respiration model; observed data, respiration model, and ground truth. **(b-d)** Laparoscopic position for MC-SLAM (green) and ground truth (red). **(e-g)** Laparoscopic position static SLAM (blue) and ground truth (red).

Table 7.1 Periodic respiration model parameters for simulated data.

	τ **(Frames)**	b **(cm)**	z **(cm)**
Simulation Estimated	32.38	3.09	0. 95
Simulation Ground Truth	31.83	3	1

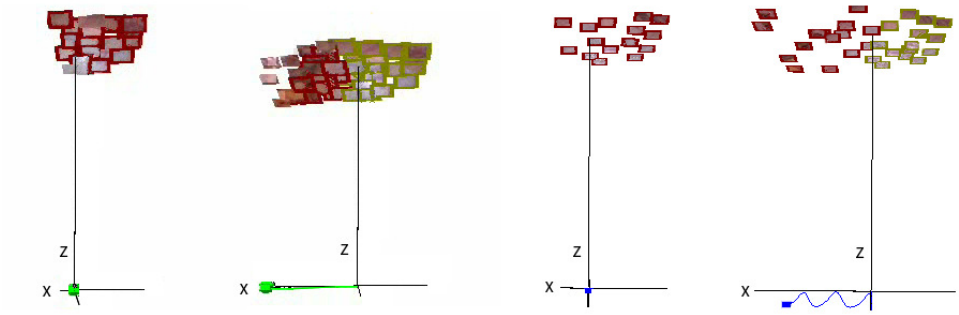


Figure 7.5 Simulated data for MC-SLAM evaluation at **(a)** frame zero and **(b)** frame 500 illustrating the dynamic map and motion compensated camera estimation (green). Static SLAM at **(c)** frame zero and **(d)** frame 500 illustrating the static map and erroneous camera estimation (blue). Tracked features are shown using a red boarder and estimated feature positions with a yellow border.

Figure 7.6 illustrates the results of MC-SLAM and static SLAM for a simulated sequence of over 1600 frames. The top row of each figure shows the simulated laparoscopic images: the position of tracked features is indicated using a black square. An ellipse demarcates the uncertainty in the estimated position. The ellipse is coloured red if the feature is successfully tracked, blue if no match can be found, and yellow if no attempt is made to match the feature. Attempts to match features are only made if the feature is visible in the left and right stereo images and not near the image perimeter. Only a selected, pre-defined number of features are matched at each frame to improve computational efficiency. In the MC-SLAM coordinate system, the estimated camera position is depicted using a green cube: a green line illustrates its trajectory. In the static SLAM coordinate system, the estimated camera position is shown with a blue cube: a blue line illustrates its trajectory. In both systems, the ground truth position is depicted using a red cube: a red line illustrates its trajectory. **Figure 7.6 (a-f)** shows the results for static SLAM. The blue line showing the static SLAM camera trajectory oscillates with a periodic error and is clearly different from the red line representing the ground truth. In **Figure 7.6 (g-l)**, the green line representing the MC-SLAM camera motion is almost indistinguishable from the ground truth.

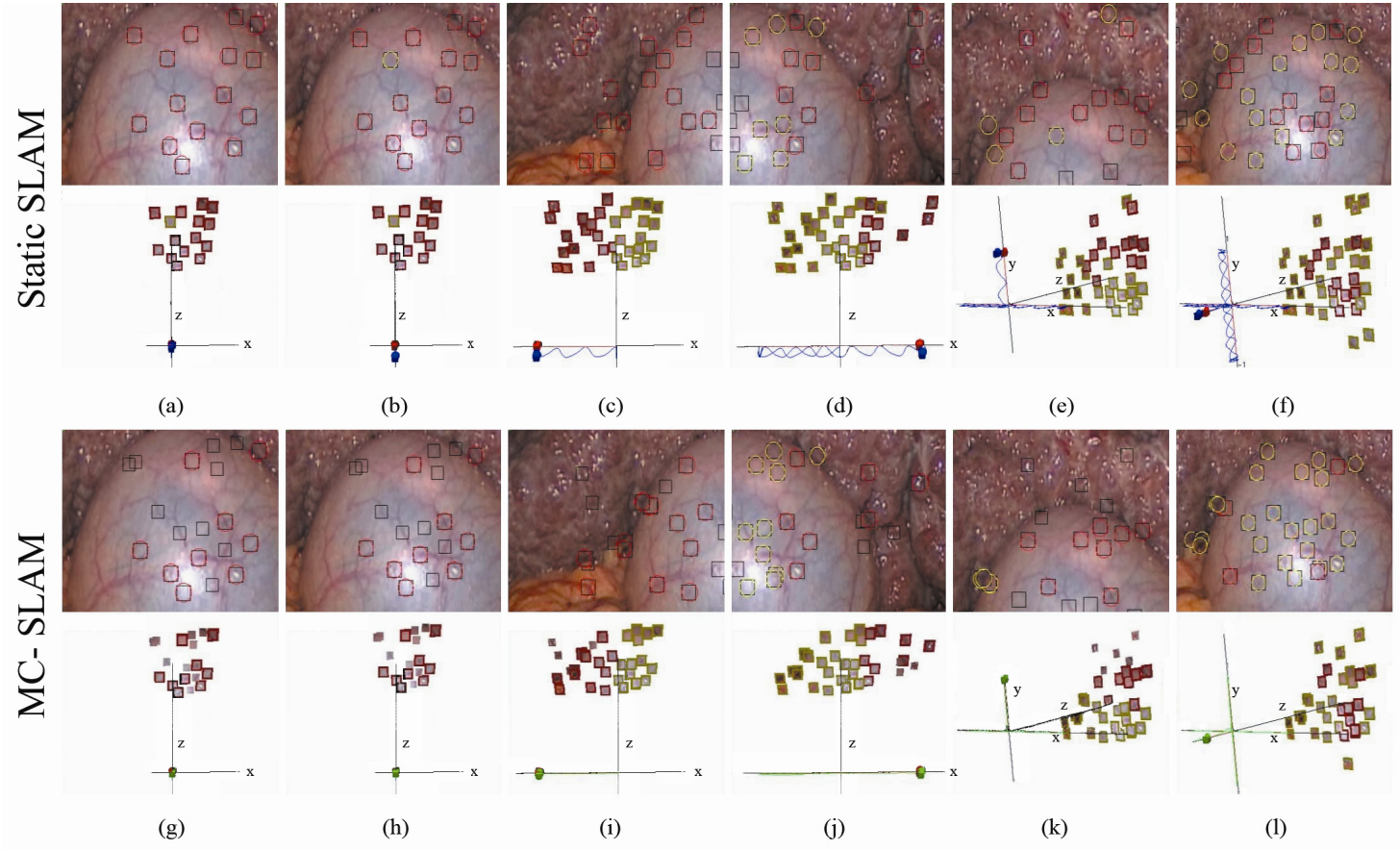


Figure 7.6 Simulated data showing the laparoscopic image (with tracked features) and the SLAM coordinate system (with map features and laparoscope position). **(a-f)** Static SLAM system with camera position shown in blue and ground truth shown in red. (g-l) MC-SLAM system with camera position shown in green and ground truth shown in red.

Table 7.2 Periodic respiration model parameters for *ex vivo* data.

	τ **(Frames)**	b **(cm)**	z **(cm)**
Ex Vivo **Estimated**	52.47	0.85	0.33
Ex Vivo **Ground Truth**	52	0.9	0.3

7.2.2 *Ex Vivo* Experiments

To validate the MC-SLAM algorithm on real tissue, an *ex vivo* experiment was set-up with induced tissue motion. The ground truth position of the stereo laparoscopic camera was obtained using the approach outlined in **Chapter 5**. An *ex vivo* porcine liver sample was used. In order to induce realistic motion, as encountered during MIS, the sample was placed on a sliding tray. A custom made device, shown in **Figure 7.7**, was attached to the tray to induce periodic motion that simulates respiration. The device consisted of a stepper motor and a cam. The cam was designed to obtain a profile, which models **Equation (7.1)** defining the parameters z, n, and b. The stepper motor was controlled by a computer and determines the τ parameter. The parameters used in the experiment are detailed in **Table 7.2**. This experimental set-up provides ground truth data for the camera's position and the dynamic motion of the environment.

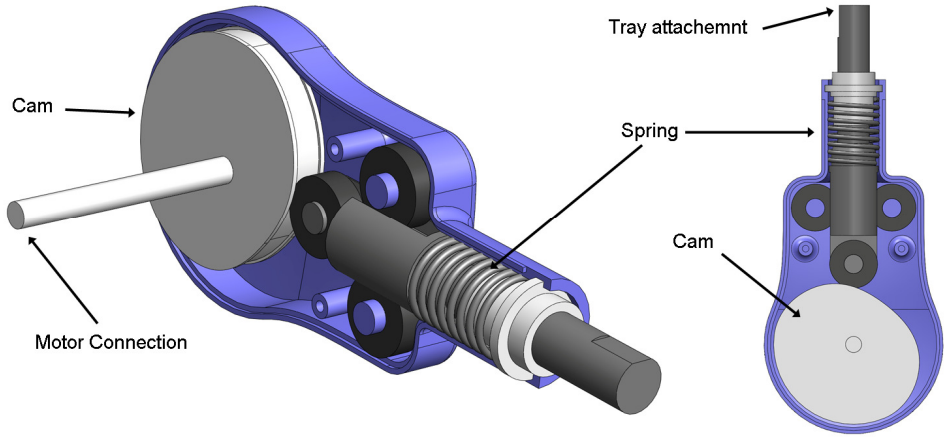


Figure 7.7 Custom made mechanical device used to replicate periodic respiration during *ex vivo* experiments. The motion is controlled by a motor, which is connected to the cam. The profile of the cam is designed to create an asymmetric motion by pushing the shaft away from the centre of the cam. The spring holds the shaft in place and maintains contact with the cam. A tray is attached to the end of the shaft upon which the tissue is fixed.

Quantitative evaluation of respiration modelling is provided in **Figure 7.8 (a)** and in **Table 7.2**. These results demonstrate how the principal axis of motion resulting from respiration can be recovered when the motion pursues an arbitrary direction. The recovered parameters in **Table 7.2**. are used to initialise the respiration model EKF in the SLAM framework. These parameters also indicate the accuracy of the dynamic motion in the map.

Quantitative evaluation of the MC-SLAM algorithm is provided in **Figure 7.8**. The estimated camera position using MC-SLAM (green) is compared to ground truth (red) in **Figure 7.8 (b-d)**. **Figure 7.8 (e-g)** demonstrates the results for static SLAM applied to the same sequence. The recovered motion of the camera using MC-SLAM closely follows the ground truth. Static SLAM periodically oscillates away from the ground truth. The mean error for MC-SLAM and static SLAM is 0.11 cm with a standard deviation of 0.07 cm and 0.56 cm with a standard deviation of 0.25 cm respectively. Rotations are accurately recovered by both systems. The MC-SLAM system does exhibit some drift after initialisation. This is shown in **Figure 7.8 (d)** by the motion along the Z-axis.

It was observed that static SLAM is more susceptible to data association errors in dynamic environments. This is due to the position of features that cannot be accurately predicted in the image space resulting in active search failure. This is particularly problematic on visually repetitive data, such as the liver surface, where regions appear similar. A data association error occurs in static SLAM between frames 800 and 1000. This is demonstrated in **Figure 7.8 (e-g)** by the change in the osculation of the error.

The performance of MC-SLAM is demonstrated on *ex vivo* data in **Figure 7.9**. **Figure 7.9 (a-e)** shows the intra-operative laparoscopic image with features tracked by the SLAM system. **Figure 7.9 (f-j)** demonstrates the MC-SLAM coordinate system where the estimated camera position is shown as a green cube for MC-SLAM, as a red cube for ground truth, and a blue cube for static SLAM - with corresponding trajectories shown as lines. **Figure 7.9 (k-o)** illustrates the intra-operative laparoscopic images augmented with a virtual tumour, which is manually and rigidly registered to the MC-SLAM map. The tumour is visualised using Augmented Reality (AR). AR is implemented with Inverse Realism [43] to improve depth perception of the virtual object.

The laparoscopic image shown in **Figure 7.9 (a-b)**, and associated visualisations, are captured from a static laparoscope and illustrate the position of the tumour at full exhale and full inhale. It is clear in the MC-SLAM coordinate space in **Figure 7.9 (f-g)** that the camera is estimated as static, however, the static SLAM system contains camera motion. The dynamic nature of the map and the AR is illustrated in **Figure 7.9 (k-l)** where **Figure 7.9 (k)** shows the position of the tumour at full exhale, and **Figure 7.9 (l)** shows its position at full inhale from a static camera.

The remaining laparoscopic video sequence contains tissue motion and camera motion. In the MC-SLAM coordinate system, the camera motion results in new features added to the map, see **Figure 7.9 (h-j)**, and demonstrates incremental mapping and the ability to add new features on the fly. In **Figure 7.9 (o)**, the laparoscope navigates away from the tumour, and its position is visualised outside the current field-of-view using Dynamic View Expansion, as described in **Chapter 6**. This illustrates the capability of the MC-SLAM to predict the dynamic 3D position of tissue, including when it is not directly measured with inter-operative imaging.

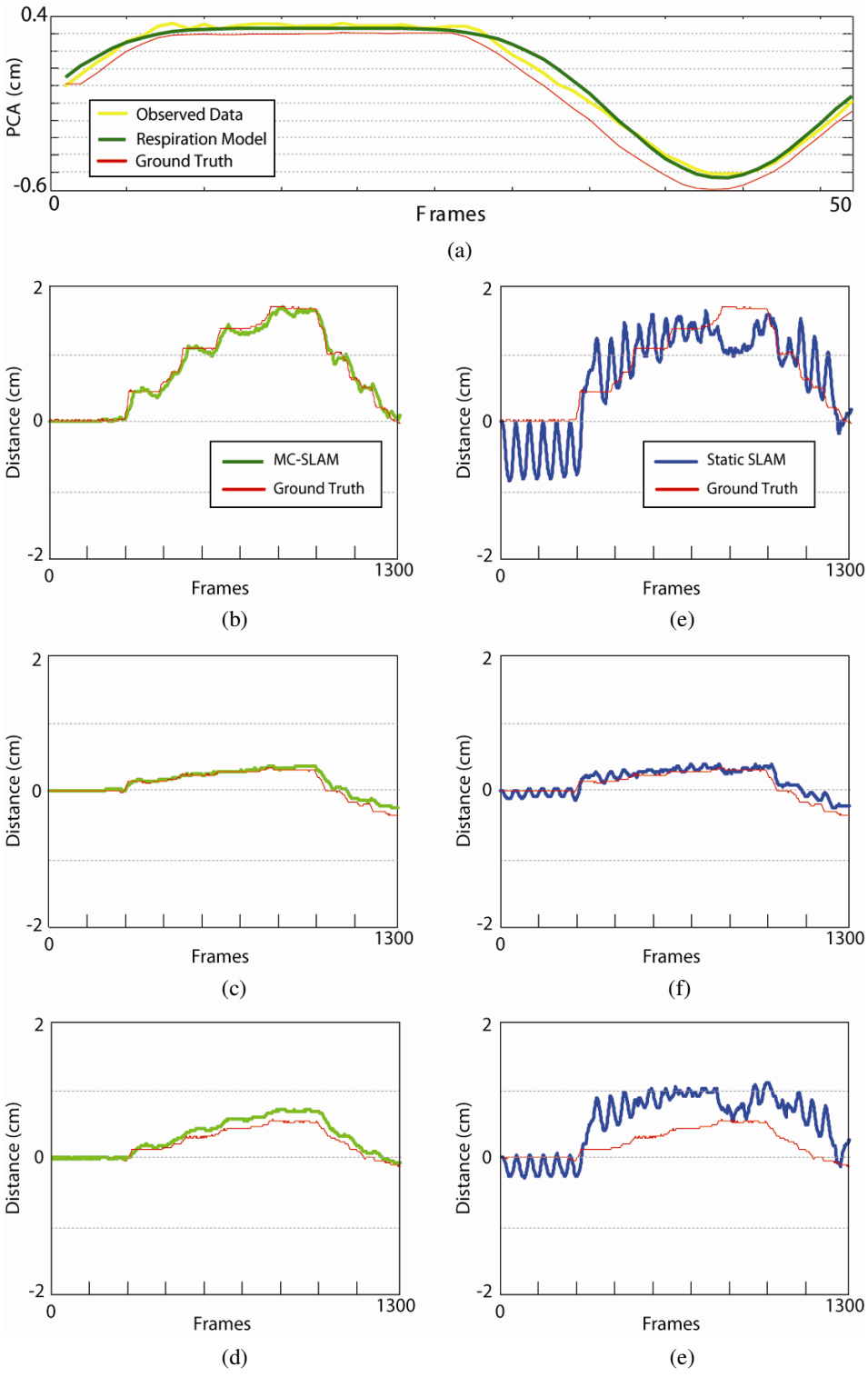


Figure 7.8 *Ex vivo* data. **(a)** Respiration data showing the observed data, respiration model and ground truth. **(b-d)** Laparoscopic position for MC-SLAM (green) and ground truth (red). **(e-g)** Laparoscopic position static SLAM (blue) and ground truth (red).

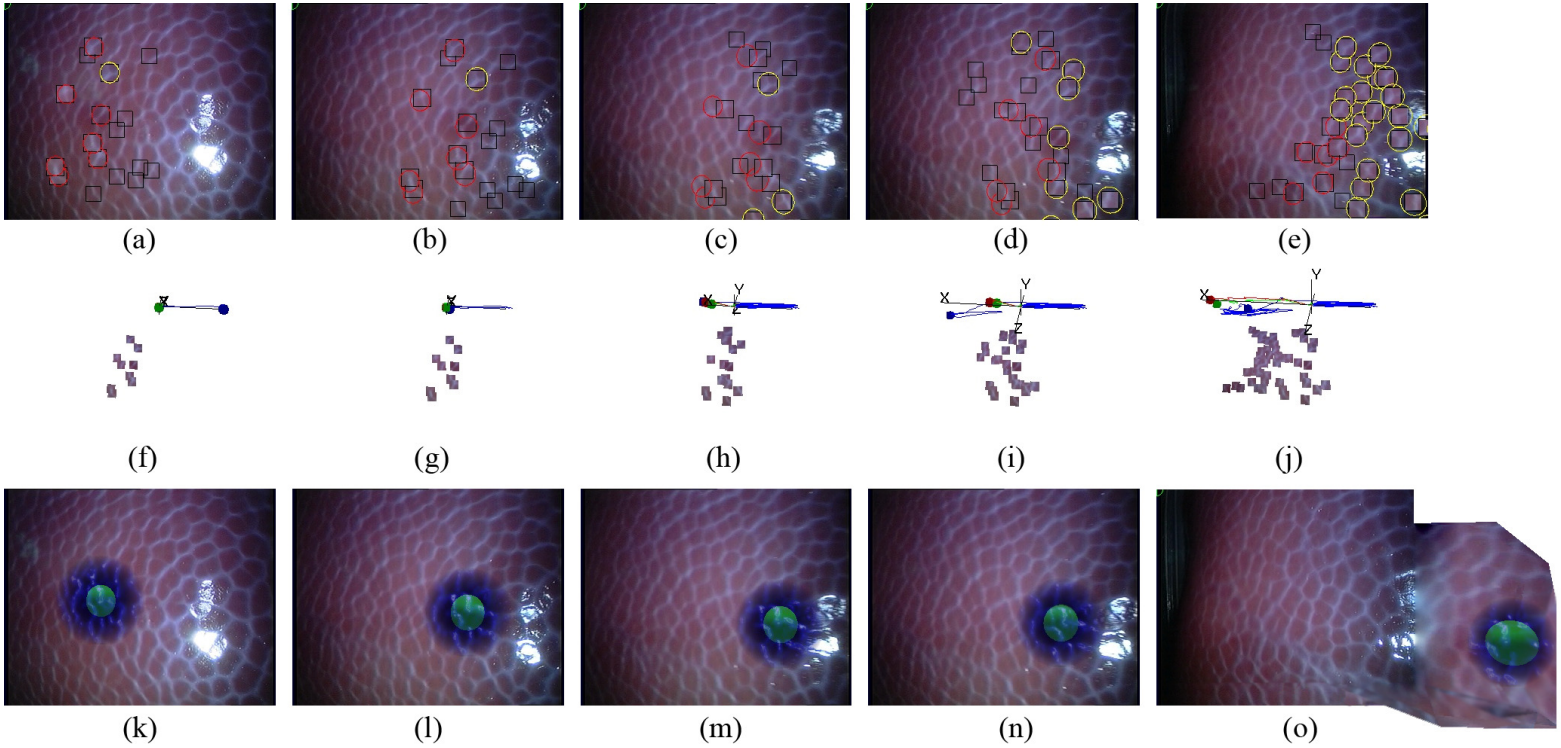


Figure 7.9 *Ex vivo* data. **(a-j)** Laparoscopic image with associated MC-SLAM map and laparoscope camera positions; MC-SLAM (green), static SLAM (blue) ground truth (red). **(k-o)** Illustration of Image Guided Surgery with pre-operative data visualised intra-operatively using Inverse Realism [43]. **(a-b)** are images taken from a static camera and illustrate the motion of the liver resulting from respiration where **(a)** is inhale position and **(b)** is exhale position. **(c-e)** Illustrate combined laparoscope and tissue motion. **(o)** laparoscope motion results in the target moving outside the current field-of-view. The dynamic target position is estimated relative to the current position of the laparoscope and visualised using view expansion described in the previous chapter.

7.2.3 *In Vivo* Experiments

An *in vivo* experiment was carried out on a MIS sequence collected during a porcine experiment. During the procedure, the surgeon navigates the laparoscope in a circular manner around the abdomen to view the liver and surrounding organs. Periodic tissue motion, resulting from respiration, is clearly visible during the procedure.

The ground truth data was not available for the *in vivo* study. The estimated respiration model is shown in **Figure 7.10 (a)**. The estimated tissue displacement resulting from respiration is 1.08 cm, and the respiration rate is estimated at 20.83 breaths per minute. **Figure 7.10 (b-d)** shows the estimated camera motion in the X, Y, and Z components using MC-SLAM and in **Figure 7.10 (e-g)** for static SLAM. The static SLAM camera position contains visible periodic oscillation that is attributed to tissue motion.

Figure 7.11 illustrates the intra-operative laparoscopic images with tracked MC-SLAM features, the MC-SLAM coordinate system, and the intra-operative images with AR visualisation. The AR visualisation is created by manually and rigidly registering a virtual tumour to the MC-SLAM map. **Figure 7.11 (k-l)** demonstrates intra-operative *in vivo* images captured using a static laparoscope. **Figure 7.11 (k)** displays the tissue position at the full exhale position, and **Figure 7.11 (l)** shows the tissue at the full inhale position of the respiration cycle. The augmented tumour's change in the position demonstrates the dynamic nature of the MC-SLAM map and progression beyond the static world assumption. In **Figure 7.11 (m-o)**, the surgeon navigates the laparoscope to explore the abdomen. Throughout this exploration, the tumour is displayed in a location being visually consistent with the surrounding tissue. This is achieved in the presence of both laparoscopic and tissue motion.

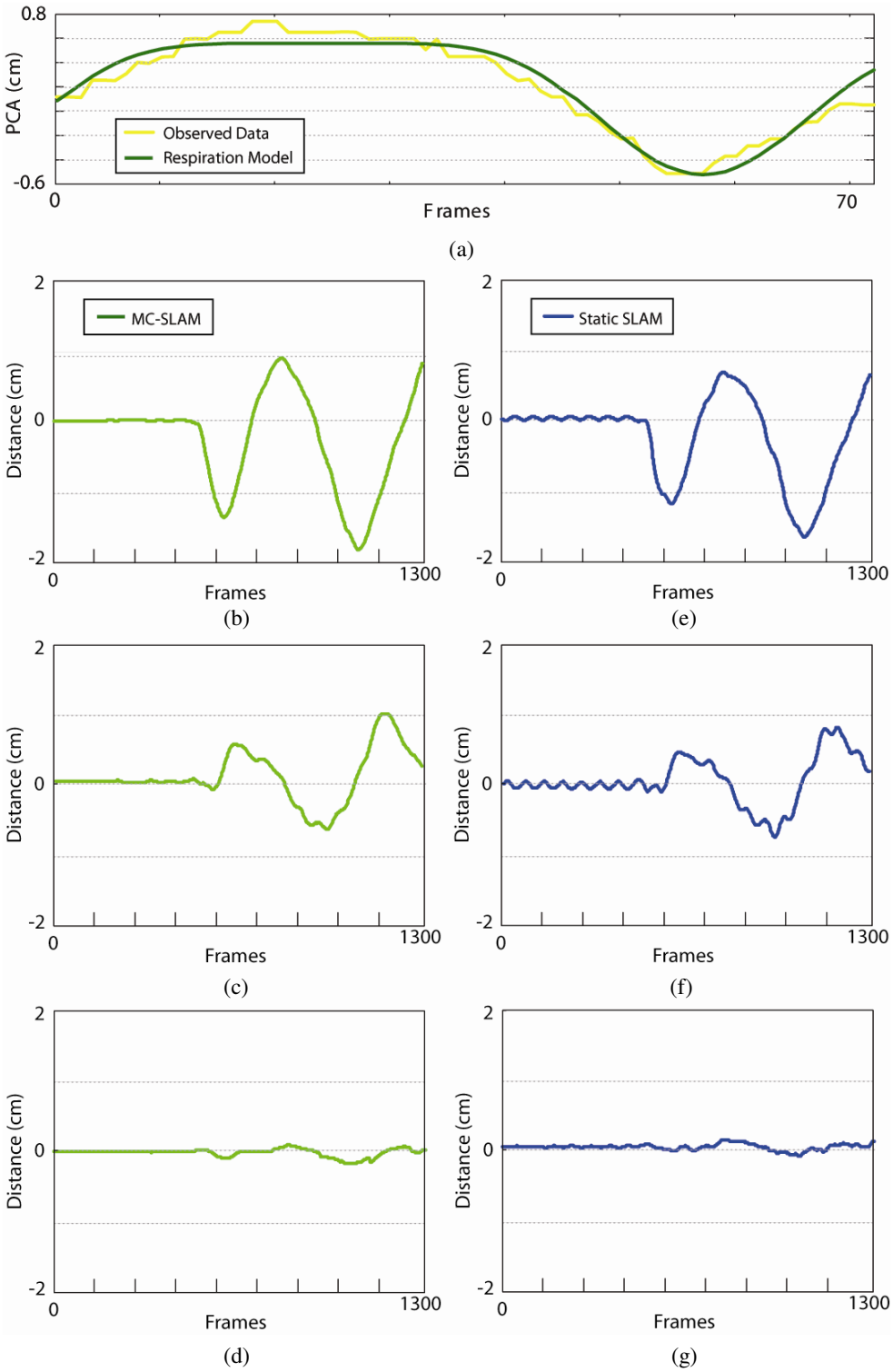


Figure 7.10 *In vivo* data. **(a)** Respiration data showing the observed data and respiration model. **(b-d)** Laparoscopic position for MC-SLAM (green). **(e-g)** Laparoscopic position static SLAM (blue)

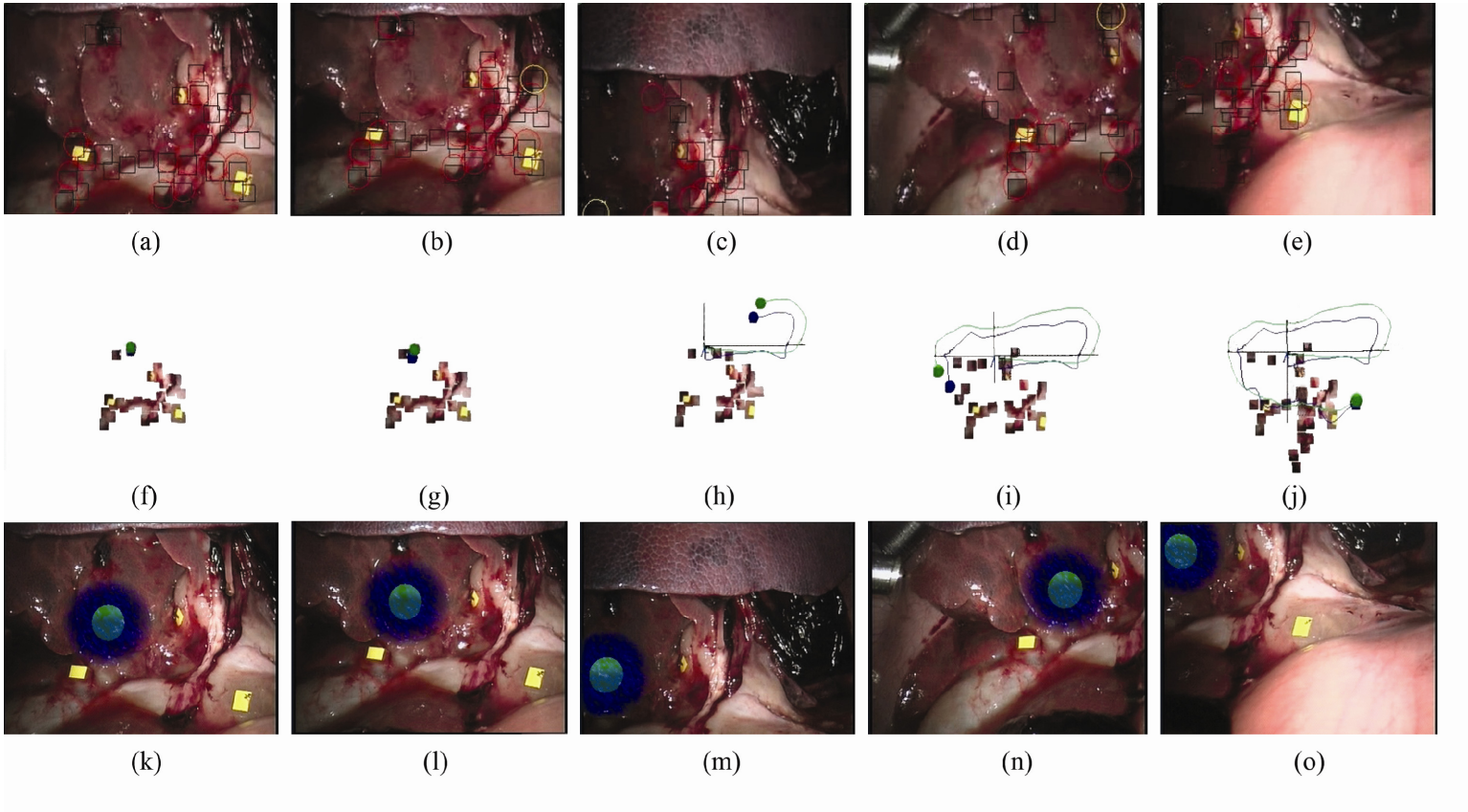


Figure 7.11 *In vivo* data showing laparoscopic images **(a-e)** with features tracked in the SLAM system. **(f-j)** The SLAM coordinate system illustrating the map features and the MC-SLAM laparoscope estimate in green and the static SLAM estimate in blue. **(k-o)** Illustration of Image Guided Surgery with pre-operative data visualised intra-operatively. Using Inverse Realism [43]. **(k-l)** show a static laparoscope and the tissue at **(k)** exhale and **(l)** inhale position. **(m-n)** combined laparoscope and tissue motion. **(o)** laparoscope motion results in the target moving outside the current field-of-view.

7.3 Discussions and Conclusion

This chapter presented MC-SLAM, a new approach to camera localisation and tissue structure estimation in a periodically deforming environment. The system extracts a model of respiration from intra-operative laparoscopic images and explicitly incorporates this high-level periodic model into the SLAM framework. This enables the method to predict and anticipate the dynamic motion of the tissue. The correlation between respiration and organ motion is exploited, thus enabling the estimation of organ motion, including when not directly observed by the intra-operative images. The system estimates the dynamic structure and camera motion both sequentially and simultaneously. This allows the system to be used for real-time applications, which is a fundamental prerequisite for IGI. Validation of the proposed method has been performed on simulated and *ex vivo* data, and its clinical relevance has been demonstrated on *in vivo* data. The current system requires an initialisation phase of at lease one respiratory cycle. Although this initialisation is short (2.88 seconds on *in vivo* data), removing or shortening this phase will make transition to the operating theatre more feasible. This may be possible by reformulating the SLAM problem, from a map containing 3D points, to a map with 3D vectors. This places additional constraints on the system and may be well-suited to deforming surfaces.

The current system uses a single model to represent the frequency and amplitude of respiration and organ motion. Initial results indicate this is suitable for modelling a region of an organ, however, the amplitude of organ motion is a function of the organ-specific tissue elasticity and distance to the diaphragm. The organ motion is approximated as locally linear. For the liver example, this model can be improved by incorporating hysteresis in the tissue motion. The dynamic tissue motion can be caused by a combination of both cardiac and respiratory cycles. The current model will not work when large instrument-tissue deformation is present. The work in this thesis has made the first useful step towards simultaneous localisation and mapping in a dynamic environment with repetitive tissue deformation, despite these challenges. Future work will focus on developing models with feature-specific amplitude and multiple frequency models with non-linear motion paths.

Chapter 8

Conclusions and Future Work

8.1 Contribution of the Thesis

The efficacy and clinical benefits of image-guided intervention are well established for procedures where manageable tissue motion, such as neurosurgery and orthopaedics, is present. Pre-operative data can be registered to patient anatomy and visualised intra-operatively in these procedures. This enables the surgeon to visualise anatomical structures below the surface of the tissue and helps guide the surgeon, thus avoiding critical structures and effectively identifying the target anatomy.

There are many situations in cardiac, abdominal, and gastrointestinal procedures where intra-operative visualisation of pre-operative data would be beneficial. Take, for example, instances involving the guidance of tumour resection margins in hepatic surgery and the identification of critical landmarks, such as the coronary artery in coronary bypass surgery. Large scale tissue deformation is common in these procedures prohibiting the accurate registration and visualisation of pre-operative and intra-operative images. Tissue deformation can be caused by the respiratory and cardiac cycles, tissue-tool interaction, organ shift, and muscle contraction during MIS. In order to successfully

co-register pre- and intra-operative data, tissue deformation must be estimated and predicted *in vivo*, *in situ*. Accurate estimation of tissue deformation and laparoscope localisation using intra-operative imaging are the main objectives of this thesis.

The work in this thesis has focussed on estimating tissue deformation from laparoscopic and endoscopic images. This approach is attractive because it does not require additional equipment in the operating theatre, surgeons are familiar with the imaging modality for navigation in MIS, and the images provide a coordinate space in which to visualise the pre-operative data using augmented reality. The task of estimating tissue deformation from laparoscopic and endoscopic cameras, however, is challenging, particularly, when combined with camera motion. The work presented in this thesis has advanced state-of-the-art procedures with the following key contributions:

A boosted tracking-by-detection framework for recovering tissue deformation using systematic image descriptor evaluation, selection, and fusion;

An algorithm for learning contextually specific information to improve tissue tracking online using unlabeled data;

A SLAM system to simultaneously estimate laparoscope motion and 3D tissue structure using stereo cameras and robust region matching;

Optical Biopsy Mapping; A method for registering multi-modality images to a common coordinate system for Augmented Reality enhanced navigation;

Dynamic view expansion; Intra-operative image enhancement using photorealistic models generated via SLAM;

A novel Motion Compensated SLAM (MC-SLAM) algorithm for laparoscopic camera localisation and dynamic mapping in a periodically deforming environment.

Chapter 3 investigates the use of region tracking to estimate tissue deformation from a static camera. Current computer vision-based approaches to tissue tracking are mainly focused on recursive techniques, which do not address re-initialisation after tracking

failure and are susceptible to error propagation. A tracking-by-detection approach is proposed for deformable tissue tracking which does not require temporal information. The performance of existing region descriptors was evaluated with respect to tissue deformation. Their relative performance was determined, and a multi-descriptor fusion framework was proposed in order to boost tracking performance. The framework used a supervised machine learning approach to select a subset of complementary, high performing descriptors and fused them in a Bayesian network. The approach demonstrated increased performance and re-initialisation of region tracking after failure. The performance of the proposed method was quantitatively evaluated on simulated data and on *in vivo* data.

An approach to tissue tracking was proposed in **Chapter 4**, which learned context-specific information online to improve performance. This method increased region density and persistency, but it also remained robust to occlusion and deformation. The region tracking problem was formulated as a classification approach, and a practical solution was proposed for learning from unlabelled data. The algorithm was able to learn what information is most discriminative to separate a region from its surroundings. The method then used this information to improve the region tracking. It was demonstrated that the proposed method was capable of learning sufficiently discriminative information such that the tracker was robust to deformation, changes in scale and orientation, occlusion, and smoke resulting from diathermy. This affirms that online learning can be used to estimate deformation from intra-operative images. The practical application of this technique was demonstrated by decoupling and modelling respiratory and cardiac motion.

One of the main contributions of the thesis is the investigation of vision-based algorithms to simultaneously estimate the position of the laparoscopic camera and the structure of tissue. IGI requires this information to be available online and, preferably, to be extracted in real-time. To achieve this, the use of a stereo SLAM algorithm for MIS was proposed. **Chapter 5**, demonstrated the use of SLAM on tissue that has little or no deformation and showed that it may be used to accurately recover the structure of the tissue and position of the laparoscope. The method was robust to error propagation when revisiting previously viewed areas, and was able demonstrate loop closure. The method was validated on simulated and phantom data and applied to *in vivo* data captured during a robotically-assisted procedure.

In **Chapter 6**, two practical, clinical applications for SLAM in MIS were proposed - Optical Biopsy Mapping and dynamic view expansion. Optical Biopsy Mapping is an intra-operative navigation system registering two intra-operative imaging modalities into a single coordinate system. Data from a micro-confocal imaging probe was visualised with augmented reality for intra-operative guidance and the re-targeting of biopsy sites. The feasibility of inferring the optical biopsy site using probe tracking in the laparoscopic image was demonstrated. It was shown that the biopsy site's position can be estimated beyond the current field-of-view by representing the biopsy site stochastically in a probabilistic framework. Dynamic view expansion enhances visualisation of intra-operative images to reduce disorientation during surgery and aid navigation. In this work, it was shown that the field-of-view of a laparoscopic camera could be dynamically extended using a 3D textured model of tissue generated from image data. The visual fidelity of the augmented intra-operative visualisation was improved through the use of texture selection and blending.

A significant contribution of this thesis is the reformulation of the SLAM problem without the static world assumption, as detailed in **Chapter 7**. Instead of using a static spatial frame of reference, the integration of periodic motion models based on biological signals in the SLAM framework, was proposed. This Motion Compensated SLAM (MC-SLAM) framework is capable of performing dynamic mapping and camera localisation in non-static environments. This addresses a fundamental problem that has prohibited the application of existing SLAM techniques to deforming tissue in MIS. Although the constraints imposed, particularly in the periodic assumption of the motion model, are still strong, this was the first attempt, to the author's knowledge, that used SLAM in dynamic *in vivo* environments.

8.2 Potential Future Work

Throughout this thesis, computer vision methods for the estimation of the morphological structure of tissue and the localisation of intra-operative imaging devices during MIS have been discussed. Using intra-operative laparoscopic images to estimate tissue deformation is a challenging task. The robustness of computer vision techniques in MIS is affected by a number of factors including image quality, occlusion, tissue deformation,

changes in scale and orientation, specular highlights, rapid camera motion, and the paucity of salient image features. The paucity of features is a significant factor that limits the methods used throughout this thesis. Organs that are well textured and contain surface features, such as vasculature, are well-suited to region tracking. Many areas of tissue, however, are homogenous and provide no distinctive, traceable information. In such cases, the current region tracking algorithms cannot be applied and alternative solutions must be proposed.

Time-of-flight cameras [49] and structured light [47] [48] are potential solutions for dealing with feature paucity. Unfortunately, these methods only produce structural information and offer no additional data for use when tracking a given point on the tissue surface. Techniques such as narrow band imaging [267], which has the capacity to extract more information than standard laparoscopic images, may offer a way forward. Regardless of the optical devices employed, the captured images can be adversely affected by interaction between the device and the MIS environment. Lens-tissue contact can obstruct the view, effect illumination, and can prevent the device from capturing useful information. It can also lead to soiling of the lens and occlusion. It is not feasible to model all the scenarios experienced during MIS nor is the use of computer vision practical in extreme situations such as lens obstruction. In these cases, recovering from system failure is essential to enable the surgeon to continue the procedure. This represents a significant research challenge in delivering practical and sustainable vision-based systems to the operating theatre, particularly for deforming environments.

In this thesis, it has been demonstrated that SLAM, formulated without the static world assumption, can be performed on periodically deforming tissue. Information from additional sensors such as gyroscopes or optical trackers may improve the camera position estimate, assuming this additional hardware can be incorporated into the surgical theatre. For deformation correlated to the respiratory or cardiac cycles, information from the ventilator, or ECG, can be incorporated to constrain the problem [137]. Tissue deformation; however, can also be caused by instrument-tissue interaction and involuntary muscular contraction. Within the MC-SLAM framework, small deformation may be identified as outliers. One of the major challenges of IGI for MIS is the theoretical treatment and modelling of large scale tissue deformation. This is likely to require prior knowledge of tissue characteristics.

Prior knowledge of patient-specific tissue morphology can be acquired from pre-operative data and represented using a variety of statistical shape and biomechanical models. The incorporation of these models into IGI will be an important research topic in future years. Full solutions to biomechanical models are challenging to achieve, and the real-time requirement places restrictions on the computations that can be performed. In this case, the use of biomechanical modelling, combined with intra-operative surface deformation, can be used to constrain the model and significantly reduce computational complexity. This research topic is pursued in [268] under the general framework of Image Constrained Biomechanical Modelling (ICBM). ICBM can be used to constrain the registration problem; however, non-rigid registration remains a challenging research area. The clinical uptake of non-rigid techniques is dependent on the development of new and rigorous validation methods [4].

There is also a requirement for systems to go beyond simply adapting to the changing surgical environment of MIS. New methods capable of understanding and predicting dynamic tissue motion are required. Cardiac procedures are particularly important examples. The stability of the operating field is affected by the cardiac and respiratory cycles. Tissue motion caused by these cycles can be modelled and predicted using periodic and quasi-periodic models. These models can also be used to control robotic instruments and to achieve motion compensation where the instruments are synchronised with the physiological motion of the tissue. Theoretically, this cancels out the periodic motion of the organs, thus enabling the surgeon to operate on a visually static heart. In practice, however, mechanical motion compensation can only be performed relative to a single point on the surface of the tissue, and non-linear tissue deformation causes residual motion in the surrounding area. The localisation of the stabilisation point can be controlled intelligently with gaze contingent motor channelling. This approach can been intuitively incorporated into the surgical work-flow, thus enabling the stabilisation point to be dynamically altered based on the surgeon's focus of attention. In addition, residual motion is observed in the peripheral vision reducing visible errors [269].

Surgical robotics has an important role to play in the future of IGI for MIS. Robotic control is fundamental to the implementation of dynamic active constraints. Dynamic active constraints use intra-operative deformation estimation to register *no-go* zones defined using pre-operative data and further refined during the intra-operative process. Robotic control is used to prevent the surgeon from manoeuvring the tools into these

zones and damaging critical structures. A fundamental component of active constraints is haptic feedback. The incorporation of haptic feedback into the robotic control interface is not trivial. The forces acting on surgical instrumentation are caused by interaction with tissue and other instruments, including the trocar. Sensing and decoupling these forces requires miniaturised, embedded sensors, which are biocompatible and easily sterilised. Current research has evaluated the influences of haptic perception [270] and demonstrated the benefits [135], however, further study is required for integrating haptics into robotic systems with clinical applications.

Human-computer or human-robot interactions are important components for consideration when translating research to the operating theatre. Traditional interfaces, such as keyboards and mice, are inappropriate for the operating theatre as the surgeons must use their hands to perform surgery. This makes developing simple and intuitive interfaces for the surgical theatre challenging and is further complicated by the complex, surgical work-flow that varies greatly between surgeons. Brain-machine interfaces and gaze-contingent motor channelling offer elegant solutions to the interface problem. Unlike traditional techniques, they have the potential to provide more information, such as the attention and focus of the surgeon. Gaze-contingent motor channelling has been proposed for visual servoing [271] and for motion compensation [272] in surgery.

A fundamental component of the interface is visualisation. Information for surgical guidance must be effectively displayed to the surgeon. AR provides an intuitive and promising mode of visualisation and data fusion for MIS. Head-mounted and auto-stereoscopic displays are attractive visualisation methods, however, they are not yet sufficiently advanced to be integrated into the surgical theatre and require the introduction of additional equipment. Augmenting the endoscopic or laparoscopic images offers the simplest solution for MIS as it can be easily incorporated into the surgical work flow. The correct visualisation of depth in AR remains a challenging research area. Incorrect or conflicting depth cues can lead to misinterpretation, nausea, and fatigue. The accuracy of the visualisation must also be taken into account. The alignment of physical and virtual objects in AR, achieved using sensors regardless of type, will contain noise and inaccuracies. Measuring the error and displaying it to the surgeon is essential to enable the formation of informed decisions [39].

This thesis has presented methods for tissue tracking and laparoscope localisation based on laparoscopic images only. Estimating tissue deformation and laparoscope position are fundamental prerequisites for the advancement of image-guided navigation of gastrointestinal, cardiac, and abdominal surgeries. IGI has the potential to increase the current functional capabilities of MIS in these procedures, thus enabling new procedures, increasing safety, and reducing operation times.

Bibliography

1. Burrows EH. Pioneers and Early Years. A History of British Radiology. Colophon Press, Alderney; 1986.
2. Horsley V, Clarke RH. The structure and functions of the cerebellum examined by a new method. Brain. 1908; 31:45–124.
3. Hounsfield GN. Computerized transverse axial scanning (tomography). 1. Description of system. British Journal of Radiology. 1973; 46(552):1016-1022.
4. Peters TM, Cleary KR. Image-Guided Interventions: Technology and Applications. Springer; 2008; p. 560.
5. Friets EM, Strohbehn JW, Hatch JF, Roberts DW. A frameless stereotaxic operating microscope for neurosurgery. IEEE Transactions on Biomedical Engineering. 1989; 36(6):608-617.
6. Roberts DW, Strohbehn JW, Hatch JF, Murray W, Kettenberger H. A frameless stereotaxic integration of computerized tomographic imaging and the operating microscope. Journal of Neurosurgery. 1986; 65(4):545-549.
7. Kosugi Y, Watanabe E, Goto J, Watanabe T, Yoshimoto S, Takakura K, et al. An articulated neurosurgical navigation system using MRI and CT images. IEEE Transactions on Biomedical Engineering. 1988; 32(2):147-152.
8. Watanabe E, Watanabe T, Manaka S, Mayanagi Y, Takakura K. Three-dimensional digitizer (neuronavigator): new equipment for computed tomography-guided stereotaxic surgery. Surgical Neurology. 1987; 27(6):543-547.
9. Galloway R, Edwards C, Haden G, Maciunas R An interactive, image-guided articulated arm for laser surgery. Strategic Defense Initiative Organization's Fourth Annual Meeting on Mefical Free Electron Lasers; 1989; Dallas, TX.
10. Adams L, Krybus W, Meyer-Ebrecht D, Rueger R, Gilsbach JM, Moesges R, et al. Medical Imaging: Computer-Assisted Surgery. IEEE Computer Graphics and Applications. 1990; 10(3):43-51.
11. Yang G-Z, Hu XP. Multi-Sensor Fusion Body Sensor Networks. In. London; 2006. p. 239 - 286.
12. Guo T, Finnis KW, Parrent AG, Peters TM. Visualization and navigation system development and application for stereotactic deep-brain neurosurgeries. Computer Aided Surgery. 2006; 11(5):231-239.
13. Thompson PM, Toga AW. Warping strategies for intersubject registration. In: Handbook of medical imaging; 2000. p. 569 - 601.
14. Fahlbusch R, Nimsky C. Intraoperative MRI developments. Neurosurgery Clinics of North America. 2005; 16(1):11-13.
15. Baumhauer M, Feuerstein M, Meinzer HP, Rassweiler J. Navigation in endoscopic soft tissue surgery: perspectives and limitations. Journal of Endourology. 2008; 22(4):751-766.
16. Van-Dam J. Novel methods of enhanced endoscopic imaging. Gut. 2003; 52(4):12-16.
17. Taylor RH, Mittelstadt BD, Paul HA, Hanson W, Kazanzides P, Zuhars JF, et al. An Image-Directed Robotic System for Precise Orthopaedic Surgery. Transactions on Robotics and Automation. 1994; 10(3):261-276.
18. Davies B.L., Harris S.J., Lin W.J., Hibberd R.D., Middleton R., Cobb J.C. Active compliance in robotic surgery—the use of force control as a dynamic constraint. Inst Mech Eng; 1997 p. 285–292.
19. Davies B.L., Hibberd R.D., Coptcoat M.J., Wickham J.E. A surgeon robot prostatectomy—a laboratory evaluation. Med Eng Technol. 1989; (13):273–277.
20. Davies B.L., Hibberd R.D., Fan K.L., Jakopec M., Harris S.J. ACROBOT-Using Robots and Surgeons synergistically in Knee Surgery. International Conference on Advanced Robotics.; 1997 p. 173-180.

21. Wengert C, Cattin PC, Duff JM, Székely G Markerless Endoscopic Registration and Referencing. Medical Image Computing and Computer-Assisted Intervention 2006 p. 816--823.
22. Kwartowitz DM, Miga M, Herrel SD, Galloway RL. Towards image guided robotic surgery: multi-arm tracking through hybrid localization International Journal of Computer Assisted Radiology and Surgery. 2009; 4:281-286.
23. Barnett GH, Kormos DW, Steiner CP, Weisenberger J. Intraoperative localization using an armless, frameless stereotactic wand. Technical note. Journal of Neurosurgery. 1993; 78(5):510-514.
24. Reinhardt HF, Horstmann GA, Gratzl O. Sonic stereometry in microsurgical procedures for deep-seated brain tumors and vascular malformations. Neurosurgery. 1993; 32(1):51-57.
25. Wengert C, Bossard L, Häberling A, Baur C, Székely G, Cattin PC Endoscopic Navigation for Minimally Invasive Suturing. Medical Image Computing and Computer Assisted Intervention; 2007 p. 620–627.
26. Tsai R, Lenz R Real Time Versatile Robotic Hand/Eye Calibration using 3D Machine Vision. International Conference on Robotics and Automation; 1988 p. 554-561.
27. Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics. 1987; 21(4):163-169.
28. Besl P, McKay N. A Method for Registration of 3-D Shapes. Pattern Analysis and Machine Intelligence. 1992; 14:239–256.
29. Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based Registration of Medical Images: A Survey. IEEE Transactions on Medical Imaging. 2003; 22(8):986-1004.
30. Wunsch P, Hirzinger G Registration of Cad-Models to Images by Iterative Inverse Perspective Matching. International Conference on Pattern Recognition; 1996 p. 78-83.
31. Gueziec A, Kazanzides P, Williamson B, Taylor RH. Anatomy-based Registration of CT-scan and Intraoperative X-ray Images for Guiding a Surgical Robot. IEEE Transactions on Medical Imaging 1998; 17(5):715-728.
32. Mori K, Deguchi D, Sugiyama J, Suenagaa Y, Toriwakia J, Jr CRM, et al. Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. Medical Image Analysis. 2002; 6(3):321-336.
33. Deligianni F, Chung AJ, Yang G-Z pq-Space Based 2D/3D Registration for Endoscope Tracking. Medical Image Computing and Computer-Assisted Intervention; 2003; Montreal, Canda. p. 311-318.
34. Bichlmeier C, Ockert B, Heining SM, Ahmadi A, Navab N Stepping into the Operating Theater: ARAV - Augmented Reality Aided Vertebroplasty. International Symposium on Mixed and Augmented Reality; 2008.
35. Jannin P, Morandi X, Fleig OJ, Rumeur EL, Toulouse P, Gibaud B, et al. Integration of Sulcal and Functional Information for Multimodal Neuronavigation. Journal of Neurosurgery. 2002; 96:713-723.
36. Edwards PJ, King AP, Jr. CRM, Cunha DAD, Hawkes DJ, Hill DLG, et al. Design and Evaluation of a System for Microscope-assisted Guided Interventions (MAGI). Transactions on Medical Imaging. 2000; 19(11):1082-1093.
37. Wacker F, Vogt S, Khamene A, Jesberger J, S SN, Elgort D, et al. An Augmented Reality System for MR Image–guided Needle Biopsy: Initial Results in a Swine Model. Radiology. 2006; 238:497-504.
38. Nicolau SA, Pennec X, Soler L, Buy X, Gangi A, Ayache N, et al. An Augmented Reality System for Liver Thermal Ablation : Design and Evaluation on Clinical Cases. Medical Image Analysis. 2009.
39. Sielhorst T, Feuerstein M, Navab N. Advanced Medical Displays: A Literature Review of Augmented Reality. Journal of Display Technology; Special Issue on Medical Displays. 2008; 4(4):451-467.
40. Hussain SA, Selway R, Harding C, Polkey CE. The urgent postoperative CT scan: a critical appraisal of its impact British Journal of Neurosurgery. 2001 15(2):116-118.
41. Aziz O, Lo B, King R, Yang G-Z, Darzi A Pervasive Body Sensor Network: An Approach to Monitoring the Post-operative Surgical Patient. Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks (BSN); 2006 p. 13-18.

42. Lo B, Atallah L, Aziz O, ElHelw M, Darzi A, Yang G-Z Real time pervasive monitoring for post operative car. Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks (BSN; 2007 p. 122-127.
43. Lerotic M, Chung AJ, Mylonas G, Yang G-Z pq -space Based Non-Photorealistic Rendering for Augmented Reality. Medical Image Computing and Computer Assisted Intervention; 2007 p. 102-109.
44. Gao L, Heath DG, Kuszyk BS, Fishman EK. Automatic liver segmentation technique for three-dimensional visualization of CT data. Radiology. 1996; 201:359–364.
45. McGahan JP, Dodd GD. Radiofrequency Ablation of the Liver: Current Status American Journal of Roentgenology. 2001; 176(1):3-16.
46. Fuchs H, Livingston MA, Raskar R, Colucci D, Keller K, State A, et al. Augmented reality visualization for laparoscopic surgery. Medical Image Computing and Computer-Assisted Intervention; 1998Springer p. 934-943.
47. Albitar C, Graebling P, Doignon C Robust Structured Light Coding for 3D Reconstruction. International Conference on Computer Vision; 2007 p. 1-6.
48. Knaus D., Friets E., Bieszczad J., Chen R., Miga M., Galloway R., et al. System for laparoscopic tissue tracking. Symposium on Biomedical Imaging: Macro to Nano; 2006 p. 498- 501.
49. Penne J, Höller K, Stürmer M, Schrauder T, Schneider A, Engelbrecht R, et al. Time-of-Flight 3-D Endoscopy. Medical Image Computing and Computer Assisted Intervention; 2009 p. 467-474.
50. Lo BPL, Visentini-Scarzanella M, Stoyanov D, Yang G-Z Belief Propagation for Depth Cue Fusion in Minimally Invasive Surgery. Medical Image Computing And Computer Assisted Intervention; 2008 p. 104-112.
51. Visentini-Scarzanella M, Mylonas GP, Stoyanov D, Yang G-Z i-BRUSH: A Gaze-Contingent Virtual Paintbrush for Dense 3D Reconstruction in Robotic Assisted Surgery Medical Image Computing and Computer-Assisted Intervention; 2009 p. 353-360.
52. Bao P, Warmath J, Galloway RJ, Herline A. Ultrasound-to-computer-tomography registration for image-guided laparoscopic liver surgery. Surgical Endoscopy. 2005; 19(3):424-429.
53. Herline AJ, Herring JL, Stefansic JD, Chapman WC, Galloway RL, Dawant BM, et al. Surface Registration for Use in Interactive Image-Guided Liver Surgery. Computer Aided Surgery. 1999; 5(1):11-17.
54. Linte CA, Moore J, Wiles AD, Wedlake C, Peters TM Targeting Accuracy under Model-to-Subject Misalignments in Model-Guided Cardiac Surgery. Medical Image Computing and Computer-Assisted Intervention; 2009 p. 361-368.
55. Mourgues F, Vieville T, Falk V, Coste-Manière È Interactive guidance by image overlay in robot assisted coronary artery bypass. Medical Image Computing and Computer-Assisted Intervention; 2003; Montréal, Canada. November. p. 173-181.
56. Falk V, Mourgues F, Adhami L, Jacobs S, Thiele H, Nitzsche S, et al. Cardio Navigation: Planning, Simulation, and Augmented Reality in Robotic Assisted Endoscopic Bypass Grafting. Annals of Thoracic Surgery. 2005; 79(6):2040-2048.
57. Figl M, Rueckert D, Hawkes D, Casula R, Hu M, Pedro O, et al. Coronary Motion Modelling for Augmented Reality Guidance of Endoscopic Coronary Artery Bypass. International Symposium on Biomedical Simulation; 2008 p. 197 - 202
58. Nicolaou M, James A, Lo BP, Darzi A, Yang G-Z Invisible shadow for navigation and planning in minimal invasive surgery. Medical Image Computing and Computer Assisted Intervention; 2005 p. 25-32.
59. Höller K, Penne J, Schneider A, Jahn J, Guttierrez J, Wittenberg T, et al. Endoscopic Orientation Correction. Medical Image Computing and Computer Assisted Intervention; 2009 p. 459-466.
60. Koppel D, Wang Y-F, Lee H Robust and Real-Time Image Stabilization and Rectification. IEEE Workshops on Application of Computer Vision; 2005 p. 350 - 355
61. Koppel D, Wang Y-F, Lee H Viewing Enhancement in Video-Endoscopy. IEEE Workshop on Applications of Computer Vision; 2002 p. 304

62. Moll M, Koninckx T, Van-Gool LJ, Koninckx PR Unrotating images in laparoscopy with an application for 30° laparoscopes 4th European Conference of the International Federation for Medical and Biological Engineering; 2009 p. 966-969.
63. Lerotic M, Chung A, Clark J, Valibeik S, Yang G-Z Dynamic View Expansion for Enhanced Navigation in Natural Orifice Transluminal Endoscopic Surgery. Medical Image Computing and Computer Assisted Intervention; 2008 p. 467 - 475
64. Nicolau SA, Soler L, Marescaux J Augmented Reality Systems for Medical Interventions: Current Limits World Congress on Medical Physics and Biomedical Engineering; 2009 p. 1635-1638.
65. Helferty JP, Sherbondy AJ, Kiraly AP, Higgins WE System for Live Virtual-Endoscopic Guidance of Bronchoscopy. Computer Vision and Pattern Recognition; 2005 p. 68-76.
66. Mori K, Deguchi D, Kitasaka T, Suenaga Y, Takabatake H, Mori M, et al. Bronchoscope Tracking Based on Image Registration Using Multiple Initial Starting Points Estimated by Motion Prediction. Medical Image Computing and Computer Assisted Intervention; 2006 p. 645-652.
67. Feuerstein M, Mussack T, Heining SM, Navab N. Intraoperative Laparoscope Augmentation for Port Placement and Resection Planning in Minimally Invasive Liver Resection. Transactions on Medical Imaging. 2008; 27(3):355-69.
68. Nicolau SA, Goffin L, Soler L A Low Cost and Accurate Guidance System for Laparoscopic Surgery: Validation on an abdominal phantom. ACM Symposium on Virtual Reality Software and Technology; 2005 p. 124-133.
69. Ukimura O, Gill IS. Imaging-Assisted Endoscopic Surgery : Cleveland Clinic Experience. Journal of Endourology. 2008; 22(4):803-809.
70. Su L, Vagvolgyi B, Agarwal R, Reiley C, Taylor R, Hager G. Augmented Reality During Robot-assisted Laparoscopic Partial Nephrectomy: Toward Real-Time 3D-CT to Stereoscopic Video Registration. Journal of Urology. 2009; 73(4):896-900.
71. Teber D, Baumhauer M, Simpfendoerfer T, Hruza M, Klein J, Rassweiler J. Augmented reality: a new tool to improve surgical accuracy during laparoscopic partial nephrectomy? Preliminary in vitro and in vivo results. European Urology Supplements. 2009; 7(3):258-258.
72. Navab N, Feuerstein M, Bichlmeier C Laparoscopic Virtual Mirror: New Interaction Paradigms for Monitor Based Augmented Reality. IEEE Virtual Reality Conference; 2007 p. 43-50.
73. Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge Press; 2000.
74. Zhang Z. A Flexible New Technique for Camera Calibration. Pattern Analysis and Machine Intelligence. 2000; 22(11):1330-1334.
75. Tsai RY An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. Conference on Computer Vision and Pattern Recognition; 1986; Miami Beach, FL. p. 364-374.
76. Stoyanov D. Camera Calibration Tools; 2009.
77. Bouguet J-Y. Camera Calibration Toolbox for Matlab; 2004.
78. Stoyanov D, Darzi A, Yang G-Z Duncan JS, Gerig G, editors. Laparoscope Self-calibration for Robotic Assisted Minimally Invasive Surgery. Medical Image Computing and Computer Assisted Intervention; 2005Springer-Verlag p. 114-121.
79. Marr JL. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: W. H. Freeman and Company; 1982.
80. Forster CHQ, Tozzi C Towards 3D Reconstruction of Endoscope Images Using Shape from Shading. Brazilian Symposium on Computer Graphics and Image Processing; 2000 p. 90-96.
81. Tankus A, Sochen N, Yeshurun Y Perspective Shape-from-Shading via Fast Marching. IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2004.
82. Prados E, Faugeras O Shape From Shading: A Well Posed Problem? Computer Vision and Pattern Recognition; 2005; San Diego, USA. p. 870-877.
83. Rashid HU, Burger P. Differential algorithm for the determination of shape from shading using a point light source. Image and Vision Computing. 1992; 10(2):119-127.

84. Okatani T, Deguchi K. Shape Reconstruction from an Endoscope Image by Shape from Shading Technique for a Point Light Source at the Projection Centre. Computer Vision and Image Understanding. 1997; 66(2):119-131.
85. Devernay F, Mourgues F, Coste-Maniere E Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. Medical Imaging and Augmented Reality; 2001.
86. Lau WW, Ramey NA, Corso J, Thakor NV, Hager GD Stereo-Based Endoscopic Tracking of Cardiac Surface Deformation. Medical Image Computing and Computer Assisted Intervention; 2004; St. Malo, France. p. 494-501.
87. Richa R, . Bo, A, P, L,. Poignet, P,. Motion prediction for tracking the beating heart. EMBC; 2008 p. 3261-3264.
88. Richa R, Poignet P, Liu C Efficient 3D tracking for motion compensation in beating heart surgery. Medical Image Computing and Computer Assisted Intervention; 2008 p. 684-691.
89. Stoyanov D, Mylonas GP, Deligianni F, Darzi A, Yang G-Z Soft-tissue Motion Tracking and Structure Estimation for Robotic Assisted MIS Procedures. Medical Image Computing and Computer Assisted Intervention; 2005 p. 139-146.
90. Mourgues F, Devernay F, Malandain G, Coste-Manière È 3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery. International Symposium on Augmented Reality; 2001; New York, USA. October.
91. Hager G, Vagvolgyi B, Yuh D Stereoscopic Video Overlay with Deformable Registration. Medicine Meets Virtual Reality; 2007.
92. Sauvée M, Noce A, Poignet P, Triboulet J, Dombre E. Three-dimensional heart motion estimation using endoscopic monocular vision system: From artificial landmarks to texture analysis Biomedical Signal Processing and Control. 2007; 2(3):199-207
93. Ginhoux R., Gangloff J.A., de Mathelin M.F., Soler L., Sanchez M.M.A., Marescaux J. Beating heart tracking in robotic surgery using 500 Hz visual servoing, model predictive control and an adaptive observer. International Conference on Robotics and Automation; 2004 p. 274- 279.
94. Sauvee M, Poignet P, Triboulet J, Dombre E, Malis E, Demaria R 3D Heart Motion Estimation using Endoscopic Monocular Vision System. 6th IFAC Symposium on Modelling and Control in Biomedical Systems; 2006.
95. Stoyanov D, Darzi A, Yang G-Z Dense 3D Depth Recovery for Soft Tissue Deformation During Robotically Assisted Laparoscopic Surgery. Medical Image Computing and Computer Assisted Intervention; 2004Springer-Verlag p. 41-48.
96. Richa R, Poignet P, Liu C Deformable motion tracking of the heart surface. International Conference on Intelligent Robots and Systems; 2008 p. 3997-4003.
97. Wu C, Narasimhan SG, Jaramaz B. A Multi-Image Shape-from-Shading Framework for Near-Lighting Perspective Endoscopes. International Journal of Computer Vision. 2009.
98. Haneishi H, Ogura T, Miyake Y. Profilometry of a gastrointestinal surface by an endoscope with laser beam projection. Optics Letters. 1994; 19(9):601-603.
99. McKinlay R, Shaw M, Park A. A technique for real-time digital measurements in laparoscopic surgery. Surgical Endoscopy. 2004; 18:709–712.
100. Hayashibe M, Suzuki N, Nakamura Y. Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management. Medical Image Analysis. 2006; 10:509-519.
101. Masson N, Nageotte F, Zanne P, Mathelin Md In Vvio Comparison of Real-time Tracking Algorithms for Interventional Felexible Endoscopy. ISBI; 2009.
102. Ott L, Zanne P, Nageotte F, Mathelin Md, Gangloff J Physiological motion rejection in flexible endoscopy using visual servoing. IEEE International Conference on Robotics and Automation; 2008 p. 2928-2933.
103. Giannarou S, Visentini-Scarzanella M, Yang G-Z Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery. IEEE international symposium on biomedical imaging; 2009 p. 1059–1062.
104. Harris CG, Stephens M A combined corner and edge detector. Alvey Vision Conference; 1988; Manchester, UK. p. 147-151.
105. Shi J, Tomasi C Good features to track. Computer Vision and Pattern Recognition; 1994 p. 593-600.

106. Stoyanov D, Yang G-Z Stabilization of Image Motion for Robotic Assisted Beating Heart Surgery. Medical Image Computing and Computer Assisted Intervention; 2007 p. 417-424.
107. Hu M, Penney GP, Edwards PJ, Figl M, Hawkes DJ 3D Reconstruction of Internal Organ Surfaces for Minimal Invasive Surgery. Medical Image Computing and Computer Assisted Intervention; 2007 p. 68-77.
108. Gröger M, Ortmaier T, Sepp W, Hirzinger G Tracking local motion on the beating heart. SPIE Medical Imaging Conference; 2002; San Diego, USA. p. 233-241.
109. Matas J., Chum O., Martin U., Pajdla T. Robust wide baseline stereo from maximally stable extremal regions. British Machine Vision Conference; 2002 p. 384-393.
110. Lindeberg T. . Feature detection with automatic scale selection. Computer Vision. 1998; 2(30):77-116.
111. Lowe D.G. Object recognition from local scale-invariant features. International Conference on Computer Vision; 1999 p. 1150-1157.
112. Bay H, Tuytelaars T, Van-Gool L SURF: Speeded Up Robust Features. 2006.
113. Lepetit V, Fua P. Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. Foundations and Trends in Computer Graphics and Vision. 2005; 1(1):1–89.
114. Baker S, Matthews I. Lucas-Kanade 20 years on: A unifying framework. International Journal of Computer Vision. 2004; 56(3):221–255.
115. Lucas BD, Kanade T An Iterative Image Registration Technique with an Application to Stereo Vision. International Joint Conference on Artificial Intelligence; 1981 p. 674–679.
116. Bradski GR, Kaehler A. Learning OpenCV Computer Vision with the OpenCV Library. 2008; p. 575
117. Hager G, Belhumeur PN. Efficient Region Tracking With Parametric Models of Geometry and Illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1999; 20(10):1-15.
118. Masson N, Nageotte F, Zanne P, Mathelin Md, Marescaux J Comparison of Visual Tracking Algorithms on In Vivo Sequences for Robot-Assisted Flexible Endoscopic Surgery. EMBC; 2009 p. 5571-5576.
119. Benhimane S, Malis E Real-time image-based tracking of planes using efficient second-order minimization. Intelligent Robots Systems; 2004 p. 943-948.
120. Comaniciu D, Ramesh V, Meer P. Kernel-Based Object Tracking. Pattern Analysis and Machine Intelligence. 2003; (25):564–577.
121. Bradski GR, . Computer video face tracking for use in a perceptual user interface. Intel Technology Journal. 1998; Q2.
122. Ott L, Zanne P, Nageotte F, Mathelin Md, Gangloff J Physiological motion rejection in flexible endoscopy using visual servoing. International Conference on Robotics and Automation; 2008 p. 2928-2933.
123. Ott L, Nageotte F, Zanne P, Mathelin Md Physiological motion rejection in flexible endoscopy using visual servoing and repetitive control : Improvements on non-periodic reference tracking and non-periodic disturbance rejection International Conference on Robotics and Automation; 2009 p. 4233 - 4238.
124. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 2004; 60(2):91-110.
125. Noce A, Triboulet J, Poignet P Efficient Tracking of the Heart Using Texture. EMBC; 2007 p. 4480 – 4483.
126. Mirota D, Wang H, Taylor R, Ishii M, Hager G Towards Video-based Navigation for Endoscopic Endonasal Skull Base Surgery Medical Image Computing and Computer Assisted Intervention 2009 p. 91–99.
127. Fischler MA, Bolles RC. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the Association for Computing Machinery. 1981; 24:381-395.
128. Riviere CN, Gangloff J, Mathelin Md. Robotic Compensation of Biological Motion to Enhance Surgical Accuracy. Proceedings of the IEEE. 2006; 94(9):1705-1716.
129. Davies SC, Hill AL, Holmes RB, Halliwell M, Jackson PC. Ultrasound quantitation of respiratory organ motion in the upper abdomen. British Journal of Radiology. 1994; 67:1096–1102.

130. Hostettler A, Nicolau S, Forest C, Soler L, Remond Y Harders M, Székely C, editors. Real Time Simulation of Organ Motions Induced by Breathing: First Evaluation on Patient Data. International Symposium on Biomedical Simulation 2006 p. 9-18.
131. Ginhoux R, Gangloff J, Mathelin Md, Soler L, Sanchez MMA, Marescaux J. Active Filtering of Physiological Motion in Robotized Surgery Using Predictive Control. IEEE Transactions on Robotics. 2005; 21(1):67-79.
132. Lujan AE, Larsen EW, Balter JM, Haken. RKT. A Method for Incorporating Organ Motion due to Breathing into 3D Dose Calculations. Medical Phisics. 1999; 26(5):715-720.
133. Pedro O, Edwards P, Rueckert D Cardio-Respiratory Motion Decomposition for Myocardial Motion Modelling in TECAB The MICCAI workshop on Image Guidance and Computer Assistance for Soft-Tissue Interventions; 2008.
134. Ortmaier T, Gröger M, Boehm DH, Falk V, Hirzinger G. Motion Estimation in Beating Heart Surgery. IEEE Transactions on Biomedical Engineering. 2005; 52(10):1729-1740.
135. Pezzementi Z, Ursu D, Misra S, Okamura AM Modeling realistic tool–tissue interactions with haptic feedback: a learning-based method. Symposium on haptic interfaces for virtual environment and teleoperator systems (Haptics 2008); 2008 p. 209–215.
136. Hu M, Penney GP, Rueckert D, Edwards PJ, Bello R, Casula R, et al. Non-rigid Reconstruction of the beating Heart Surface for Minimally Invasive Cardiac Surgery. Medical Image Computing and Computer Assisted Intervention; 2009 p. 34-42.
137. Cuvillon L, Gangloff J, deMathelin M, Forgione A Toward Robotized Beating Heart TECABG: Assessment of the Heart Dynamics Using High-Speed Vision. Medical Image Computing and Computer Assisted Intervention; 2005 p. 551-558.
138. Stoyanov D, Darzi A, Yang G-Z. A Practical Approach Towards Accurate Dense 3D Depth Recovery for Robotic Laparoscopic Surgery. Computer Aided Surgery. 2005; 10(4):199-208.
139. Noce A, Triboulet J, Poignet P, Dombre E Texture Features Selection for Visual Servoing of the Beating Heart. BioRob 2006 p. 335 – 340.
140. Richa R, Bo APL, Poignet P Motion Prediction for Tracking the Beating Heart. EMBC; 2008 p. 3261-3264.
141. Gröger M, Hirzinger G Optical flow to analyse stabilised images of the beating heart. International Conference on Computer Vision Theory and Applications 2006 p. 237 - 244.
142. Ginhoux R, Gangloff JA, Mathelin Md, Soler L, Sanchez MMA, Marescaux J Beating heart tracking in robotic surgery using 500 Hz visual servoing, model predictive control and an adaptive observer. International Conference on Robotics and Automation; 2004 p. 274- 279.
143. Koppel D, Wang YF, Lee H. Image-based rendering and modeling in video-endoscopy. IEEE International Symposium on Biomedical Imaging: Macro to Nano. 2004:269-272 Vol. 1.
144. Igarashi T, Suzuki H, Naya Y. Computer based endoscopic image processing technology for endourology and laparoscopic surgery. International Journal of Urology. 2009; (16):533–543.
145. Garcıa O, Civera J, Gueme A, Munoz V, Montiel JMM Real-time 3D Modeling from Endoscope Image Sequences. ICRA Workshop on Advanced Sensing and Sensor Integration in Medical Robotics; 2009
146. Atasoy S, Noonan DP, Benhimane S, Navab N, Yang G-Z A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis. Medical Image Computing and Computer Assisted Intervention; 2008 p. 850-7.
147. Koppel D, Chen C-I, Wang Y-F, Lee H, Gu J, Poirson A, et al. Toward automated model building from video in computer-assisted diagnoses in colonoscopy. SPIE; 2007.
148. Zhou J, Das A, Li F, Li B Circular generalized cylinder fitting for 3D reconstruction in endoscopic image based MRF. Computer Vision and Pattern Recognition Workshops; 2008 p. 1-8.
149. Seshamani S, Lau W, Hager G Real-time endoscopic mosaicking. Medical Image Computing and Computer Assisted Intervention; 2006 p. 355-63.
150. Wu C-H, Sun Y-N, Chang C-C. Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. IEEE Transactions on Biomedical Engineering. 2007; 54(7):1199-1211.

151. Miranda-Luna R, Daul C, Blondel WCPM, Hernandez-Mier Y, Wolf D, Guillemin F. Mosaicing of Bladder Endoscopic Image Sequences: Distortion Calibration and Registration Algorithm. IEEE Transactions on Biomedical Engineering. 2008; 55(2):541-553.
152. Behrens A. Creating Panoramic Images for Bladder Fluorescence Endoscopy. Acta Polytechnica Journal of Advanced Engineering. 2008; 48(3):50-54.
153. Olijnyk S, Mier YH, Blondel WCPM, Daul C, Wolf D, Bourg-Heckly G. Combination of panoramic and fluorescence endoscopic images to obtain tumor spatial distribution information useful for bladder cancer detection. Progress in Biomedical Optics and Imaging. 2007; 8.
154. Seibel EJ, Carroll RE, Dominitz JA, Johnston RS, Melville CD, Lee CM, et al. Tethered Capsule Endoscopy, A Low-Cost and High-Performance Alternative Technology for the Screening of Esophageal Cancer and Barrett's Esophagus. IEEE Transactions on Biomedical Engineering. 2008; 55(3):1032-1042.
155. Carroll RE, Seitz SM Rectified Surface Mosaics. International Conference on Computer Vision; 2007 p. 1-8.
156. Castaneda V, Atasoy S, Mateus D, Navab N, Meining A Reconstructing the Esophagus Surface from Endoscopic Image Sequences. Russian Bavarian Conference on Bio-Medical Engineering; 2009.
157. Wang H, Mirota D, Hager GD, Ishii M. Anatomical reconstruction from endoscopic images: toward quantitative endoscopy. American Journal of Rhinology. 2008 (22):47-51.
158. Wang H, Mirota D, Ishii M, Hager GD Robust Motion Estimation and Structure Recovery from Endoscopic Image Sequences With an Adaptive Scale Kernel Consensus Estimator. Computer Vision and Pattern Recognition; 2008 p. 1-7.
159. Burschka D, Li M, Ishii M, Taylor R, Hager GD. Scale-invariant Registration of Monocular Endoscopic Images to CT-scans for Sinus Surgery. Medical Image Analysis. 2005; 9(5):413 - 426.
160. Burschka D, Li M, Taylor R, Hager GD Scale-invariant registration of monocular stereo images to 3D surface models. International Conference on Intelligent Robots and Systems; 2004 p. 2581-2586.
161. Burschka D, Li M, Taylor R, Hager GD Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery. Medical Image Computing and Computer Assisted Intervention; 2004 p. 413–421.
162. Konen W, Naderi M, Scholz M Endoscopic image mosaics for real-time color video sequences. Computer Assisted Radiology and Surgery; 2007; Berlin. p. 329-338.
163. Atasoy S, . Noonan, D, P,. Benhimane, S,. Navab, N,. Yang, G-Z,. A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis. In Proc MICCAI; 2008 p. 850-7.
164. Koppel D, Wang YF, Lee H. Image-based rendering and modeling in video-endoscopy. Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on. 2004:269-272 Vol. 1.
165. Hu M, . Penney, G,. Edwards, P,. Figl. M,. Hawkes, M,J,. 3D Reconstruction of Internal Organ Surfaces for Minimal Invasive Surgery MICCAI 2007 p. 68-77.
166. Wu C-H, . Sun, Y-N., Chen, Y-C,. Chang, C-C,. . Endoscopic Feature Tracking and Scale-Invariant Estimation of Soft-Tissue Structures. IEICE TRANSACTIONS on Information and Systems E91-D(2):351-360.
167. Brand M Morphable 3d models from video. Conference on Computer Vision and Pattern Recognition; 2001 p. 456-463.
168. Xiao J, Chai J, Kanade T A Closed-Form Solution to Non-Rigid Shape and Motion Recovery. European Conference on Computer Vision; 2004.
169. Torresani L, Yang DB, Alexander EJ, Bregler C Tracking and Modeling Non-Rigid Objects with Rank Constraints. Computer Vision and Pattern Recognition p. 493.
170. Durrant-Whyte H, Bailey T. Simultaneous localization and mapping (SLAM): Part I: The essential algorithms. Robotics & Automation Magazine. 2006; 13(2):99-108.
171. Bailey T. Mobile Robot Localisation and Mapping in Extensive Outdoor Environments thesis]; 2002.

172. Smith R., Cheeseman P. On the Representation and Estimation of Spatial Uncertainty. International Journal of Robotics Research. 1986; 5:56-68
173. Castellanos JA, Montiel JMM, Neira J, Tardos JD. The SPmap: a probabilistic framework for simultaneous localization and map building. IEEE Transactions on Robotics and Automation. 1999; 15(5):948-952.
174. Dissanayake MWMG, Newman P, Clark S, Durrant-Whyte HF, Csorba M. A solution to the simultaneous localization and map building (SLAM) problem. IEEE Transactions on Robotics and Automation. 2001; 17(3):229-241.
175. Leonard J. J., Feder H. J. S. A computationally efficient method for large-scale concurrent mapping and localization. International Symposium on Robotics Research; 1999 p. 169-176.
176. Se S., Barfoot T. D., Jasiobedzki P. Visual Motion Estimation and Terrain Modelling for Planetary Rovers. International Symposium on Artificial Intelligence for Robotics and Automation in Space 2005.
177. Sim R., Elinas P., Griffin M., Little J. J. Vision-based SLAM using the Rao-Blackwellised Particle Filter. IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR); 2005.
178. Davison AJ, Reid I, Molton N, Stasse O. MonoSLAM: Real-Time Single Camera SLAM. Pattern Analysis and Machine Intelligence. 2007; 29(6):1052-1067.
179. Burschka D, Hager GD. V-GPS(SLAM): vision-based inertial system for mobile robots. International Conference on Robotics and Automation. 2004; 1:409- 415
180. Burschka D., Hager G.D. V-GPS(SLAM): vision-based inertial system for mobile robots. International Conference on Robotics and Automation. 2004; 1:409- 415
181. Burschka D, Li M, Taylor R, Hager GD Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery. International Conference on Medical Image Computing and Computer Assisted Intervention; 2004; St. Malo, France. p. 413–421.
182. Brown CM, Ballard DH. Computer Vision. Englewood Cliffs, NJ: Prentice-Hall; 1982.
183. Thrun S, Burgard W, Fox D. Probabilistic Robotics. MIT Press; 2005.
184. Thrun S. Robotic mapping: A survey. Morgan Kaufmann. 2002.
185. Abdel-Hakim A.E., Farag A.A. CSIFT: A SIFT Descriptor with Color Invariant Characteristics. CVPR; 2006 p. 1978- 1983.
186. In: http://www.robots.ox.ac.uk/~vgg/research/affine.
187. In: http://people.cs.ubc.ca/~lowe/keypoints/.
188. Van-Gool L, Moons T, Ungureanu D Affine/Photometric Invariants for Planar Intensity Patterns. European Conference on Computer Vision; 1996 p. 642-651.
189. Florack L., ter Haar Romeny B., Koenderink J., Viergever M. General Intensity Transformations and Second Order Invariants. Scandinavian Conference on Image Analysis; 1991 p. 338-345.
190. Freeman W., Adelson E. The Design and Use of Steerable Filters. Pattern Analysis and Machine Intelligence. 1991; 13(9):891-906.
191. Lazebnik S., Schmid C., Ponce J. Sparse Texture Representation Using Affine-Invariant Neighborhoods Computer Vision and Pattern Recognition; 2003.
192. Ling H, Jacobs DW Deformation invariant image matching. International Conference on Computer Vision p. 1466- 1473.
193. Geusebroek J. M., van den Boomgaard R., Smeulders A. W. M., H.. G. Color invariance. Pattern Analysis and Machine Intelligence. 2001; 12(23):1338–1350.
194. K. Mikolajczyk CS. A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence. 2005; 10(27):1615-1630.
195. Funt BV, Finlayson GD. Color constant color indexing. Pattern Analysis and Machine Intelligence. 1995; 17(5):522 - 529.
196. van de Weijer J, Schmid C Blur Robust and Color Constant Image Description. International Conference on Image Processing; 2006 p. 993 - 996.
197. Gevers T, Smeulders AWM. Color Based Object Recognition. Pattern Recognition. 1999; 32:453 - 464.
198. Johnson A., Hebert M. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. Pattern Analysis and Machine Intelligence. 1999; 5(21):433-449.

199. Swain MJ, Ballard DH. Color Indexing. International Journal of Computer Vision. 1991; 7(1):11-32.
200. Yang GZ, Hu XP. Multi-Sensor Fusion. In: Body Sensor Networks. London; 2006. p. 239 - 286.
201. Koller D, Sahami M Towards optimal feature selection. In: Proc. ICML; 1996 p. 284 - 292.
202. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97:273 - 324.
203. Hu XP. Feature selection and extraction of visual search strategies with eye tracking [Thesis thesis]; 2005.
204. Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(10):1615 - 1630.
205. Van-Trees HL. Detection Estimation and Modulation Theory. New York: Wiley and Sons; 1971.
206. Thiemjarus S. A Framework for Contextual Data Fusion in Body Sensor Networks thesis]; 2008.
207. Amit Y, Geman, D. Shape Quantization and Recognition with Randomized Trees. Neural Computation; 1997 p. 1545–1588.
208. Rosten E, Drummond T Leonardis A, Bischof, H., Pinz, A, editor. Machine learning for high-speed corner detection. European Conference on Computer Vision; 2006 p. 430–443.
209. Meltzer J, Yang M-H, Gupta R, Soatto S Multiple View Feature Descriptors from Image Sequences via Kernel Principal Component Analysis. European Conference on Computer Vision; 2004 p. 215-227.
210. Lepetit V, Pilet J, Fua P Point Matching as a Classification Problem for Fast and Robust Object Pose Estimation. Computer Vision and Pattern Recognition; 2004 p. 244-250
211. Lepetit V, Lagger P, Fua P Randomized trees for real-time keypoint recognition. Computer Vision and Pattern Recognition; 2005 p. 775–781.
212. Lepetit V, Fua P. Keypoint Recognition Using Randomized Trees. Pattern Analysis and Machine Intelligence. 2006; (9):1465 - 1479
213. Özuysal M, . Lepetit, V,. Fleuret, F,. Fua, P,. Feature Harvesting for Tracking-by-Detection. European Conference on Computer Vision; 2006 p. 592-605.
214. Lucas BD, Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. IJCAI; 1981 p. 674–679.
215. Quinlan JR. Induction of decision trees. Machine Learning 1. 1986.
216. Collins R, Liu, Y., Leordeanu, M. On-Line Selection of Discriminative Tracking Features. IEEE Trans Pattern Analysis and Machine Intelligence. 2005; 10(27):1631–1643.
217. Yuen S, G,. Novotny, P,M,. Howe, R,D,. Quasiperiodic predictive filtering for robot-assisted beating heart surgery. International Conference on Robotics and Automation; 2008 p. 3875-3880.
218. Crothers I, . Gallagher, A,. McClure, N,. James, D,T,D,. McGuigan, J,. Experienced laparoscopic surgeons are automated to the "fulcrum effect": an ergonomic demonstration. Endoscopy. 1999; 318:365-369.
219. Masson N, . Nageotte,F,. Zanne, P,. de Mathelin, M,. In Vvio Comparison of Real-time Tracking Algorithms for Interventional Felexible Endoscopy. ISBI; 2009.
220. Forss PE Maximally Stable Colour Regions for Recognition and Matching. Computer Vision and Pattern Recognition; 2007 p. 1-8.
221. Ayache N, Faugeras OD. Building, Registrating, and Fusing Noisy Visual Maps. International Journal of Robotics Research. 1988; 7(6):45–65.
222. Sim R., Elinas P., Griffin M., Shyr A., Little J. J. Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters. IIntelligent Robots and Systems 2006 p. 2082-2089.
223. Il-Kyun J, Lacroix S High resolution terrain mapping using low attitude aerial stereo imagery. Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on; 2003 p. 946-951 vol.2.
224. Pupilli M. L., Calway A. D. Real-Time Camera Tracking Using a Particle Filter. British Machine Vision Conference; 2005 p. 519-528.

225. Kwok NM, Dissanayake G Bearing-only SLAM in indoor environments using a modified particle filter. Australasian Conference on Robotics and Automation 2003 p. 1-3.
226. Eade E., Drummond T. Scalable Monocular SLAM. Computer Vision and Pattern Recognition; 2006 p. 469- 476.
227. Klein G, Murray D Parallel Tracking and Mapping for Small AR Workspaces. International Symposium on Mixed and Augmented Reality; 2007 p. 225-234.
228. Bosse M, Newman P, Leonard J, Teller S. Simultaneous localization and map building in large-scale cyclic environments using the Atlas framework. The International Journal of Robotics Research 2004; 23(12):1113–1139
229. Montemerlo M, Thrun S, Koller D, Wegbreit B FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. 2002; Edmonton, Canada.
230. Montemerlo M, Thrun S Simultaneous localization and mapping with unknown data association using FastSLAM. International Conference on Robotics and Automation; 2003 p. 1985-1991 vol.2.
231. Elfes A. Using occupancy grids for mobile robot perception and navigation. Computer. 1989; 22(6):46-57.
232. Moravec H., Elfes A. High resolution maps from wide angle sonar. International Conference on Robotics and Automation; 1985 p. 116-121.
233. Paz LM, Piníes P, Tardós JD, Neira J. Large Scale 6DOF SLAM with Stereo-in-Hand. Transactions on Robotics. 2008; 24(5):946-957.
234. Saez J. M., Escolano F. Entropy Minimization SLAM Using Stereo Vision. International Conference on Robotics and Automation; 2005 p. 36-43.
235. Tomono M Robust 3D SLAM with a stereo Camera Based on an Edge-Point ICP Algorithm. International Conference on Robotics and Automation; 2009 p. 4306–4311.
236. Sim R., Griffin M., Shyr A., Little J. J. Scalable real-time vision-based SLAM for planetary rovers. IROS Workshop on Robot Vision for Space Applications; 2005.
237. Barfoot TD Online visual motion estimation using FastSLAM with SIFT features. International Conference on Intelligent Robots and Systems 2005 p. 579-585.
238. Miro J. V., Dissanayake G., Weizhen Z. Vision-based SLAM using natural features in indoor environments. International Conference on Intelligent Sensors, Sensor Networks and Information; 2005 p. 151-156.
239. Davison AJ Active Search for Real-Time Vision. International Conference on Computer Vision; 2005 p. 66-73.
240. WooYeon J., Mu L.K. CV-SLAM: a new ceiling vision-based SLAM technique. International Conference on Intelligent Robots and Systems; 2005 p. 3195-3200.
241. Williams B, Klein G, Reid I. Real-time SLAM Relocalisation thesis]; 2007.
242. Sim R., Elinas P., Griffin M., Shyr A., Little J. J. Design and analysis of a framework for real-time vision-based SLAM using Rao-Blackwellised particle filters. Canadian Conference on Computer and Robotic Vision 2006 p. 21.
243. Karlsson N., di Bernardo E., Ostrowski J., Goncalves L., Pirjanian P., Munich M. E. The vSLAM Algorithm for Robust Localization and Mapping. International Conference on Robotics and Automation; 2005 p. 24-29.
244. Eustice R., Singh H., Leonard J., Walter M., Ballard R. Visually Navigating the RMS Titanic with SLAM Information Filters. Robotics: Science and Systems 2005 p. 57-64.
245. Williams S., Mahon I. Simultaneous localisation and mapping on the Great Barrier Reef. International Conference on Robotics and Automation; 2004 p. 1771-1776.
246. Yokokohji Y, Kurisu M, Takao S, Kudo Y, Hayashi K, Yoshikawa T Constructing a 3-D map of rubble by teleoperated mobile robots with a motion canceling camera system. Intelligent Robots and Systems; 2003 p. 3118-3125.
247. Burgard W., Cremers A. B., Fox D., Hahnel D., Lakemeyery G., Schulz D., et al. Experiences with an Interactive Museum Tour-Guide Robot. Artificial Intelligence. 1999; 114(1-2):3-55.
248. Davison A.J., Mayol-Cuevas w., Murray D.W. Real-Time Localisation and Mapping with Wearable Active Vision. International Symposium on Mixed and Augmented Reality. 2003.
249. http://www.doc.ic.ac.uk/~ajd/software.html.

250. Kalman RE. A new approach to linear filtering and prediction problems. Journal of Basic Engeneering. 1960.
251. Greg Welch GB. An Introduction to the Kalman Filter. In.
252. Swaminathan R, Nayar SK. Nonmetric Calibration of Wide-Angle Lenses and Polycameras. Pattern Analysis and Machine Intelligence. 2000; 22(10):1172-1178.
253. Dickens MM, Bornhop DJ, Mitra S Removal of Optical Fiber Interference in Color Micro-Endoscopic Images. 11th IEEE Symposium on Computer Based Medical Systems; 1998 p. 246.
254. Winter C, Rupp S, Elter M, Munzenmayer C, Gerhauser H, Wittenberg T. Automatic adaptive enhancemnt for images obtained with fiberscopic endoscopes. IEEE Transactions on Biomedical Engineering. 2006; 53(10):2035-2046.
255. Noonan D, Mountney P, Elson D, Darzi A, Yang G-Z A Stereoscopic Fibroscope for Camera Motion and 3D Depth Recovery During Minimally Invasive Surgery. International Conference on Robotics and Automation; 2009 p. 4463-4468.
256. Meining A, Bajbouj, M., von Delius, S., Prinz, C., . Confocal Laser Scanning Microscopy for in vivo Histopathology of the Gastrointestinal Tract. Arab Journal of Gastroenterology. 2007; 8:1–4.
257. Thiberville L, Moreno-Swirc, S., Vercauteren, T., Peltier, E., Cavé, C., Bourg Heckly, G.,. In Vivo Imaging of the Bronchial Wall Microstructure Using Fibered Confocal Fluorescence Microscopy. American Journal of Respiratory and Critical Care Medicine 2007; 175:22-31.
258. Van Dam J, . . Novel methods of enhanced endoscopic imaging. GUT. 2003 52 (4).
259. Allain B, Hu MX, Lovat LB, Ourselin S, Cook R, Hawkes DJ Biopsy Site Re-localisation based on the Computation of Epipolar Lines from Two Previous Endoscopic Images. Medical Image Computing and Computer Assisted Intervention; 2009 p. 491-498.
260. Wengert C., Cattin P.C., Duff J.M., Székely G. Markerless Endoscopic Registration and Referencing. Medical Image Computing and Computer-Assisted Intervention 2006 p. 816--823.
261. Giannarou M, Elson DS, Yang GZ Tracking of spectroscopic and microscopic optical probes in endoscopy using the endoscope image field. The Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy (OPTIMHisE) Workshop at Medical Image Computing and Computer Aided Intervention; 2009.
262. Brown M, Lowe D Recognizing Panoramas. International Conference on Computer Vision; 2003 p. 1218–1225.
263. Delaunay B. Sur la sphère vide. Otdelenie Matematicheskikh i Estestvennykh Nauk. 1934; 7:793–800.
264. Perez P, Gangnet M, Blake A. Poisson image editing. ACM Trans. Graph 2003; 3(22):313–318.
265. Montemerlo M, Thrun S. FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics Springer Berlin / Heidelberg; 2007.
266. Wang C-C. Simultaneous Localization, Mapping and Moving Object Tracking thesis]: Carnegie Mellon University; 2004.
267. Atasoy S, Glocker B, Giannarou S, Mateus D, Meining A, Yang GZ, et al. Probabilistic Region Matching in Narrow-Band Endoscopy for Targeted Optical Biopsy. Medical Image Computing and Computer Assisted Intervention; 2009 p. 499-506.
268. Lee S-L, Huntbatch A, Pratt P, Lerotic M, Yang G-Z. In Vivo, In Situ Image Guidance and Modelling in Robotic Assisted Surgery. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2010.
269. Mylonas GP, Stoyanov D, Darzi A, Yang G-Z Assessment of Perceptual Quality for Gaze-Contingent Motion Stabilization in Robotic Assisted Minimally Invasive Surgery. Medical Image Computing and Computer Assisted Intervention; 2007 p. 660-667.
270. Okamura AM. Methods for haptic feedback in teleoperated robot-assisted surgery. Industrial Robot: An International Journal. 2004; 31(5):499–508.
271. Noonan DP, Mylonas GP, Darzi A, Yang G-Z Gaze Contingent Articulated Robot Control for Robot Assisted Minimally Invasive Surgery. International Conference on Intelligent Robots and Systems; 2008 p. 1186-1191.

272. Mylonas GP, Stoyanov D, Deligianni F, Darzi A, Yang G-Z Duncan JS, Gerig G, editors. Gaze-Contingent Soft Tissue Deformation Tracking for Minimally Invasive Robotic Surgery. Medical Image Computing and Computer Assisted Intervention; 2005 p. 843-850.