3D non-rigid SLAM in minimally invasive surgery

by

Jingwei Song

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

at the
Centre for Autonomous Systems
Faculty of Engineering and Information Technology
University of Technology Sydney

March 2020

Declaration of Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by an Australian Government Research Training Program Scholarship.

Production Note:

Signed: Signature removed prior to publication.

Date: 22 March 2020

UNIVERSITY OF TECHNOLOGY SYDNEY

Abstract

Faculty of Engineering and Information Technology
Centre for Autonomous Systems

Doctor of Philosophy

by Jingwei Song

Aiming at reducing trauma and morbidity associated with large incisions in open surgery, minimally invasive surgery (MIS) has been widely acquired in clinical practice as a powerful tool enabling patients with less pain, shorter hospital stay, and fewer complications. However, MIS narrows the surgeon's field of view which confines visual information when implementing MIS. Therefore, a stereoscope or monocular scope is an essential tool for capturing and transmitting 2D images during the procedure.

Although numbers of special sensors including laser, structured light, time-of-flight cameras have been applied or investigated in MIS, RGB scope is still widely applied in the intro-operative system because it is non-invasive and cheap to be installed. Thus it is an important topic to rebuild and visualize the latest deformed shape of soft-tissue surfaces to mitigate tissue damages from stereo or monocular scopes. This research aims at proposing innovative robocentric simultaneous localization and mapping (SLAM) algorithm for deformable dense reconstruction of soft-tissue surfaces using a sequence of images obtained from a stereoscope or monocular camera. In this paper, we try to solve the problem by introducing a warping field based on the embedded deformation (ED) nodes which makes full use of the 3D shapes recovered from consecutive pairs of stereo images by deforming the last updated model to the current live model. Our robocentric SLAM system (off-line and tested on stereo videos) can: (1) Incrementally build a live model by progressively fusing new observations with vivid accurate texture. (2) Estimate the deformed shape of the unobserved region with the principle As-Rigid-As-Possible. (3) Perform the dynamic model shape deformation. (4) Estimate the current relative pose between the soft-tissue and the scope.

We further improve and optimize the proposed robocentric deformable SLAM algorithm to MIS-SLAM: a complete real-time large scale robocentric dense deformable SLAM system with stereoscope in MIS based on heterogeneous computing by making full use of CPU and GPU. Idled CPU is used to perform ORB-SLAM for providing robust global pose. Strategies are taken to integrate modules from CPU and GPU. We solve the key problem raised in previous work, that is, fast movement of scope and blurry images make the scope tracking fail. Benefiting from improved localization, MIS-SLAM can achieve large scale scope localizing and dense mapping in real-time. It transforms and deforms the current model and incrementally fuses new observation while keeping the vivid texture. In-vivo experiments conducted on publicly available datasets presented in the form of videos demonstrate the feasibility and practicality of MIS-SLAM for potential clinical purpose.

In MIS-SLAM, however, it remains challenging to keep constant speed in deformation nodes parameter estimation when the model grows larger. In practice, the processing time grows rapidly in accordance with the expansion of the maps. Therefore, we propose an approach to decouple nodes of deformation graph in large scale robocentric dense deformable SLAM and keep the estimation time to be constant. We discover that only partial deformable nodes in the graph are connected to visible points. Based on this principle, the sparsity of the original Hessian matrix is utilized to split parameter estimation into two independent steps. With this new formulation, we achieve faster parameter estimation with amortized computation complexity reduced from $O(n^2)$ to closing $O(1)$. As a result, the computation cost barely increases as the map keeps growing. By our strategy, the bottleneck of limited computation in estimating deformation field in large scale environment has been overcome. The effectiveness is validated by experiments, featuring large scale deformation scenarios.

In addition to robocentric SLAM, this thesis also aims at developing a general SLAM which estimates the scope poses correctly. An elaborate observability analysis is conducted on the ED graph. We demonstrate and prove that the ED graph widely used in such scenarios is unobservable and leads to multiple solutions unless suitable priors are provided. Example, as well as theoretical prove, are provided to show the ambiguity of ED graph and scope pose. Different from robocentric SLAM, in modeling non-rigid scenario with ED graph, motion priors of the deforming environment is essential to separate robot pose and deforming environment. The conclusion can be extrapolated to any free form deformation formulation. In guaranteeing the observability, this research proposes a preliminary deformable SLAM approach to estimate robot pose in complex environments that exhibits regular motion. A strategy that approximates deformed shape using a linear combination of several previous shapes is proposed to avoid the ambiguity in robot movement and rigid and non-rigid motions of the environment. Fisher information matrix rank analysis is performed to prove the effectiveness. Moreover, the proposed algorithm is validated using Monte Carlo simulations and real experiments. It is demonstrated that the new algorithm significantly outperforms conventional SLAM and ED based SLAM especially in scenarios where there is large deformation.

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor A/Prof. Shoudong Huang and Dr. Liang Zhao for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I hope I can continue to collaborate with you in the future.

Besides my supervisor, I would like to thank Dr. Teresa Vidal Calleja, A/Prof. Jaime Valls Miro, Prof. Gamini Dissanayake, and A/Prof. Sarath Kodagoda for their guidance and friendship, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also go to Dr. Antonio Agudo and Prof. Francesc Moreno-Noguer, who provided me an opportunity to join their team as an exchange student at Institut de Robótica i Informática Industrial (IRI) of the Spanish Council for Scientific Research and the Technical University of Catalonia. And also for supporting me to win IRI-Maria de Maeztu intern scholarship. During my 4 months visit, I have terrific experience and memory in IRI as well as Barcelona. I am also grateful to Dr. Mitesh Patel for hosting me at FXPAL as an intern; the friendly environment and living in California touches me very much.

I acknowledge the scholarships International Research Training Program Scholarship (IRTP) funded by the Australian Government, Department of Education and Training to support 3.5 years for my research and make my study possible. I would also like to acknowledge Centre for Autonomous Systems and School of Mechanical and Mechatronic Engineering of the Faculty of Engineering and Information Technology at University of Technology Sydney for providing me funding to attend conferences held in London, Barcelona, Madrid, Brisbane, Vancouver and being an exchange student in Barcelona.

Many thanks go to my fellow friends and colleagues at the Centre for Autonomous Systems for the numerous stimulating and helpful discussions and their friendship. My special thanks go to Dr. Jun Wang for inspiring discussions on solving technical issues as well as reporting the state-to-art technologies. We work closely in achieving good publications. I also appreciate Dr. Raphael Falque for useful discussions on theories of deformation technologies. Raphael is senior to me and I benefit a lot from him. I would also like to thank my colleague Kristian Villavicencio for helping with the ElasticFusion experiments.

I thank my friends and labmates Karthick Thiyagarajan, Mahdi Hassan, Lakshitha Dantanarayana, Phillip Quin, Leo Shi, Nalika Ulapane, Buddhi Wijerathna, Asok Aravinda, Julien Collart, Alexander Virgona, Maani Ghaffari Jadidi, Kasra Khosoussi, Daobilige Su, Brendan Emery, Katherine Waldron, Cédric Le Gentil, Kanzhi Wu, Teng Zhang and many other colleagues. I thank my fellow labmates for the stimulating discussions and for all the fun we have had in the last four years, as well as their hands-on help on setting up the experimental environment to collect and process data; Many thanks to Herni Winarta and Katherine Waldron for their help on administrative work.

Additional thanks go to my other friends in Sydney, I enjoy the 3 years in this sunny city. The last but most important thanks give to my parents, who are always proud of me for my achievements, and supported and cared for me throughout these years.

Contents

List of Figures

List of Tables

Acronyms & Abbreviations

CPU	Central Processing Unit
CT	Computed Tomography
CUDA	Compute Unified Device Architecture
DFF	Distance Field Function
ED	Embedded Deformation
EKF	Extended Kalman Filter
ELAS	Efficient Large-scale Stereo
EM sensor	Electromagnetic sensor
FEM	Finite Element Method
FIM	Fisher Information Matrix
GPGPU	General-Purpose computing on Graphics Processing Units
GPU	Graphical Processing Unit
ICP	Iterative Closest Point
MIS	Minimally Invasive Surgery
NRSfM	Non Rigid Structure from Motion
PI	Points Irrelevant

PR	Points Relevant
RANSAC	RANdom SAmple consensus
RMSE	Root-Mean-Square Error
SfM	Structure from Motion
SfS	Shape from Shading
SfT	Structure from Template
SIFT	Scale-invariant feature transform
SLAM	Simultaneous Localization And Mapping
SURF	Speeded Up Robust Feature
TSDF	Truncated Signed Distance Function
TSDW	Truncated Signed Distance Weight

Nomenclature

	General Notations
O	Centers of the left image.
O'	Centers of the right image.
p	Pixel of the projected point on left image.
p'	Pixel of the projected point on right image.
u	First coordinate of 2D pixel.
v	Second coordinate of 2D pixel.
$\mathbf{K}$	Camera intrinsic matrix.
$\alpha(u, v)$	Surface albedo.
$\hat{I}(u, v)$	The measured pixel intensity at pixel (u, v).
$O(\cdot)$	Computational complexity.
Ω	Linear elastic solid parameter.
λ and $\mathbf{G}$	Lam'e parameters that define the material elastic properties.
$\mathcal{U}$	Poisson's ratio.
$\mathcal{E}$	Young's modulus.
$a_{i,jj}$ and $a_{j,ij}$	Displacement vectors share the same edge.
$\mathbf{g}_j$	Position of node j.
$\mathbf{A}_j$	Affine matrix of node j.
$\mathbf{v}$	Position of a point.
$\tilde{\mathbf{v}}$	Target deformed vertex of $\mathbf{v}$.
$\mathbf{R}_c$	Global rotation of the scope.
$\mathbf{T}_c$	Global translation of the scope.

$w(\mathbf{v})$	The quantified weight for transforming $\mathbf{v}$ exerted by each related ED node.
d_{max}	The maximum distance of the vertex to $k+1$ nearest ED node.
m	The number of ED nodes.
$\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$	The column vectors of $\mathbf{A}$.
E_{rot}	The affine matrix close to $SO(3)$.
E_{reg}	The sum the transformation errors from each ED node.
α_{jk}	The overlap influence of the two ED nodes.
$\mathbb{N}(j)$	The set of neighboring node to node $\mathbf{g}_j$.
$\boldsymbol{\mathcal{F}}(\cdot)$	A general function defining a point to target distance.
$\mathcal{D}(\cdot)$	The corresponding voxel value recorded in DFF.
E_{data}	The sum distance error.
$\mathbb{L}$	The set for all the visible points.
$\mathbb{D}$	Depth scan.
$\Gamma(\cdot)$	Lift 2D depth pixel up to 3D point.
$\mathrm{H}(\cdot)$	Lift 2D normal pixel up to 3D normal.
$P(\mathbf{v})$	The projective function projecting 3D vertex to 2D pixel.
ϵ_d	Threshold for the distance.
ϵ_n	Threshold for the angle.
$\tilde{\mathbf{V}}_i$	The 3D points of current frame from last frame of ORB features.
$\mathbf{V}_i$	The 3D points of the deformed points from last frame of ORB features.
$\omega(\mathbf{v}_i)$	Weight of model point $\mathbf{v}_i$.
$\mathbf{C}_i$	Color of model point.
$d_{min}(\tilde{\mathbf{v}}_i)$	The minimum distance of model point.
$\tilde{\mathbf{v}}_i$	to its corresponding nodes.
ϵ	The average grid size of nodes.
$\tilde{\mathbf{v}}_i\|_z$	The value of point $\tilde{\mathbf{v}}_i$ on the z direction.
t_i	The time stamp of vertex $\mathbf{v}_i$.
$\mathbf{n}_i$	The normal of vertex $\mathbf{v}_i$.
$\mathbf{s}_i$	The boolean variable stability of vertex $\mathbf{v}_i$.

E_{corr}	Sum errors of distances between deformed key points and target key points.
E_r	The $SO(3)$ distance between estimated orientation and intilial orientation.
E_p	The Euclidean distance between estimated position and intilial position.
ϵ_{dv}	The threshold for extracting visible points based on distance.
ϵ_{nv}	The threshold for extracting visible points based on angle.
$\mathbf{R}_c^n$	Estimated rotation in step n.
$\mathbf{T}_c^n$	Estimated translation in step n.
$\tilde{\mathbf{R}}_c^n$	Rotation in step n from ORB-SLAM.
$\tilde{\mathbf{T}}_c^n$	Translation in step n from ORB-SLAM.
$\mathbb{V}^n$	Set of visible points in step n.
$\mathbb{D}^n$	Observed depth in step n.
$\mathbb{P}^n$	Fused point set.
ϵ_{df}	The threshold for fusing points based on distance.
ϵ_{nf}	The threshold for fusing visible points based on angle.
$\tilde{\mathbf{v}}_i^n\|_z$	The value of deformed point $\tilde{\mathbf{v}}_i^n$ in the z direction.
$\tilde{\mathbf{n}}_i^n$	The deformed normal of $\mathbf{n}_i^n$.
ω_{max}	The maximum weight for each model point.
$\mathbf{P}$	A group of predefined key source points.
$\tilde{\mathbf{P}}$	Deformed key points set of $\mathbf{P}$.
$\otimes$	The Kronecker product.
$\|\| \cdot \|\|_F^2$	The Frobenius norm.
X_i	The state vector is denoted as .
$skew(\cdot)$	The skew symmetric operator.
$\mathbf{I}$	A 3 by 3 identity matrix.
$\odot$	The Hadamard product.
$\boldsymbol{\mathcal{H}}_{ed}$	Hessian matrix of ED formulation.
$\mathbf{H}_1$	One submatrix of $\boldsymbol{\mathcal{H}}_{ed}$.
$\mathbf{H}_2$	One submatrix of $\boldsymbol{\mathcal{H}}_{ed}$.

$\mathbf{f}^n$	A feature position in step n.
N	The number of features.
F	The number of steps.
$\mathbf{B}$	The combination of all valid features.
E_{obs}	The sum error of robot to feature observations.
$\mathbf{m}_i^j$	The observation from robot to location of feature i in step j.
$\mathcal{F}(\cdot)$	The estimated observation from robot pose to feature position.
E_f	The error between current feature and its estimation from historical locations.
E_{ini}	The initial robot pose keeps static in the period size t.
$\ominus$	Inverse retraction defining $SO(3)$ distance.
$\mathbf{c}$	The coefficient matrix .

Chapter 1

Introduction

Minimally Invasive Surgery (MIS), which is an indispensable tool for modern surgery, greatly benefits the patients with reduced incisions, trauma and less hospitalization time [2]. A typical MIS setup is made up of one scope and one surgical instrument manipulated by the surgeons or robots. Normally, the surgeons work through a set of holes approximately 1 cm in diameter. Long handled instruments cut and grip tissue within the body with a video camera providing a view of the internal operating field. In the field of computer assisted surgery, advanced surgical robots even develop “hand in hand ”with MIS.

Although being an indispensable tool for modern surgery for the ability to mitigate postoperative infections, MIS also narrows the surgeons’ field of view and limit their perception. Thus, MIS introduces significant challenges to surgeons as they are required to perform the procedures in a narrow space with elongated tools without direct 3D vision [3]. Therefore, it is helpful if a dynamic 3D morphology could be generated and rendered for the surgeons intra-operatively. However, the small field of view of the scopes and the deformation of the soft-tissue limit the feasibility of using traditional structure-from-motion and image mosaic methods. Even worse, the rigid and non-rigid movement caused by the motion of camera pose, breathing, heartbeat and instrument interaction increases difficulty in soft-tissue reconstruction and visualization. Therefore, this thesis focuses on incrementally recovering the morphology and motion of soft-tissues with a stereoscope and monocular scope intra-operatively.

The first MIS case is believed to be conducted by Dr. J. Barry McKernan in 1988. He made only a 10mm incision and inserted a laparoscope (or miniature camera) into a patient's abdomen and removed a gall bladder. Thanks to small invasions, the patient recovered in days rather than weeks or months. This is the first laparoscopic cholecystectomy performed in the U.S. and the beginning of the minimally invasive movement in surgery. From that time onward, MIS is widely applied in modern surgeries and plays an important role in substituting traditional open surgeries.

Recently, MIS is combined with state-of-the-art technologies like robotics and automation intending to minimize trauma and incisions in the process. For example, Fig. 1.1 is a comparison of open Transforaminal lumbar interbody fusion (TLIF) and MIS-TLIF [4]. TLIF is a surgical procedure that removes a painful lumbar disc and replaces it with either a block of bone or a fusion device to allow the bone to grow across the disc space creating a fusion. The left figure shows the "open" surgery which is done by making a large incision in the middle of the back and operate the process. Or it can be implemented in the right figure with smaller incisions (several inches) made on each side against the back and insert two tubes. A miniature camera (usually a laparoscope or endoscope) is placed through one of the trocars so the surgical team can view the procedure as a magnified image on video monitors in the operating room. After that, the specialized instruments are placed through another tube to perform the procedure.

One important issue in the MIS process is knowing the position and direction of the miniature camera instrument while map the inner environment. And this can be formulated as a visual simultaneous localization and mapping (SLAM) problem.

Many research activities have been devoted to deal with 3D soft-tissues shape reconstruction, camera navigation or both. A structure from motion (SfM) pipeline [5] is proposed to partial 3D surgical scene reconstruction and localization. And in the work of Stoyanov [6], stereo images were used to extract sparse 3D point cloud. Haouchine et al. [7] and Malti et al. [8] extract whole tissue surface of organs from stereo or monocular images. Contrary to feature extraction based methods, Du et al. [9] employ an optical flow based

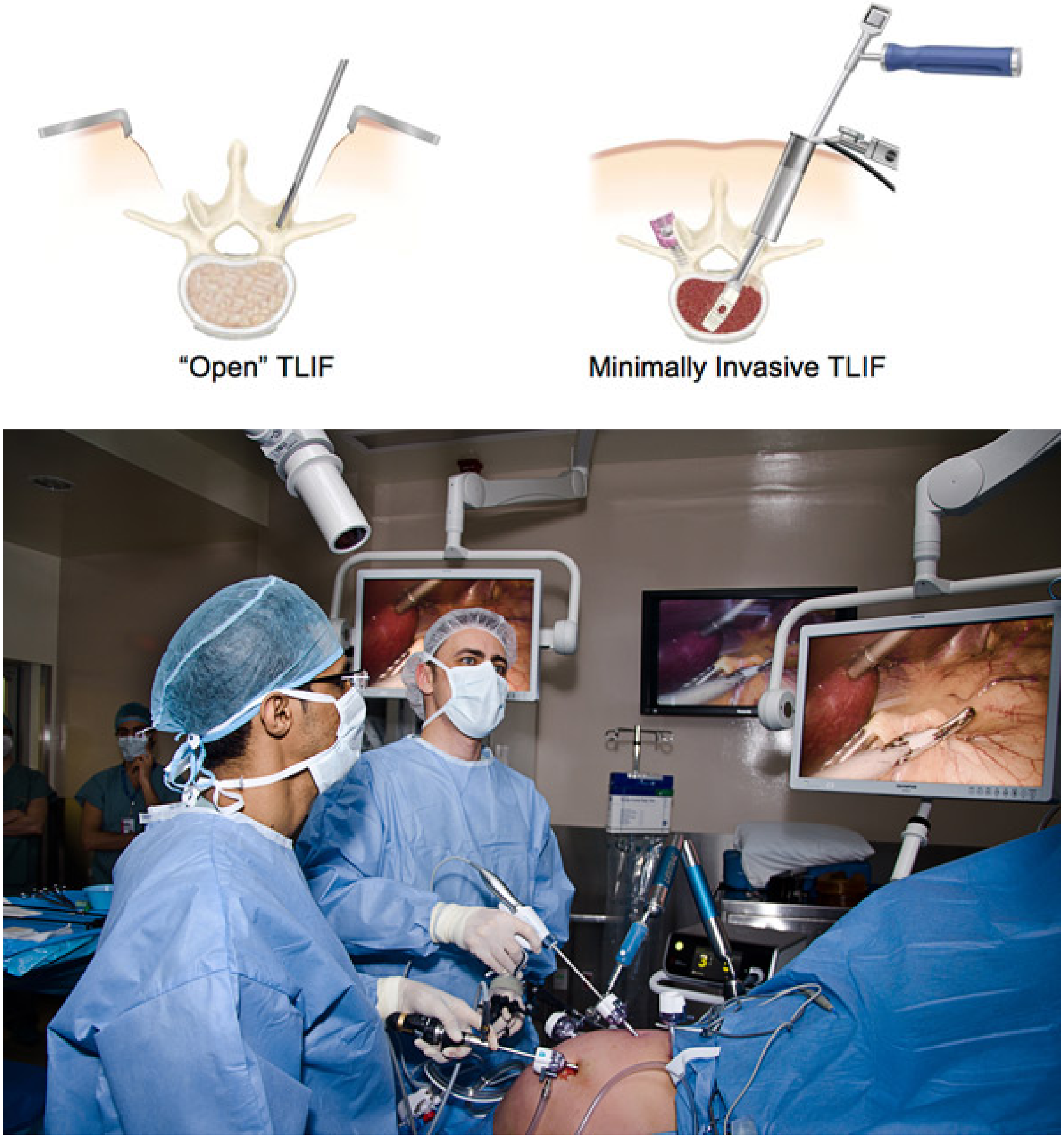


Figure 1.1: The difference between open surgery and MIS.

approach namely deformable Lucas-Kanade for tracking tissue surface. All the methods described above contribute greatly to enable implementing augmented reality or virtual reality in computer assisted interventions which greatly promote the accuracy and efficiency of MIS. Yet, these works mainly focus on tracking key feature points for localization and no work has been devoted to geometry based registration and dynamic soft-tissue surface reconstruction and dynamic deformation visualization.

1.1 Motivation

This research is inspired by the current development of RGB-D SLAM in the computer vision community. With the development of consumer-based RGB-D cameras like Microsoft Kinect and Intel Realsense, volume based template free reconstruction method has been proposed in reconstructing deformable objects and mainly part or whole human body. All related works follow the basic ideas presented in KinectFusion [10] which makes use of truncated signed distance function (TSDF) for fusing and smoothing rigid objects in real-time. Inspired by this idea, research efforts are devoted to transferring the TSDF fusion approach into modeling non-rigid objects. By dynamically warping TSDF volumes, Zollhöfer et al. [11], Newcombe et al. [12] and Innmann et al. [13] achieved real-time non-rigid model deformation and incremental reconstruction. These template-free techniques can process slow motion without occlusions because the sensor used is a single depth camera. Meanwhile, Dou et al. [14] proposed a multi-view RGB-D camera set as a substitution so that a real-time colored, fast moving and close-loop model can be built. This work is mainly integrated into the Holoportation system whose robustness in handling fast movement benefits from multi-view cameras with fixed positions. Although promising results can be achieved, these depth sensors are seldom applied in MIS which makes RGB-D based approaches impossible to be directly applied to the surgical scenario. And all the works mentioned above are a fixed volume based data management approach that requires spatial limits of the scenario. Thus, none of these methods can be directly applied to the computer-assisted interventions in MIS.

Since point cloud can be acquired from the disparity of stereo images [15], the goal of this thesis is to propose a new framework for implementing SLAM using a stereoscope.

There are two major requirements in the clinical application which limits applying DynamicFusion like pipeline into surgical vision. First, due to the spatial and computational limitations, the DynamicFusion pipeline requires a predefined volume and only allows the target object to move within this boundary. While in the MIS scenario, due to the limited field of view of the scope, the surgeons always require the scope moves freely in the space to observe more areas of the tissue during interventions. Volume deforming approach used in [12], [13] and [14] makes computation and unnecessary data storage increases exponentially as the volume size increases and there is a trade-off between model details depended on the grid size and computational cost in volume based data management. Second, different from obvious topologies in dynamic human body modeling, the smoothness of organs makes the algorithm easily converges to a local minimum. Considering the small field of view of the scope, the drifts of reconstruction caused by mismatching can seldom be corrected. This is different from the scenario of a large field of view since they can frequently re-observe the target as loop closure [12] or even reset the model [14] if multiple cameras are provided (8 sets of depth cameras were used in [14]). Even a slight drift leads to misalignment in textures especially on vascular.

1.2 Research aims

After analyzing the differences between our scenario (MIS) and similar works in the computer vision community, this thesis aims at proposing an innovative robocentric SLAM framework to recover the deformed 3D structure of the soft tissues.

We aim at solving major challenges in applying SLAM in the MIS scenario. The first and the most important challenge is the fast movement of the scope. The state-of-art methods do not address the tracking when the camera is moving fast. Similar to the traditional SLAM approaches [16] [17], serious consequences of fast motion are the blurry images and relevant disorder of depths. These phenomenon happen especially when current constructed model deforms to match the depth with false edges suffering from image blurring. That's why the proposed pipeline visualizes periodic deformation like respiration and heartbeat clearly on central regions but shows obvious drifts on the edges. Fast motion is a challenging issue as the only data source is the blurry images. Another issue is the accuracy of texture. A laparoscope with a narrow field of view results in obvious drifts and gaps on texture, especially in blurry images. In this thesis, we aim at integrating some image enhancement techniques to increase the robustness and accuracy of our pipeline.

In addition to robustness, another key issue in the embedded deformation (ED) graph, a widely applied deformation modelling method, is that when a new observation is incorporated, the number of nodes increases dramatically, posing a heavy computational burden. The estimation state space keeps expanding quadratically upon the increase of ED nodes. This thesis tries to propose a strategy to decrease the scale of the problem and convert computational complexity from $O(n^2)$ to $O(1)$.

Aside from building a technical framework enabling a real-time sequential robocentric SLAM system in MIS, this thesis also aims to theoretically testing the ambiguity in robot movement and the rigid and non-rigid motions of the environment. In the SLAM problem, pose estimation is crucial and we, therefore, focus on how to accurately estimate the robot pose. Particularly, the question is 'Is global pose of robot observable in a deformable environment unique?'. If the answer is no, then 'How can we enable observability of pose in a deformable environment?'. In this thesis, we extensively discuss the observability in the popular ED based SLAM algorithm. A counterexample is provided when analyzing the ED graph based visual SLAM system in the deformable environment. We clearly demonstrate that the global pose of the robot can be embedded into environment deformation formulation which is not separable. To solve this, we introduce a priori that theoretically deformation is a mixture of base shapes. Typical deformations we try to process include heartbeat, breath, periodic body movement. The rest can also be approximated by several historical basis shapes with rigid movement. Based on this priori, we propose an innovative back-end SLAM system that can efficiently calculate accurate pose as well as the deforming environment. The proposed time series basis formulation explicitly enforces correct observability constraints to overcome rigid pose mixing with the non-rigid deformation field. The result is compared with conventional rigid SLAM and ED formulation.

1.3 The structure of the thesis

The major contributions of this thesis are as follows:

- We introduce a warping field based on the ED nodes with 3D shapes recovered from consecutive pairs of stereo images. The warping field is estimated by deforming the last updated model to the current live model. Our robocentric SLAM system can: (1) Incrementally build a live model by progressively fusing new observations with vivid accurate texture. (2) Estimate the deformed shape of the unobserved region with the principle of As-rigid-as-possible. (3) Show the consecutive shape of models. (4) Estimate the current relative pose between the soft-tissue and the scope.
- On the basis of the deformable system, We propose a minimally invasive surgery simultaneous localization and mapping (MIS-SLAM) system: a complete real-time large scale dense deformable robocentric SLAM system with stereoscope in MIS based on heterogeneous computing by making full use of central processing unit (CPU) and graphical processing unit (GPU). Idled CPU is used to perform ORB-SLAM for providing initial rigid target transformation. Strategies are taken to integrate modules from CPU and GPU. We solve the key problem raised in previous work, that is, fast movement of scope and blurry images make the reconstruction fail. Benefiting from the improved rigid transformation, MIS-SLAM can achieve large scale scope to soft-tissue localizing and dense mapping in real-time. It transforms and deforms the current model and incrementally fuses new observation while keeping the vivid texture.
- Aiming at solving rapid growing time consumption in deformation nodes parameter estimation of ED formulation in deformable geometry and graphical problems, we propose an approach to decouple node of deformation graph in large scale dense deformable SLAM and keep the estimation time to be constant. Our approach fully exploits the fact that only a limited number of deformable nodes are related to visible points. We theoretically prove that the computation complexity is reduced from

$O(n^2)$ to closing $O(1)$ meaning the computation cost barely increases as the environment gets larger. Based on our strategy, the bottleneck of limited computation in estimating the deformation field in a large scale environment has been solved.

- To address the unobservability issue hindering the deformable SLAM approach, we also theoretically analyze the problem of general worldcentric SLAM in the deformable environment, where robots localize themselves and track multiple deforming features using their onboard sensor measurements. The main contribution is a novel deformable SLAM approach to estimate robot pose in complex environments that exhibit periodic motion. This thesis demonstrates that the widely used ED based formulation is unobservable and leads to multiple solutions unless suitable priors are available. A strategy that approximates deformed shape using a linear combination of several previous shapes is proposed to avoid the ambiguity of rigid and non-rigid motions of the robot and the environment.

1.4 Publications

The work on the introducing ED graph to present deformation of template based soft-tissue was presented first in 2016 Australasian Conference on Robotics and Automation Song et al. [18]. Then it was developed into the real-time template free 3D reconstruction framework published in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems [19]. A complete 3D robust heterogeneous 3D reconstruction method, named MIS-SLAM, was proposed and published in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems [20]. The list of publications is as follows:

- Song, J., Wang, J., Zhao, L., Huang, S. and Dissanayake, G. (2016). 3D Shape Recovery of Deformable Soft-tissue with Computed Tomography and Depth Scan. In Australasian Conference on Robotics and Automation (ACRA2016).
- Song, J., Wang, J., Zhao, L., Huang, S. and Dissanayake, G. Deformable Soft-tissue Reconstruction using Stereo Scope for Minimally Invasive Surgery. Computer Assisted Radiology and Surgery (CARS2017).

- Song, J., Wang, J., Zhao, L., Huang, S. and Dissanayake, G. Robust Shape Recovery of Deformable Soft-tissue Based on Information from Stereo Scope for Minimally Invasive Surgery. Hamlyn Symposium on Medical Robotics 2017

- Zhang, T., Wu, K., Song, J., Huang, S. and Dissanayake, G. (2017). Convergence and consistency analysis for a 3-d invariant-ekf slam. IEEE Robotics and Automation Letters, 2(2), 733-740.

- Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G. (2017). Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robotics and Automation Letters, 3(1), 155-162.

- Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G. (2018). MIS-SLAM: Real-Time Large-Scale Dense Deformable SLAM System in Minimal Invasive Surgery Based on Heterogeneous Computing. IEEE Robotics and Automation Letters, 3(4), 4068-4075.

- Wang, J., Song, J., Zhao, L. and Huang, S. A submap joining based RGB-D SLAM algorithm using planes as features[C] Field and Service Robotics. Springer, Cham, 2018: 367-382.

- Wang, J., Song, J., Zhao, L., Huang, S. and Xiong, R. A submap joining algorithm for 3D reconstruction using an RGB-D camera based on point and plane features[J]. Robotics and Autonomous Systems, 2019, 118: 93-111.

- Song, J., Bai, F., Zhao, L., Huang, S. and Xiong, R. Efficient two step optimization for large embedded deformation graph based SLAM. (In preparation)

- Song, J., Zhao, L. and Huang, S. An observable time series based SLAM algorithm for deforming environment. (In preparation)

- Song, J., Zhao, L., Huang, S., Moreno-Noguer, F., and Agudo, A. Non-rigid structure from motion with isometric constraint. (In preparation)

Chapter 2

Related works

This section covers the state-of-art single frame shape recovery approaches, rigid and non-rigid SLAM in deformable system, template based SLAM system, pose graph system acceleration and non-rigid SLAM with prior.

2.1 Sensors and single frame shape recovery approaches

In the past, various sensors for recovering 3D surface structures of surgical spaces have been proposed including stereo miniature scope, monocular miniature scope, laser, structured light, and time-of-flight cameras. But the active sensors including laser, structured light, and time-of-flight cameras are seldom applied in clinical applications due to sensor size, effectiveness or popularity [5]. Till now, the most widely used sensors are still the passive monocular scopes and stereoscopes. Monocular scopes are widely applied in conventional surgery in the form of laparoscope or endoscope, which provide surgeons with 2D images and is in the lowest cost among all the potential sensors. With respect to single monocular scope 3D shape recovery, a well-studied 3D reconstruction method in computer vision named shape from shading (SFS) provides researchers a possible solution. SFS is based on the assumption of a single source light reflectance model to estimate the shape. In recent decades, accompanied with the development of computer assisted intervention, stereo cameras have also gained popularity in robot-assisted surgeries, among them is the da Vinci system [21][22]. For stereo visions, a simple but practical application is to feed left and right images to each eye separately, while surgeons send two separate images to the brain for creating an imaginary 3D shape. However, this imaginary 3D model is not enough for automatic scope pose estimation and computer assisted surgery. Numerous single frame 3D shape recovery algorithms have been proposed and applied in estimating shape from stereo images. This section reviews and analyses state-of-art stereo shape estimation and monocular SFS shape recovery.

2.1.1 Stereo shape recovery

Stereo shape recovery is a typical stereoscopic vision problem aiming at extracting 3D information from digital images. Fig. 2.1 is an illustration of the basic principle of how stereo vision works. $\mathbf{O}$ and $\mathbf{O}'$ are the centers of the left and right camera coordinates and $\mathbf{p}$ and $\mathbf{p}'$ are corresponding pixels of the projected 3D point on both images. In order to obtain the desired depth information, a noticeable disparity between the two images needs to be calculated. That is the different image coordinate $\mathbf{p}$ and $\mathbf{p}'$ of the projected 3D point resulted from a different observing angle. If we can obtain the relative disparity between points in a scene across the two different images, a depth map can be generated. Given a set of point correspondences between the left and right images, the depth map of the scene is determined.

Practically, the 3D stereo vision system requires neither additional hardware like active sensors nor slow and complicated depth generation approaches in monocular vision. This makes stereo vision popular in modern MIS and computer assisted intervention. Lau et al. [23] proposes the first template based soft-tissue tracking and recovery with the B-spline based method. Following general stereo vision development in the computer vision community, Stoyanov [6] proposes a method robust to specular reflections and surgical instrument occlusion. The basic idea of this approach is to extract a set of sparse salient feature points and then propagate the registration to the nearby features. Utilizing recent development in parallel computing called general-purpose computing on graphics processing units (GPGPU), Kowalczuk et al. [24] implements a traditional stereo vision algorithm on GPU and achieves real-time 3D shape recovery. Similarly, Totz et al. [25]

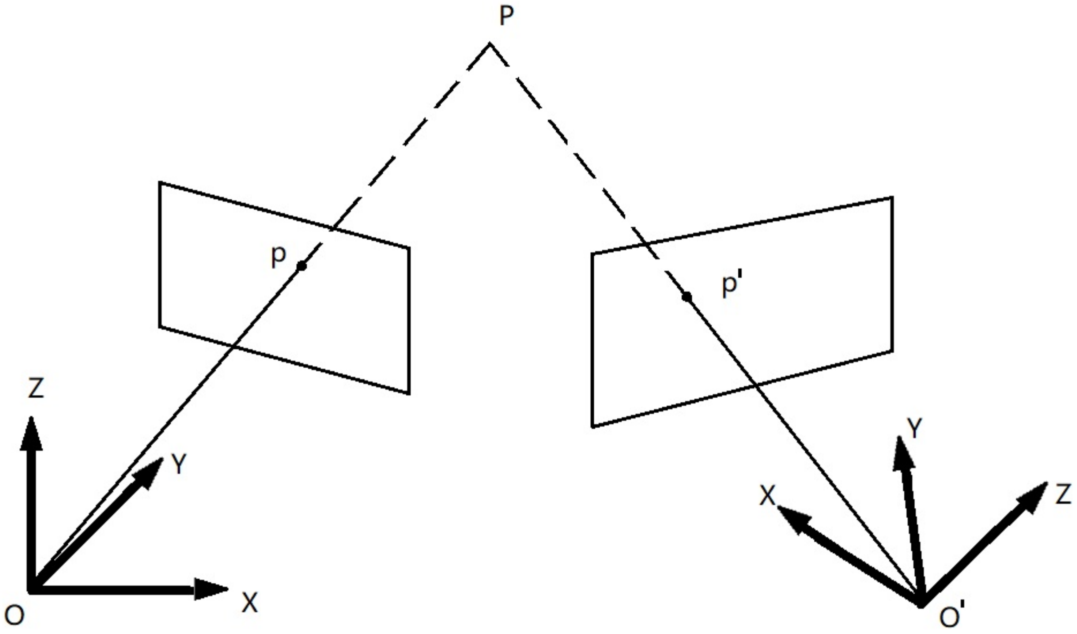


Figure 2.1: Depth estimation from stereo images.

also proposes a real-time GPU based semi-dense stereo reconstruction method for liver surface reconstruction, where a coarse-to-fine pyramidal strategy is adopted.

At the time of writing this thesis, the widely used open source stereo 3D shape recovery code is ELAS [26], which is similar to [6] by building a sparse triangular mesh and extending the triangulation to the rest of depth. Researches in MIS scenario [27][28] adopt this approach as the 3D depth generator.

2.1.2 Shape from shading

Even though stereo vision is the most practical way of 3D shape recovery, efforts are devoted to analyzing monocular scopes due to its broad applicability. Researchers try to enhance the monocular scope to a full imaging device to provide better quantitative and qualitative data. As one of the well-studied 3D reconstruction methods in computer vision, SFS is widely test in this domain.

SFS is a method aiming at reconstructing the 3D structure of a scene from the lighting of the object based on some assumption. It is based on the principal that the surface of

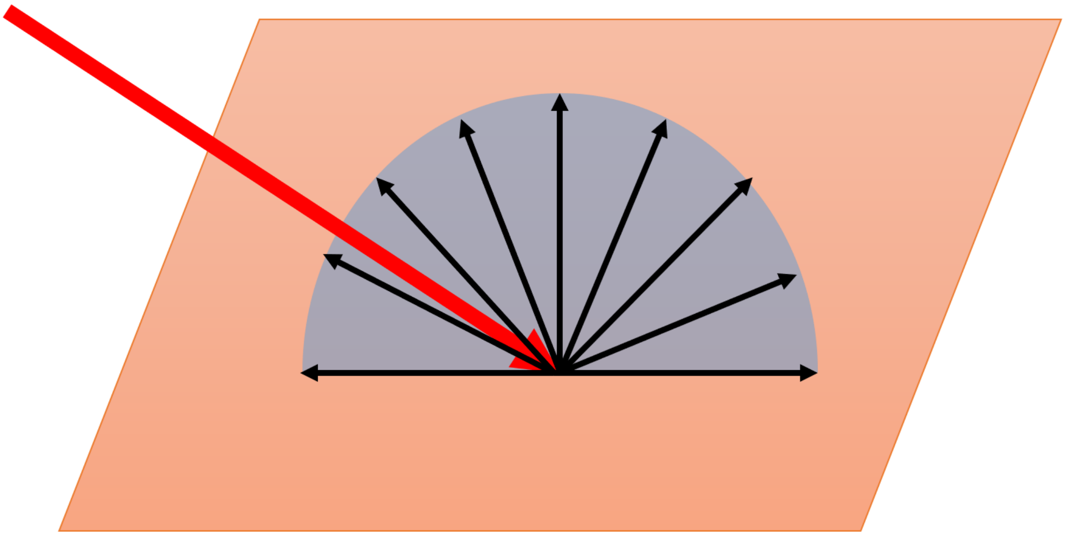


Figure 2.2: Illustration of Lambertian reflectance model.

the scene exhibits Lambertian reflectance (see Fig. 2.2) which means that the intensity reflected by a surface to the observer is the same regardless of the observing angle. SFS approach tries to find the optimal depth to minimize the following object function to obtain best depth function [29]:

$$d^*(u,v) = \arg\min_{d(u,v)} \int_{\Omega} (\alpha(u,v)L(f(u,v)) \cdot n(u,v) - g(\hat{I}(u,v)))^2 dudv. \tag{2.1}$$

where $f(u,v) = d(u,v)\mathbf{K^{-1}}(u,v,1)^T$ is the 3D mapping function for converting the 2D pixel to 3D point in camera coordinate ($\mathbf{K}$ is the camera intrinsic matrix). $d^(u,v)$ is the pixel-wise depth function and $d^*(u,v)$ is the optimized depth. $\alpha(u,v)$ denotes the surface albedo and $n(u,v)$ stands for the surface normal. The illumination vector $L(f(u,v))$ is a direction vector describing the direction of the light source. $g(\cdot)$ is the camera response function which converts image intensity to irradiance with the Lambertian reflectance model. $\hat{I}(u,v)$ denotes the measured pixel intensity at pixel (u,v).

2.1.3 Electromagnetic tracking device for navigation

Although approaches like visual odometry [30] can generate camera pose free of extra equipment, it should be addressed that the global navigation devices with better accuracy

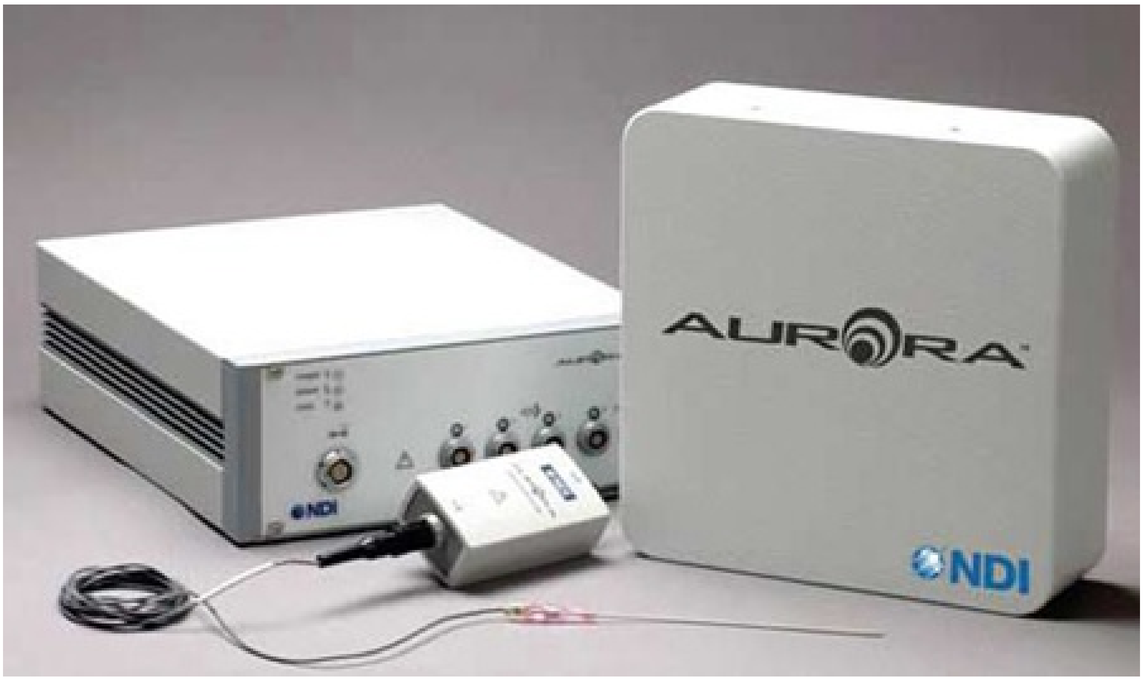


FIGURE 2.3: NDI Aurora ®EM tracking system [1]

and unaffected by the image quality, fast movement, and non-rigid environment is more suitable for MIS system. Among them, electromagnetic sensor (EM sensor) is the most widely used and researched device for position tracking of camera, instrument or other equipments in MIS [31][32][33]. According to a previous surveys and experiments, the EM sensor provides the global pose with higher accuracy and more steady output than other devices. Take the equipment 'NDI Aurora (Fig.2.3) ®EM tracking, 5 DoF' for example, the accuracy (2-sigma confidence interval) is 1.40 mm (position) and 0.35°. It is so accurate that can even be treated as ground truth in a deformable environment.

Before applying the EM sensor, a hand-eye calibration is necessary to estimate the transformation from the EM sensor coordinate to camera coordinate. Fig. 2.4 shows the workflow for implementing the hand-eye calibration between the camera and the EM sensor. The experimenter measures all the corners of the checkerboard and estimates an optimal position of the checkerboard by taking advantage of the know squares size. After that, the objective function illustrated in Fig. 2.4 is applied to find the best transformation from EM sensor to camera.

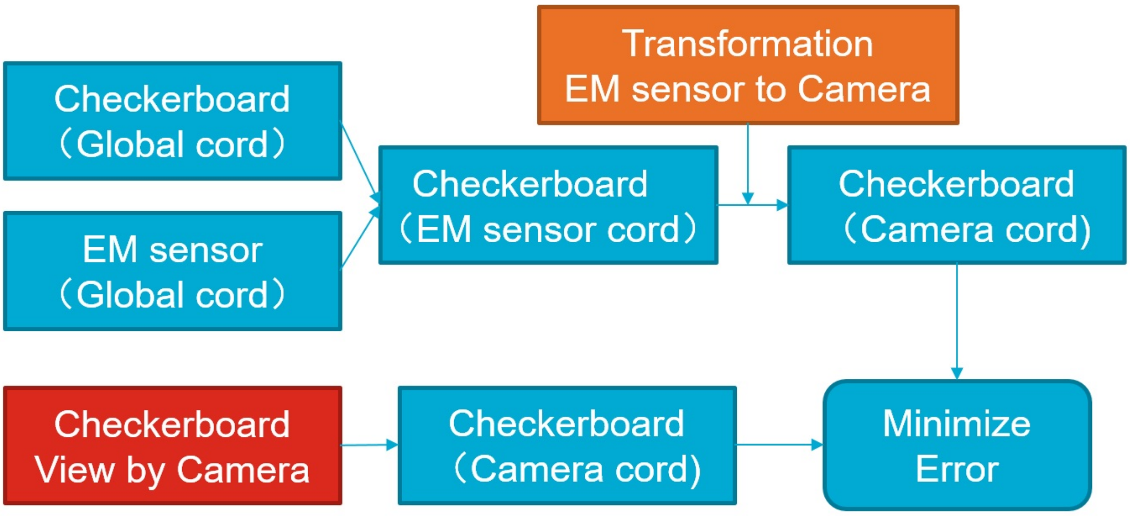


FIGURE 2.4: The framework of Handeye Calibration.

2.2 Rigid SLAM in non-rigid environment.

While 3D shape estimation algorithms provide basic ingredients for mapping the intra-operative space, visual odometry technique or EM sensor are widely adopted to provide the global pose of the scope. With the achievement of tracking technique in surgical navigation, consecutive 3D shapes may be merged or mosaic for the whole soft-tissue reconstruction. To solve the problem of camera pose navigation and 3D reconstruction, SLAM has been widely used. However, the major difference between the non-rigid SLAM and the general rigid SLAM in the scenery with the deformable object is that in surgical vision all the environment is prone to deform and visual odometry cannot be directly applied to estimate the global pose of the camera. Conventionally, moving/deforming object are masked from the static background and global camera pose is estimated only with static objects [34].

2.2.1 Sparse rigid SLAM in MIS

Hu et al. [35] applies a probabilistic principal component analysis based non-rigid SfM technique to estimate the monocular camera pose and reconstruct the beating heart. To avoid the complexity of deformation, the image sequence is deliberately arranged and the framework is run offline. Grasa et al. [16] proposes an extended Kalman filter (EKF) based

SLAM with the modified 1-Point Random sample consensus (1-Point RANSAC) algorithm for spurious points detection and rejection. This method is based on the assumption that large parts of the scenery are static and the camera moves steadily. With the efficiency of the EKF-SLAM formulation, this framework runs in real-time and achieves a good result on slight deformation and slow camera movement. Collins et al. [36] proposes a sliding window based rigid SfM approach to reconstruct 3D shapes off-line. Strictly speaking, just as Lin et al. [5] pointed out, [37] is the first research on non-rigid SLAM of soft-tissue in MIS because the motion of the liver is estimated with a periodic respiration model and current state of the model is estimated by temporally tracking the 3D points on the liver surface with stereo cameras. But this work is a template based approach for it is based on prior knowledge of the 3D model and deformation patterns. Thus this approach is limited to be applied for general use. Lin et al. [17] extends monocular PTAM [38] to stereoscope and proposes to RANSAC to detect the deforming points based on the fact that only rigid points satisfy a global Euclidean transformation. The removal of those deforming points resulted in a more accurate and stable camera pose estimation. Mahmoud et al. [39], Mahmoud et al. [40], Turan et al. [41], Chen et al. [42] and Marmol et al. [43] exploit and tune a complete and widely used large scale SLAM system named ORB-SLAM [44]. They analyze and prove that ORB-SLAM is also suitable for scope localization in MIS. In [40], a quasi-dense map is generated off-line based on pose imported from ORB-SLAM. Similar to the work of Grasa et al. [16], these approaches are also based on separating static and deforming points for better camera pose estimation.

Moreover, special purposes like visual servoing, pick-up or operation, 2D or 3D image fusion without focusing on camera pose estimation has also been analyzed. Efforts like [45] and [28] concentrate on generating a 2D mosaic image, then fuse these images into 3D shape and perform visualization.

2.2.2 Dense rigid SLAM in MIS

With the evolution of computational power of GPU, rendering high-definition graphics scenes with tremendous inherent parallelism becomes possible. Furthermore, parallel stream processing ability is fully exploited in numerous fields demanding heavy parallel computation ability. The numerical computation power is named GPGPU including bioinformatics, medical imaging, machine learning, statistics, physics, etc. [46]. In the SLAM and computer vision community, multiple processing unit integrated on GPU can process simple but heavy point cloud management in an efficient manner; a normal scenario consists more than 100000 vertices for manipulation. Efficient point cloud management in such a scale is a mission impossible for conventional sequential CPU processor.

Dense rigid SLAM can be classified as point cloud based and volumetric mesh presentation in terms of mapping. Occupancy map [47] is introduced in the 3D SLAM system to ensure expressing 3D space in volumetric occupancy map. Later, KinectFusion [10] is introduced to simultaneously localize camera as well as build a high quality 3D map [10]. The point cloud method, or defined as surfel, is a straight forward way to model the environment because the input is in the form of the point cloud (RGB-D or stereo sensors). The model point is defined with 3D location and associated weight, activeness property, etc. Henry et al. [48], Keller et al. [49] and Whelan et al. [50] have demonstrated that the reconstructed surface and texture is vivid and accurate while keeps memory consumption in reasonable scale. Fig. 2.5 shows an example of 3D reconstruction of ElasticFusion. McCormac et al. [51] further improves ElasticFusion [50] with a semantic mapping technology.

2.3 3D non-rigid SLAM

2.3.1 Non-rigid RGB-D SLAM

Much progress has been reported on the incremental 3D model reconstruction of deformable objects or moving human bodies. After the pioneering work of KinectFusion [10] which makes use of RGB-D, efforts have been devoted on making full use of the real-time RGB-D information for obtaining the current shape and the pose of the model. With reference to KinectFusion in the static and rigid scenario, Zollhöfer et al. [11] first attempts to transfer KinectFusion's idea in non-rigid body construction and simulation. Later on, DynamicFusion [12] and VolumeDeform [13] are proposed for more accurate 3D object reconstruction and simulation. Their template free work has achieved great success in both reference model construction and model deformation prediction. A compelling Fusion4D method is demonstrated in [14], where the topology changes are considered comparing to DynamicFusion. While different from previous work, multi-view RGB-D cameras are used instead of a single RGB-D camera. These DynamicFusion like techniques may be applied to a new way of sports broadcasting or immersive telepresence in other geographic locations in the future.

Despite the amazing result, these works cannot be directly applied in surgical vision due to the limitations of sensors and the high accuracy standard in surgery. There are two major requirements in clinical applications limiting applying DynamicFusion [12] like pipeline into surgical vision. First, due to the spatial and computational limitations, the TSDF pipeline requires a predefined volume and only allows the target object to move within this boundary. While in the MIS scenario, due to the limited field of view of the scope, the surgeons always require the scope to move freely in the space in order to observe more areas of the tissue in the interventions. Volume deforming approach used in [12], [13] and [14] makes computation and unnecessary data storage increases exponentially as the volume size increases, and there is a trade-off between model details (depending on the grid size) and computational cost in volume based data management. Second, different from obvious topologies in dynamic human body modeling, the smoothness of organs makes the algorithm easily converges to a local minimum. Considering the small field of view of the scope, the drifts of reconstruction caused by mismatching remains difficult to be corrected. This is different from the scenario of the large field of view since they can frequently re-observe the target as loop closure [12] or even reset the model [14] if multiple cameras are provided (8 sets of depth cameras are used in [14]). Even a slight drift leads to misalignment in textures especially on vascular. Thus, none of these methods are used in the application of computer-assisted interventions in MIS.

The basic framework for [10], [13] and [14] consists of RGB-D image preprocessing, deformation graph estimation and map fusion. The model is initialized with the depth from the first frame. Then, each time when new depth is acquired, potential visible points from the model are extracted and projected onto the 2D RGB images. Some approaches [52] may apply keypoint matching algorithms like scale-invariant feature transform (SIFT) to extract a set of key points for running an initial rigid point cloud transformation estimation.

It transforms the model to a good initial position for better warping field estimation. The optimal warping field is estimated by minimizing the energy function in the form of a sum of squared distance between sparse key points and dense visible points. After the estimation of the warping field, the last updated model is deformed to fit the new observation, predicted to deform the unobserved tissues by the 'As-rigid-as-possible' principle [53] and fused with new observation to the target scan. Through this pipeline, a live model with deformation can be aligned with new observations and built incrementally.

2.3.2 Implementing ED graph SLAM in large scale SLAM

Overall, ED graph based formulation is the most applicable and widely used approach for modeling deformations. This formulation can be optimized and updated in batch, enabling fast sequential or parallel implementation. However, ED graph formulation also comes with disadvantages and limitations. One major issue is that when a new observation is incorporated, the number of nodes increases dramatically, posing a heavy computational burden. Little attention has been paid in the field of truncated signed distance function based 3D human reconstruction because the target size, as well as the map extent, are predefined. State estimation and map updating are all confined within a volume. In more general cases, however, as reported by [19], when reconstructing geometry without a predefined volume, the size is unbounded due to the non-stop growth of the graph and an amortized $O(n^2)$ complexity with respect to the number of nodes in the graph. Equivalently, optimizing an expanding ED graph in an unconstrained space significantly limits the performance of the system.

Even though little is known in fastening ED based system performance, numerous researches are devoted to optimizing rigid pose graph, similar to the undirected graph with a different physical meaning. Pose graph defines nodes as robot poses (and landmark positions in the case of feature-based SLAM), while edges as measurements between nodes [54]. Parameters of nodes (robot poses or feature poses) are the states to be estimated. Pose graph sparsification is the most widely applied technique to marginalize subsets of nodes [55] [56] [57]. The key process in this topic is the sparsification of edges and marginalization of nodes based on some indicators like Kullback–Liebler divergence [57]. Identically, conclusions from numerous nodes marginalization method are of great value to enable efficiency in ED based SLAM. This research tries to solve the problem raised by [19] that the computation complexity in an expanding environment is $O(n^2)$. We analyze the spatial relationship between ED graph and observation and discover the inherent sparsity of the Hessian matrix. Based on this discovery, we classify ED nodes into points relevant (PR) nodes and points irrelevant (PI) nodes and propose a decoupled optimization strategy.

2.3.3 Template based non-rigid SLAM

In the meantime, a different deformation formulation named monocular finite element method (FEM) is proposed by discretizing a geometry into elements presented with 3D locations. Fig. 2.6 shows one example of curved shape representation in the form of FEM. FEM is composed of the grid with nodes and the connecting edges. Node displacements define the deformation of the grid. Stiffness is exerted by the parameters controlling behaviors of nodes sharing the same edge. Obviously, space within the grid is interpolated to simulate dense deformable surfaces. A linear elastic solid (Ω) with the steady state Navier's equations with Eq. 2.2 and the boundary conditions Eq. 2.3 model the solid deformation are presented here. Both are shown in Einstein's index notation [58], where $a_{i,jj}$ and $a_{j,ij}$ are displacement vectors share the same edge j. Ω is the solid boundary and Γ is the boundary. λ and $\mathbf{G}$ are the Lam'e parameters that define the material elastic properties, both of them are defined in terms of the Young's modulus, $\mathbf{E}$ and the Poisson's ratio, $\mathcal{U}$, being $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}$ and $G = \frac{E}{2(1+\nu)}$.

$$(\lambda + G)a_{j,ij} + Ga_{i,jj} + f_i = 0 \text{ in } \Omega \tag{2.2}$$

$$a_i = a_i \quad in\ \Gamma \tag{2.3}$$

Several works have reported adopting FEM as a way to formulate deformation in SLAM problem [59] [60].

Similar to FEM, structure from template (SfT) [61][18][62][63][64] is also proposed to simulate deformation. Like grid based FEM, they directly adopt the triangular mesh template to describe the soft-tissue. These templates are generated from other methods like SfM or computed tomography (CT). However, to the best of our knowledge, it is hard to apply these methods when the map is incrementally built, and no complete real-time implementation has demonstrated how it will effectively be applied in large scenarios.

In the MIS community, due to sensors and application requirements, CT scanning is a standard process before surgery. Thus, the pre-operative CT data provides an ideal detailed prior model for recovering the deformation. The template based methods overtake traditional non-rigid SLAM in terms of the localization error. In SLAM systems, the error in the propagation step accumulates, making the deformation parameter sets more unreliable. These template based methods, however, successfully avoid mapping errors. However, to the best of our knowledge, all template based methods are difficult to incrementally build the map and no complete implementations have demonstrated how it will effectively be applied in large scenarios.

2.4 Ambiguity in deformable surface 3D motions

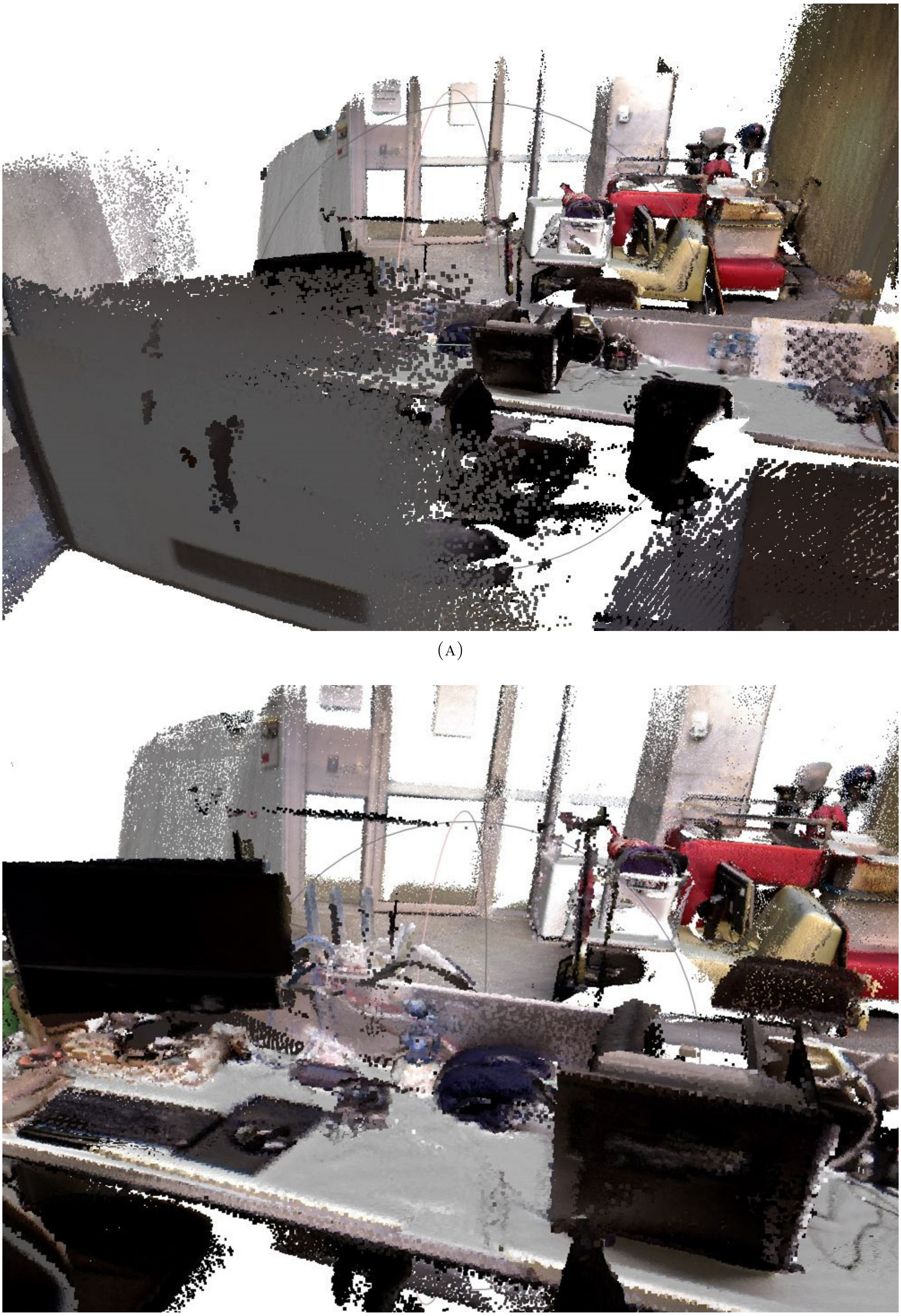
Recovering pose and dynamic three dimensional (3D) shape of a soft and deforming object from multiple images is one of the central research topics in the computer vision community. Time-varying 3D deformable structure recovery enables the virtual 3D model for CAD modeling, virtual reality, mixed reality, and robot motion planning. In addition to an on-line non-rigid formulation like ED graph, FEM and template based methods, non rigid structure from motion (NRSfM) is also a heavily researched tool aiming at recovering 3D deforming points from 2D tracked pixels on series monocular images. Early researches like Bregler et al. [65] solve the NRSfM scenario with a factorization framework that is widely used in SfM. Later researches discover that different from SfM with strong rigidity constraint, NRSfM suffers greatly from a high degree of freedom and ambiguous solvability [66]. Since these works, researchers realize that NRSfM is an ill-posed problem that multiple different deformation and shape generate the same 2D observations. Therefore, different priori are introduced to constrain the problem into a low-rank space in order to achieve solving the problem with a unique solution. A very important theoretical breakthrough [67] proves that a unique shape structure can be achieved with orthonormality constraint with given base shapes. However, the only orthonormality constraint is not enough for recovering unique base shapes and combinations.

Throughout all these state-of-art deformation formulations, when modeling deformation and global motion of a soft surface, the pose can be mixed with deformation formulation, making the rigid and non-rigid motions non-separable. Fig. 2.7 shows a toy scenario demonstrating typical ambiguity between rigid motion non-rigid deformation. In both examples, the camera observes exactly the same images because the relative movement between the camera and the soft-tissue remains the same. Apparently, the large rotation within two models is simultaneously caused by rotation of camera and the same rotation of heart. Since all methods describe rotation as the movement of model points, the rotation can both be rigid and non-rigid. Thus, there is an inner-connection within the global camera pose and the local deformation formulations. Theoretically, their interaction makes them not separable and cannot be uniquely determined in conventional formulations. Rigid rotation and transformation are embedded in local deformation. To overcome ambiguity and to reduce difficulty in the monocular dataset, researchers in NRSfM turn to add priori in NRSfM problems.

Base shape priori is a well-studied area. The basic idea is that all 3D deformed shape can be expressed with a linear transformation of base shapes. Therefore, the shape with a high degree of freedom is constrained in low-rank subspace ensuring smooth surface as well as motions. Bregler et al. [65], Xiao et al. [68] and Dai et al. [69] adopt this idea with different innovations. Torresani et al. [70] extends base shape decomposition with probabilistic principal component analysis for enforcing low rank on space. Bartoli et al. [71] introduces 'coarse-to-fine' for reducing high ambiguity and automatically choose the best number of base shapes. Lee et al. [72] enforces two consecutive shapes are aligned with generalized procrustean analysis. As a duality formulation to base shape presentation, base trajectory analysis is also proposed by Akhter et al. [73] and Valmadre and Lucey [74]. Later, more complex form shape-trajectory bases [75] enters the vernacular of NRSfM researches.

Instead of the implicit low rank formulation as bases shape, trajectories or shape-trajectory, Fragkiadaki et al. [76] and Dai et al. [69] explicitly impose low rank constraint for spatial-temporal smoothness of shape. Further works extend this idea by spanning the model with a union of low dimensional shape subspace [77] [78] [79].

After observing deformation is a result of forces in the physical world, numerous researches turn to physical priori for more technical sound methods. Agudo et al. [80] and Wuhrer et al. [81] use finite element for modeling the dynamic motions of soft tissues. These methods use inextensibility based or linear elastics to model the whole soft surface deformation based on the observation. Agudo and Moreno-Noguer [82] even model the relationship between the force and its corresponding deformation. This method proves to be a great success in understanding the mechanics between force and deformation.



(A)

(B)

Figure 2.5: A demonstration of 3D reconstruction of ElasticFusion.

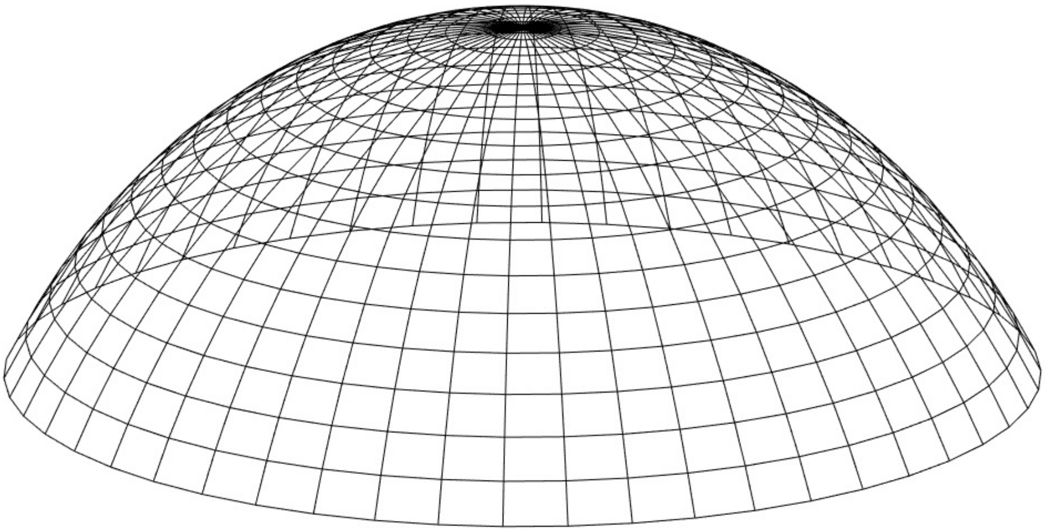


Figure 2.6: An example of FEM based curve surface representation.

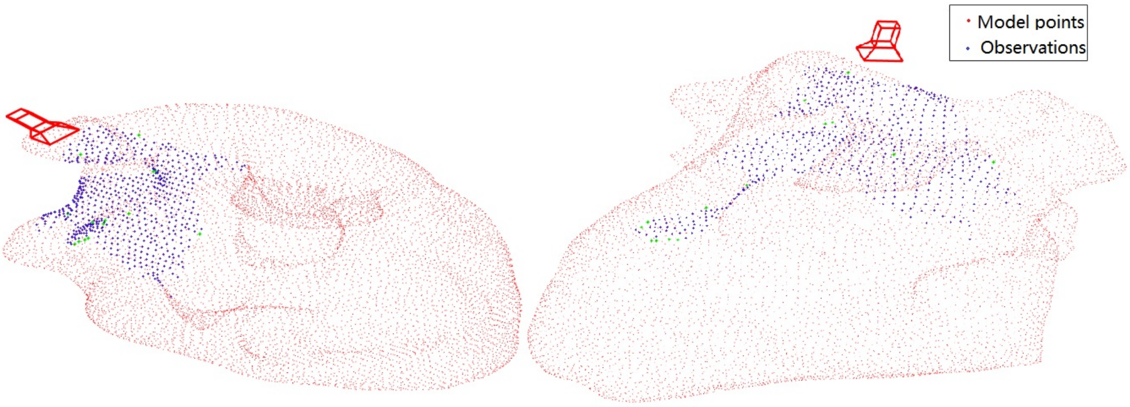


Figure 2.7: A toy model demonstrating ambiguity between camera and soft-tissue. Red dots constitute the heart and blue dots are the observation from camera.

Chapter 3

Modeling soft-tissue deformation with ED graph

The purpose of this chapter is to propose a formulation to describe the deformation of the soft-tissue in MIS. Clinically, MIS narrows surgeons' field of view when they conduct operation with elongated equipment [3]. To solve this problem, 3D laparoscopy is applied to provide two images to create an 'imagined 3D model' for surgeons. Inspired by the fact that stereo vision can generate shapes for qualitative and quantitative purposes, stitching all the 3D shape by taking account deformation will make better use of 3D information. Moreover, it is helpful if a dynamic 3D morphology is incrementally generated and rendered for the intra-operative surgery, future autonomous surgical robots (known as computer assisted intervention), implementing surgical operation and navigation. However, the small field of view of the scopes and the deformation of the soft-tissue limit the feasibility of using traditional SfM and image mosaicking methods. Even worse, rigid and non-rigid movement, scope rotation and translation, breathing, heartbeat and instrument interaction increase the difficulty in soft-tissue reconstruction and visualization. Therefore, a robocentric deformable SLAM system involving incrementally recovering the morphology and motion of soft-tissues with intra-operative stereoscope manipulation is necessary. **This chapter focuses on theoretically solving the key issue, modeling soft-tissue deformation, to enable a complete robocentric MIS SLAM system.** Specifically, instead of the monoscope, this research is based on the stereoscope since point cloud can be directly recovered from the disparity of stereo image pairs [15].

Utilizing valuable insights from previous researches [83][12][14], this work introduces ED graph as a tool to model deformation of soft-tissue. To validate the effectiveness of modeling deformation, we apply ED graph deformation formulation on two scenarios: robocentric template based SLAM and template free SLAM.

In robocentric template based SLAM, this work makes use of the morphology information of the soft-tissues from X-ray or CT and deforms it with the ED graph deformation. Here, the key is to build a distance field function of the scan from the depth sensor, which can be used to perform accurate model-to-scan deformation together with robust non-rigid shape registration in the same go.

In robocentric template free SLAM, point cloud based method is proposed to substitute the volume based model management, which not only avoids the blurring surface but also manages the texture/color information including model rendering, feature points extraction and fusing new observation. This research also applies dense speeded up robust feature (SURF) descriptors for providing a mass number of pair-wise registering key points, which can greatly overcome the texture gaps caused by the error of the reconstruction of the deformable tissue. In all, this is the first research in the MIS community that can dynamically reconstruct the deformable dense RGB model.

3.1 Revisit ED deformation graph

Before embarking on this endeavor, let's first revisit structure of ED graph invented by Sumner et al. [83] and see how to deform the model to fit the target shape. Fig. 3.1 illustrates an example of ED graph. Nodes and their shared edges compose the ED graph. The ED graph is made up of a set of uniformly scattered sparse ED nodes accompanied by an affine matrix in $\mathbb{R}^{3\times3}$ and a translation vector in $\mathbb{R}^3$. Each source vertex on the original model is transformed to the target position by several nearest ED nodes and the influence from the node depends on the distance to the ED nodes.

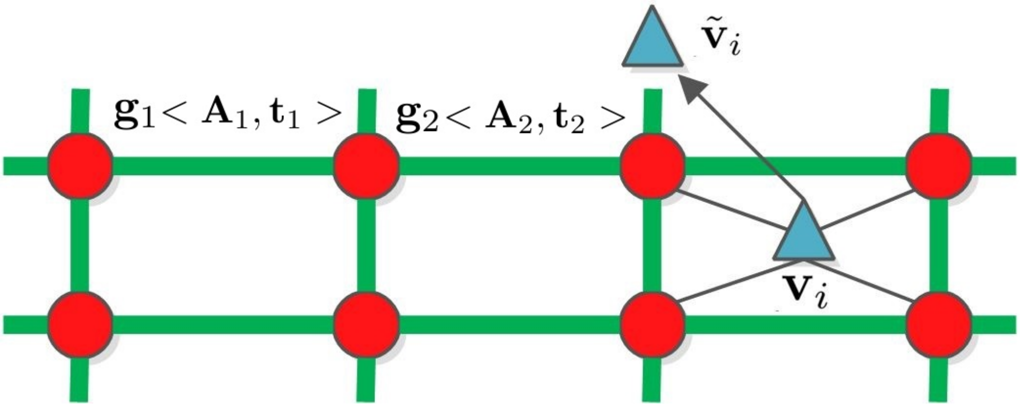

$\tilde{\mathbf{v}}_i$
$\mathbf{g}_1 < \mathbf{A}_1, \mathbf{t}_1 >$
$\mathbf{g}_2 < \mathbf{A}_2, \mathbf{t}_2 >$
$\mathbf{v}_i$

FIGURE 3.1: A toy example of an ED graph. The red circles are ED nodes, say node j, encoding a geometric position $\mathbf{g}_j$, and an affine transformation given by $\mathbf{A}_j$ and $\mathbf{t}_j$. The blue triangle is a vertex, that can be deformed from $\mathbf{v}_i$ to $\tilde{\mathbf{v}}_i$, through the impact of its neighboring ED nodes.

The jth ED node is described by a position $\mathbf{g}_j \in \mathbb{R}^3$, a corresponding quasi rotation (affine) matrix $\mathbf{A}_j \in \mathbb{R}^{3\times3}$ and a translation vector $\mathbf{t}_j \in \mathbb{R}^3$. The minimal form is a given point $\mathbf{v}$ deformed with one ED node $\mathbf{g}_j$, $\mathbf{v}$ is mapped to a new locally deformed vertex $\tilde{\mathbf{v}}$ by one ED node$\mathbf{g}_j$ in following form:

$$\tilde{\mathbf{v}} = \mathbf{A}_j(\mathbf{v} - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j. \tag{3.1}$$

In practice, this minimal non-rigid transformation can be extended to any vertex mapped by k neighboring nodes, in a mixture of deformation and rigid transformation:

$$\tilde{\mathbf{v}}_i = \mathbf{R}_c \sum_{j=1}^{k} \omega_j(\mathbf{v}_i)[\mathbf{A}_j(\mathbf{v}_i - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j] + \mathbf{T}_c, \tag{3.2}$$

where $\mathbf{R}_c$ and $\mathbf{T}_c$ are global rotation and translation relating to camera motion. $w_j(\mathbf{v}_i)$ is quantified weight for transforming $\mathbf{v}_i$ exerted by each related ED node. The number of nearest nodes is confined by defining the weight in Eq. (3.3). Deformation of each vertex in the space is limited locally by setting the weight as:

$$w_j(\mathbf{v}_i) = 1 - ||\mathbf{v}_i - \mathbf{g}_j||/d_{max}, \tag{3.3}$$

where d_{max} is the maximum distance of the vertex to $k + 1$ nearest ED node.

Eq. (3.2) formulates rigid transformation and deformation in the form of source and target vertex pairs. Given these arbitrary vertex pairs, in turn, the parameters of the ED graph can be estimated. To estimate the optimal ED graph, the problem is formulated with three terms: rotation constraint, regularization, and the point to plane distances between the visible points:

$$\underset{\mathbf{A}_1,\mathbf{t}_1...\mathbf{A}_m,\mathbf{t}_m,\mathbf{R}_c,\mathbf{T}_c}{\operatorname{argmin}} \quad w_{rot}E_{rot} + w_{reg}E_{reg} + w_{data}E_{data}, \tag{3.4}$$

where m is the number of ED nodes. Here, all the variables in the state vector for this energy function are the $[\mathbf{A}_j, \mathbf{t}_j]$ from each ED node.

To prevent the optimization converging to an unreasonable deformation, this research follows the method proposed in the ED graph [83] which constrains the model with rotation and regularization.

Rotation E_{rot} enforces the affine matrix close to $SO(3)$ by minimizing the following function of the column vectors $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$ of $\mathbf{A}$:

$$E_{rot} = \sum_{j=1}^{m} Rot(\mathbf{A}_j), \tag{3.5}$$

$$\begin{aligned} Rot(\mathbf{A}) = (\mathbf{c_1} \cdot \mathbf{c_2})^2 + (\mathbf{c_1} \cdot \mathbf{c_3})^2 + (\mathbf{c_2} \cdot \mathbf{c_3})^2 + \\ (\mathbf{c_1} \cdot \mathbf{c_1} - 1)^2 + (\mathbf{c_2} \cdot \mathbf{c_2} - 1)^2 + (\mathbf{c_3} \cdot \mathbf{c_3} - 1)^2 \end{aligned} \tag{3.6}$$

Regularization. The basic idea for this term is to prevent divergence of the neighboring nodes exerts on the overlapping space. It corresponds to the widely accepted idea 'As-rigid-as-possible' in the computer vision community [84]. The quantity for this term represents the difference of deformation exerted by the neighboring node and itself should be close. Otherwise, the deformed surface is not smooth. Therefore, Sumner et al. [83] introduces the term E_{reg} to sum the transformation errors from each ED node.

$$E_{reg} = \sum_{j=1}^{m} \sum_{k \in \mathbb{N}(j)} \alpha_{jk} ||\mathbf{A}_j(\mathbf{g}_k - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j - (\mathbf{g}_k + \mathbf{t}_k)||^2, \tag{3.7}$$

Similar to the ED graph [83], α_{jk} is defined as the overlap influence of the two ED nodes but in practice is set to 1. $\mathbb{N}(j)$ is the set of the neighboring node to node $\mathbf{g}_j$.

Data Term. Data term defines the arbitrary errors from the deformed shape to the target shape. In the original ED graph [83], it is the Euclidean distance between the deformed key points and the target key points. Other researchers Newcombe et al. [12], Innmann et al. [13] and Dou et al. [14] define the model to frame with average point to plane distance, and the pairing alignment is defined with 'back-projection'. There are various data term formulations measuring the deformed shape and the targets. The general form is:

$$E_{data} = \sum_{i=1}^{m} \boldsymbol{\mathcal{F}}(\mathbf{v}_i), \tag{3.8}$$

where $\boldsymbol{\mathcal{F}}(\cdot)$ is a general function defining a point to target distance.

3.2 Template based SLAM with ED deformation graph

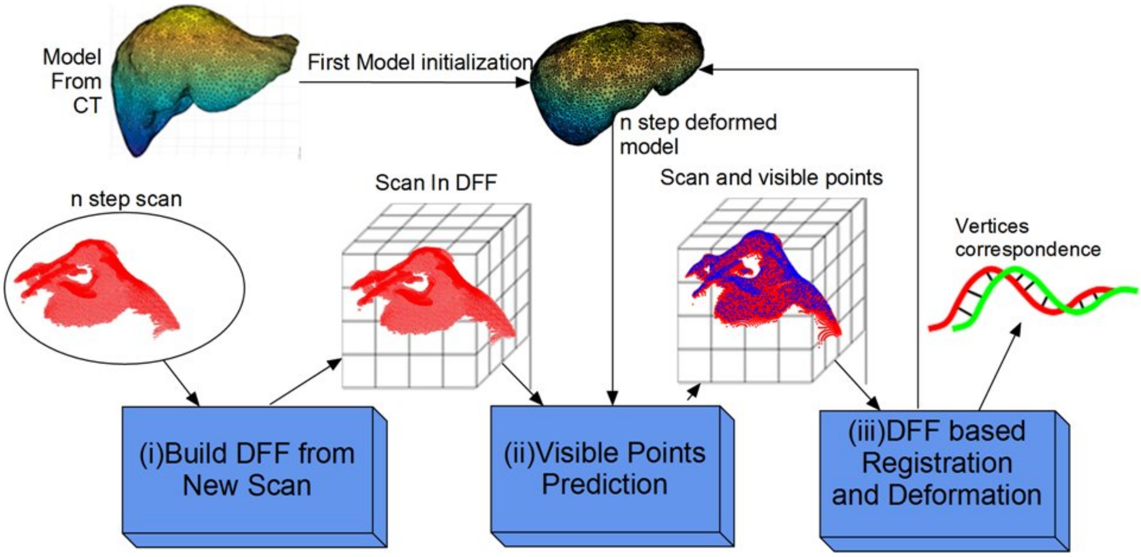


Figure 3.2: The framework of the proposed template based deformable soft-tissue reconstruction based on DFF and pre-operative CT model.

3.2.1 The framework of template based structure

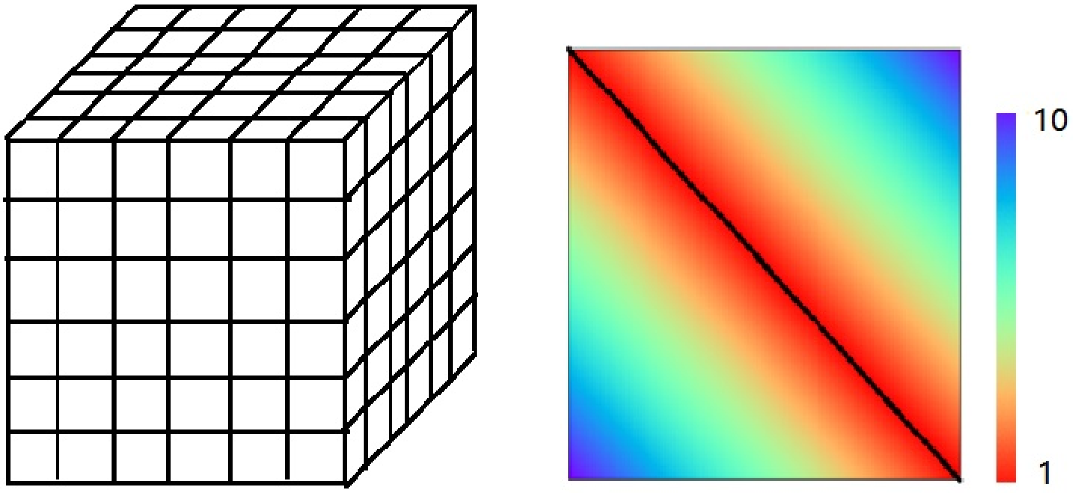
3D mesh models of the soft-tissue can be segmented from the pre-operative CT scans [85]. In this section, a synthetic experiment is conducted to simulate the robocentric template based SLAM with the ED deformation graph. This test is based on 3D segmented CT scanning and RGB-D scope. Several researches have adopted structure light [86] or time of flight sensors [87] for 3D intra-operative imaging. Therefore, in this test, we simulate a 3D mesh model generated from the pre-operative CT scan. This research also simulates the sequence of the model deformation and the partial observations from a perspective RGB-D scope. The aim is to deform the model with regard to the RGB-D scan observed. The proposed framework for recovering the deformation of the soft-tissue consists of three steps (Fig. 3.2):

(I) Compute the distance field function (DFF) for the new scan.

(II) Predict visible points from the latest update of the deformed model.

(III) Deform the current model to fit the new scan. Both the deformation of the model and the non-rigid registration between the model and the new scan are accomplished simultaneously. The model is initialized using the reference model segmented from the pre-operative CT scan at the very beginning.

Step I is an on-line pre-process. As shown in Fig. 3.3, the DFF is a volume, with each voxel recording the distance to its nearest point on a new scan. DFF is not only employed in the model-to-scan deformation and registration processes (Step III), but also in the selection of the visible points in Step II. As DFF is only built on the latest scan, dynamically building DFF does not require a heavy computational cost. Different from most existing approaches which traverse all the point to plane distances and use a threshold to decide visible points [14], this research uses regularized DFF volume directly by looking up the value of each vertex in DFF and the derivative functions generated from DFF, and compare it with the threshold for determining point visibility. This strategy reduces the computational cost of the visible points selection process significantly. After selecting the visible points, a cost function is formulated adjusting the ED graph parameters, to deform the visible



(A) (B)

Figure 3.3: (A) is the DFF volume recording distance field values. (B) is a section of the volume. The black line is an example planar and each voxel records its distance to the planar.

points close to the target scan. As described in DynamicFusion [12], a spatial ED graph deformation, as well as a source to target correspondence, is built. Based on the optimized ED graph, this research deforms the latest updated model to the current scan, not only considers current observation but also obeys 'As-rigid-as-possible' principle proposed by Sorkine and Alexa [53] in surface deformation. 'As-rigid-as-possible' principle enables that invisible part of the model can be inferred from current observation.

3.2.2 Technical details

Following the general ED graph estimation energy described in Section 3.1, the main formulation of the template based method consists of three terms: Rotation, Regularization, and Data. Specifically, Data is defined by the distances between the model and the target scan. Rotation and Regularization are defined in Eq. (3.5) and Eq. (3.7).

Distance to the target scan. This work modifies the general data term Eq. (3.8). After determining the rotation matrix and transformation vector of ED nodes, all vertices in the mesh are transformed to their new positions, and the distances between these vertices on the deformed model to the target scan are minimized. These distance values can be easily looked up from a predefined loss function DFF. The lower the value is, the closer the deformed vertices to the target surface. Note that the deformed vertices are not necessarily to the correct correspondences but at least close to the surface. The positions of vertices are transformed and compared in Eq. (3.9). Minimizing this term is equivalent to deforming the transformed model close to the target surface of the scan:

$$E_{data} = \sum_{i \in \mathbb{L}} ||\mathcal{D}(\tilde{\mathbf{v}}_i)||^2, \tag{3.9}$$

$\mathcal{D}(\cdot)$ is the corresponding voxel value recorded in DFF. $\mathbb{L}$ defines the set for all the visible points for calculating sum distance error E_{data}.

This section modifies the directional distance function proposed by [88] as a DFF by ignoring the directions. Unified volume based distance function provides a robust and efficient target loss function for surface matching. This template based research records at each voxel its distance $\mathcal{D}(\cdot)$ to the closest point on the surface of the target scan.

3.3 Robocentric template free SLAM with ED deformation graph

3.3.1 The robocentric template free SLAM framework

Different from the template based SLAM, the goal of the proposed template free SLAM framework is to recover and fuse the deformation of the soft-tissue consists of depth estimation from stereo images without prior shapes. The process includes sparse key points extraction and matching, ED graph estimation and new data fusion (see Fig. 3.4). In the proposed framework, first, depth is estimated from the stereo RGB images captured from the scope intra-operatively. And the model is initialized with the colored point cloud from the first frame. Then, each time when new stereo images are acquired, the system extracts potential visible points from the model and projects them onto 2D RGB images. Dense SURF algorithm is applied to extract a massive number of key points for running an initial rigid point cloud transformation estimation. This sets the model to a good initial

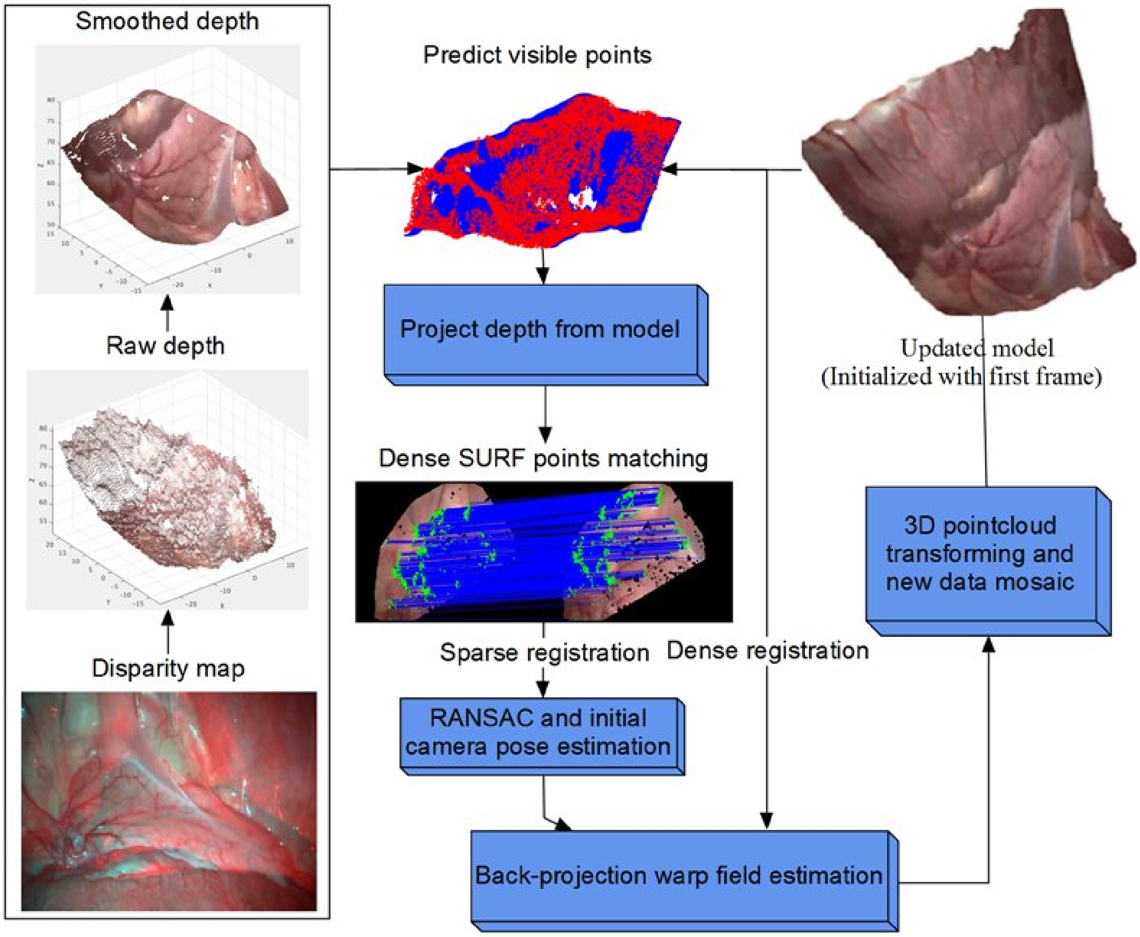

Smoothed depth
Raw depth
Disparity map
Predict visible points
Project depth from model
Dense SURF points matching
Sparse registration
Dense registration
RANSAC and initial camera pose estimation
Back-projection warp field estimation
Updated model (Initialized with first frame)
3D pointcloud transforming and new data mosaic

FIGURE 3.4: The framework of the robocentric template free SLAM with ED deformation graph.

position for better ED graph estimation. The optimal ED graph is estimated by minimizing energy function in the form of a sum of squared distance between sparse key points and dense visible points. After the estimation of the ED graph, this research deforms the last updated model to fit with new observation, predicts the unobserved tissues by the 'As-rigid-as-possible' principle [53] and fuses new observation in the model. Through this pipeline, a live model with deformation can be kept track with new observations and built incrementally.

In comparison with conventional scenarios in [12] [13] [14], there are three major challenges in surgical vision: First, scope has very narrow field of view which makes the observed information in each frame limited. Second, most of the soft-tissues have a smooth surface and do not have many distinct geometric features that can be applied in the registration process. In practice, the registration without key points results in great drifts. Last, since scope has a small field of view, the proposed method encounter blurry images in the process of key points extraction.

3.3.2 Technical details

Efficient large-scale stereo (ELAS) algorithm [89] is adopted as the depth estimation method. This research also applies similar strategies as in VolumeDeform [13] by adding control anchor points to enhance the robustness. Dense SURF is used to yield dense feature points descriptors [90] for the model to frame registration. The basic idea of the dense SURF is to directly set dense grid of locations on a fixed scale and orientation instead of detecting spatially invariant corner points. In this way, dense SURF provides much denser key points than conventional scale-invariant feature transform (SIFT) or SURF image descriptor extractors. Dense SURF is more robust in handling low-quality images and provides more points to enhance robustness. Another difference from DynamicFusion [12] is that for stability this research projects the colored point cloud to RGB and depth map in the last scope coordinate and run dense SURF between projected RGB map and new left RGB scope image. Visible points from point cloud map with RGB colors are projected onto a 'model RGB image' and matched with new RGB image observation.

After acquiring key feature point correspondences from dense SURF, this work launches an iterative rigid global transformation estimation based on RANSAC. The massive amount of key points not only provide information for rigid translation and rotation which will be used as the initial guess in the optimization to estimate the ED graph, but also filter the outliers to enhance the accuracy. After the first estimation and outliers filtering, rigid global transformation is estimated again to gain more accurate initial global transformation. The global transformation provides a good initial input for later ED graph parameter estimation.

Following the general formation defined in Eq. (3.4), we formulate the problem as:

$$\underset{\mathbf{A}_1,\mathbf{t}_1\ldots\mathbf{A}_m,\mathbf{t}_m,\mathbf{R}_c,\mathbf{T}_c}{\operatorname{argmin}} \quad w_{rot}E_{rot} + w_{reg}E_{reg} + w_{data}E_{data} + +w_{corr}E_{corr}, \tag{3.10}$$

where E_{rot} and E_{reg} are defined in Eq. (3.5) and Eq. (3.7). Similar to [12], [13] and [14], this research adopts the back-projection approach as a practical model registration strategy that penalizes misalignment of the predicted visible points $\mathbf{v}_i$ and current depth scan $\mathbb{D}$. Data term denotes the sum of point to plane errors in the form of:

$$E_{data} = \sum_{i=1}^{N} (\mathrm{H}(P(\tilde{\mathbf{v}}_i))^T (\tilde{\mathbf{v}}_i - \Gamma(P(\tilde{\mathbf{v}}_i))))2 \tag{3.11}$$

where $\Gamma(\cdot) = \Pi(P(\mathbf{v}))$ and $\mathrm{H}(\cdot)$ is the corresponding normal to the pixel u in the depth $\mathbb{D}(u)$ ($\mathbb{R}^2 \to \mathbb{R}^3$). $P(\mathbf{v})$ is the projective ($\mathbb{R}^3 \to \mathbb{R}^2$) function for projecting visible points to depth image.

This research also adopts the strategy for extracting visible points from last model, by filtering points with distance and normal to current depth with thresholds. Where ϵ_d and ϵ_n are thresholds of the distance and angle.

$$||\mathbf{v}_i - \Gamma(P(\mathbf{v}_i))|| < \epsilon_d, \quad \mathrm{H}(\mathbf{v}_i) \cdot \mathrm{H}(P(\mathbf{v}_i)) < \epsilon_n. \tag{3.12}$$

As shown in [14], back-projection, and point to plane strategies make full use of the input depth image so that the Jacobians can be calculated in 2D which leads to fast convergence and robustness to outliers. As depth generated from stereo images are not as accurate as of that from depth sensors like Kinect, the visual hull terms recommended by Dou et al. [14] is not applied because the empty space and free space are not actually observed due to the misalignment of disparity maps.

Correspondence term E_{corr} is measured by the Euclidean distance between pair-wise sparse key points generated from dense SURF in the following form:

$$E_{corr} = ||\tilde{\mathbf{V}}_i - \mathbf{V}_i|| \tag{3.13}$$

where $\tilde{\mathbf{V}}_i$ and $\mathbf{V}_i$ are the 3D points of current frame and deformed points from last frame of SURF features.

Model update by new observation. Previous works adopt TSDF volume to store and fuse models [10] [12] [13] [14]. A fine mesh can be generated in real-time with the marching cube algorithm. Nevertheless, all these volume based approaches are unable to work in surgical vision because of the unknown spatial range of the target soft-tissues. To overcome this restriction, this research proposes a weighted point cloud to represent the built model. In the proposed algorithm, each point records the exact surface location with weight showing how certain it believes the record. In this work, each point stores 3 properties: position $\mathbf{v}_i$, weight $\omega(\mathbf{v}_i)$ and color $\mathbf{C}_i$.

After acquiring an appropriate ED graph, this research transforms all the points to their deformed positions and predict visible points again. For each updated point, a truncated signed distance weight (TSDW) is assigned to each pixel of new depth:

$$\omega(\tilde{\mathbf{v}}_i) = \begin{cases} d_{min}(\tilde{\mathbf{v}}_i)/(0.5 * \epsilon) & \text{if } abs(\tilde{\mathbf{v}}_i|_z - \mathbb{D}(P(\tilde{\mathbf{v}}_i))) < \tau \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

where $d_{min}(\tilde{\mathbf{v}}_i)$ is the minimum distance of model point $\tilde{\mathbf{v}}_i$ to its corresponding nodes and ϵ is the average grid size of nodes. $\tilde{\mathbf{v}}_i|_z$ is the value of point $\tilde{\mathbf{v}}_i$ on the z direction. The vertex $\tilde{\mathbf{v}}_i$ is ignored if the z directional difference is too large because of inaccurate ED graph estimation. Depth generated from current model is fused with new depth by:

$$\mathbb{D}^{n+1}(P(\tilde{\mathbf{v}}_i)) = \frac{\tilde{\mathbf{v}}_i|_z \omega(\tilde{\mathbf{v}}_{i-1}) + \mathbb{D}^n(P(\tilde{\mathbf{v}}_i))}{\omega(\tilde{\mathbf{v}}_{i-1}) + 1} \tag{3.15}$$

$$\omega(\tilde{\mathbf{v}}_i) = min(\omega(\tilde{\mathbf{v}}_{i-1}) + 1, \omega_{max}), \tag{3.16}$$

Different from rigid transformation where uncertainty of all the points in 3D space are considered as equal, in the case of non-rigid fusion, if a point is further away to the nodes of ED graph, it has small chance to be registered to the depth [12]. Therefore, this research practically measures this certainty by using the minimum distance from point to nodes and regularize it with half of the unified node distance. The upper bound of weight ω_{max} is set to 10.

3.4 Results and discussion

We thoroughly validate the proposed framework qualitatively and quantitatively. The quantitative comparison is achieved by comparing the dense recovered shape and texture, while the quantitative comparison is carried out by comparing the average Euclidean distance error between estimated points and the target depth

$$\mathrm{e} = \frac{1}{N} \sum_{i \in \boldsymbol{\Theta}} ||\Theta(\mathbf{v}_i) - \mathbf{v}_i||_2 \tag{3.17}$$

where $\Theta(\cdot)$ is an abstract function defining the corresponding registered point on the target surface (defined as the ground truth), $\boldsymbol{\Theta}$ is the domain of valid output from $\Theta(\cdot)$ and N is the number of the valid points. In template based approach, $\Theta(\cdot)$ is the $\boldsymbol{\mathcal{F}}(\cdot)$ defined by DFF while in template free approach, it is the $\Gamma(P(\cdot))$ in back-projection registration. Please note that it is difficult to align the recovered map to the ground truth with point-wise registration, thus we relax the comparison by registering all points with DFF or back-projection function. It is feasible because visual comparison validates the spectral difference based on texture while Eq. (3.17) validates the accuracy in geometric perspective.

3.4.1 Template based approach

Synthetic datasets are generated from the different real soft-tissue models to demonstrate the effectiveness of the deformation recovery algorithm proposed in this research. Three different soft-tissue models (heart, liver and right kidney) are downloaded from OpenHELP [85], which are segmented from a CT scan of a healthy, young male undergoing shock room diagnostics. In the simulation, models are randomly deformed as the ground truth by using the ED deformation graph [83]. The deformation of the soft-tissue is simulated by randomly exerting a 2-3 mm movement on a random vertex on the model with respect to the shape of the deformed model from the last frame. Then, scope poses with trajectories looped around the model are simulated to generate the point cloud scan from the randomly deformed model. Gaussian noises are added to the scope poses to simulate the data from the EM tracking system. The distance from the scope center to the model is around

200mm. Pinhole model is used to simulate the RGB-D, and the simulation scope has the intrinsic parameters:

$$\begin{bmatrix} 520 & 0 & 640 \\ 0 & 520 & 320 \\ 0 & 0 & 1 \end{bmatrix}$$

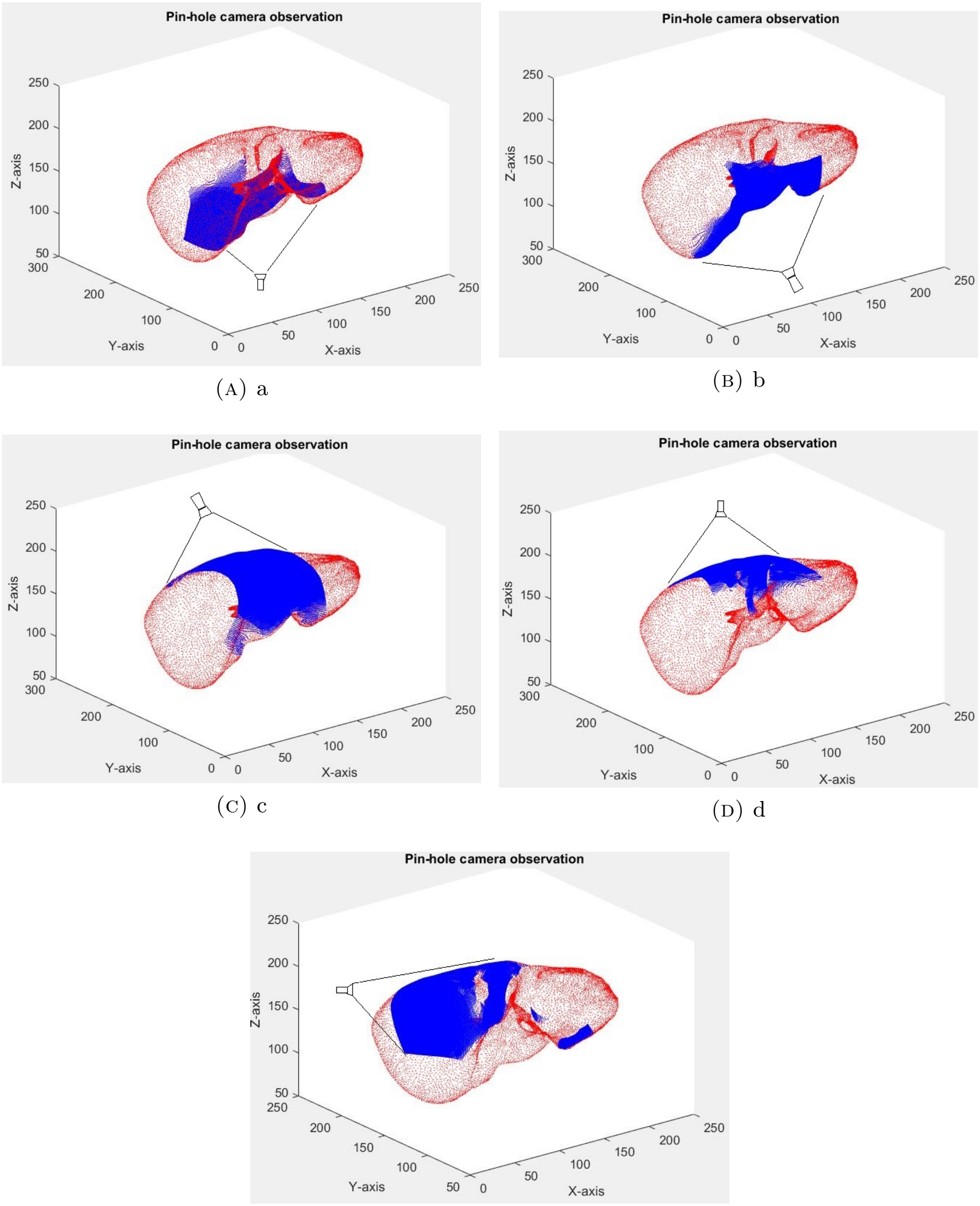
Fig. 3.5 is an example observation of a liver model. In each frame, which the recovery of the soft-tissue challenging because the scope only observes part of the deformed model.

In the model-to-scan deformation and registration process, the size of the downsampled grid is set to 20 mm to obtain the ED nodes, and the number of neighboring points is set to 4. We follow the default parameter settings of the ED [83]. The weights used in the optimization proposed in Eq. (3.4) are set to 1, 20000 and 100 for E_{rot}, E_{reg} and E_{data} respectively, which is proposed in [12] as the hyperparameter.

Fig. 3.6 illustrates the visible points-to-scan registration error map which is generated by taking corresponding value in the voxel of DFF. Results show that most points are correctly matched to the reference model and the maximum error is about 4 mm. Some significant errors are due to the deformation that cannot be described with a sparse ED graph. To solve this problem, the density of ED nodes should be increased which poses a heavy computational burden. Thus, there is a trade-off between computation and accuracy.

To illustrate the effectiveness of the robocentric template based SLAM system proposed in this research, as a comparison, the back-projection approach used in DynamicFusion [12] and Fusion4D [14] is also implemented using the same datasets. Similar to the proposed algorithm, this research defines the error from the back-projection approach to be the minimum distance from a transformed point to the closest point from the scan. The mean errors from Eq. (3.17) are used to compare the effectiveness of the two methods and the quantitative comparison of the accuracy is shown in Table. 3.1.

Fig. 3.7 presents the comparisons between the models generated from the proposed algorithm and the corresponding ground truth which are used to generate the scans from the scope. It is clear that the deformed models are close to the ground truth in the area



(A) a

(B) b

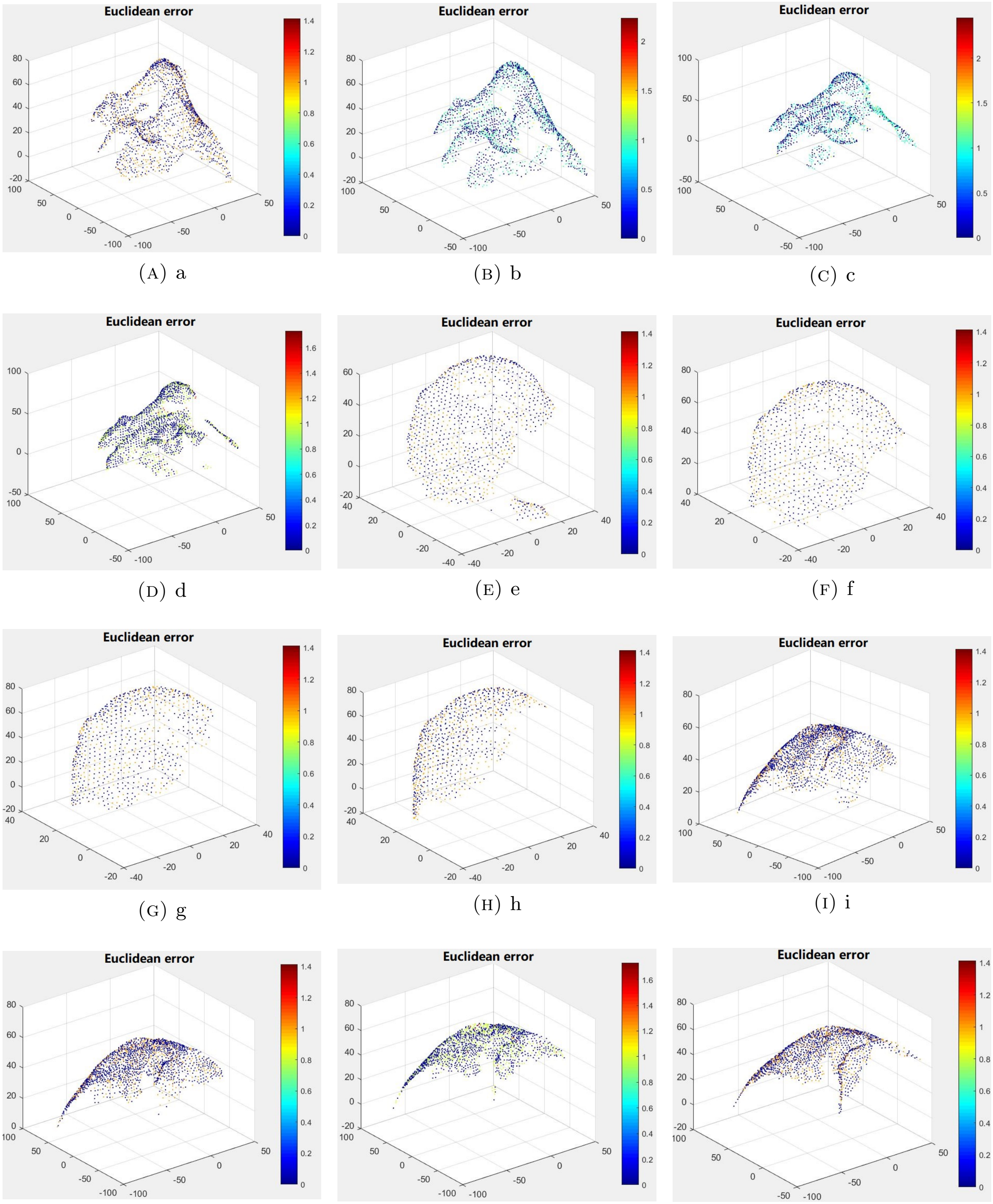
(C) c

(D) d

(E) e

Figure 3.5: (a) to (e) are the simulations of generating the depth scan observation from the deformed liver model. The blue points are the simulated depth observations. The points in red is the deformed model.

where the models are observed. On the contrary, the further the model point is away from the observation, the larger the error it has. This is due to the lack of information and the smoothness in the proposed energy function exerting on the unobserved part of the model. In other words, these unobserved points are predicted through the minimization



(A) a

(B) b

(C) c

(D) d

(E) e

(F) f

(G) g

(H) h

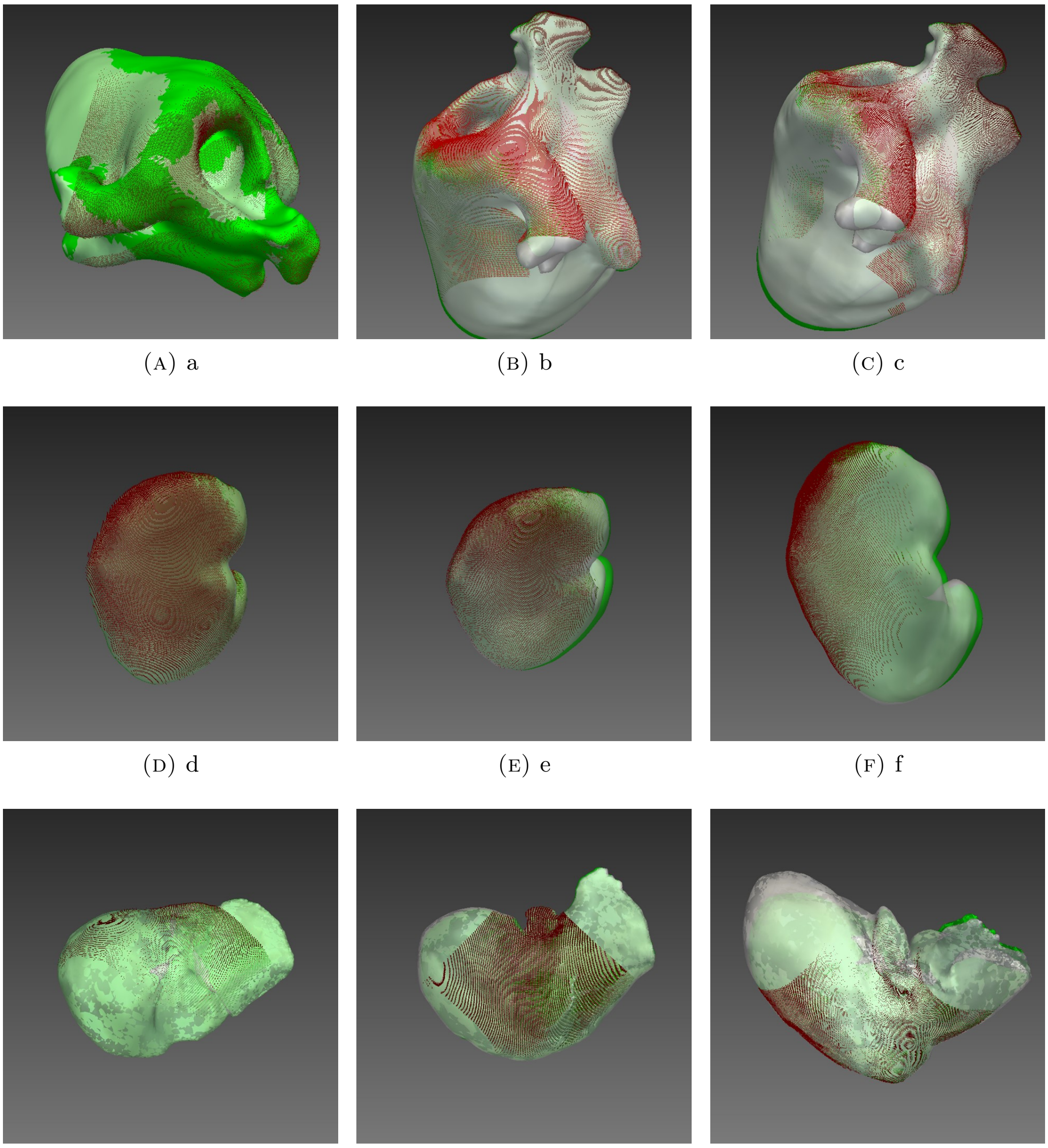
(I) i

(J) j

(K) k

(L) l

FIGURE 3.6: The results of model-to-scan registration colored by the matching error (mm) which is directly obtained from the DFF. (a)-(d) are selected error map from the heart model; (e)-(h) are selected error map from the right kidney model; (i)-(l) are selected error map from the liver model.



(A) a

(B) b

(C) c

(D) d

(E) e

(F) f

(G) g

(H) h

(I) i

FIGURE 3.7: The comparison between the deformed models recovered from the robocentric template free SLAM and the ground truth used for generating the depth observations, by using the heart model (a) - (c), the kidney model (d) - (f) and the liver model (g) - (i) respectively. The models in green are the ground truth, while the models in white are the recovered soft-tissues.

of the proposed energy function. Even though the prediction is not as accurate as of the observed tissues, it is still suitable to help surgeons (refer to the video on youtube [1]). Fig. 3.11 shows the last frame of the deformed model which is presented in the form of Axial,

[1] https://youtu.be/5HHedlXgqTE

Table 3.1: Accuracy comparison between the proposed DFF approach and the back-projection approach (mm). Each value is calculated by averaging all the points of all the frames.

	DFF based approach	Back-projection based approach
Heart	0.36	0.91
Liver	0.30	0.60
Right Kidney	0.35	0.76

Coronal, and Sagittal map. All the results demonstrate that the deformed models get significantly close to scan but areas far away to the observation show obvious errors.

In the optimization process of all the experiments, using the DFF makes the optimization, Levenberg-Marquardt algorithm, converges within 3-8 iterations. There exist no issues like singularity, divergence or bad fitting.

However, there is a limitation in the robocentric template based SLAM framework proposed in this chapter, that is it needs pre-operative data (typically the CT scan) as the initial model and EM sensor to provide global transformation of the scope. Different from DynamicFusion, in the MIS scenario, the scope is very close to the object (it is set to 200-300 mm in simulation) which limits the field of view. If only small parts of the model are observed (Fig. 3.5), the scan can be easily initialized to a different area, thus fused to the wrong shape. Considering the easy access to CT and EM sensor, this work makes full use of them for better accuracy.

3.4.2 Template free approach

Experimental setup. The robocentric template free SLAM framework is validated using the in-vivo stereo video datasets provided by the Hamlyn Centre for Robotic Surgery [91]. No extra sensing data other than stereo videos from the laparoscope is utilized in the process. The frame rate and size of in-vivo porcine dataset (model 2 in Fig. 3.10) is 30 frames per second and 640 × 480 while the rest is 25 frames per second and 360 × 288. The distance of scope to the surface of soft-tissue is between 40 to 70 mm. Due to the poor quality of obtained images and some extremely fast movement of the scope, this research deliberately chooses the videos tested on porcine with relative slow motion and some deformation caused by respiration and tissue-tool interaction.

Furthermore, to estimate accuracy, the robocentric template free SLAM is tested on two ex-vivo phantom datasets from Hamlyn [91] with ground truth from CT scan. The phantom dataset shares the same property of other Hamlyn datasets, which is 25 frames per second and 360 × 288. Moreover, synthetic data are also generated for better validation of the proposed method. In the simulation validation process, three different soft-tissue models (heart, liver and right kidney) are downloaded from OpenHELP [85], which are segmented from a CT scan of a healthy, young male undergoing shock room diagnostics. The deformation of the soft-tissue is simulated by randomly exerting a 2-3 mm movement on a random vertex on the model with respect to the status of the deformed model from the last frame [18]. This research randomly picks up points in the model as the accuracy is measured by averaging all the distances from the source points to target points. Besides, the human deformation model in VolumeDeform [13] is also employed for qualitative comparison.

The point cloud density is set to 0.2 mm and node density is set to 4 mm. The point cloud downsampling process is carried out by setting a fixed box to average points fill inside each 3-D box. The parameters for optimization are chosen as $w_{rot} = 1000, w_{reg} = 10000, w_{data} = 1, w_{corr} = 1$. The accuracy is measured by subtracting the projected model image and the observed depth image.

Note that different from kitty datasets [92], the depths generated from the fast-moving scope are of low quality. Therefore, thresholds are applied to discard some frames when their average errors are larger.

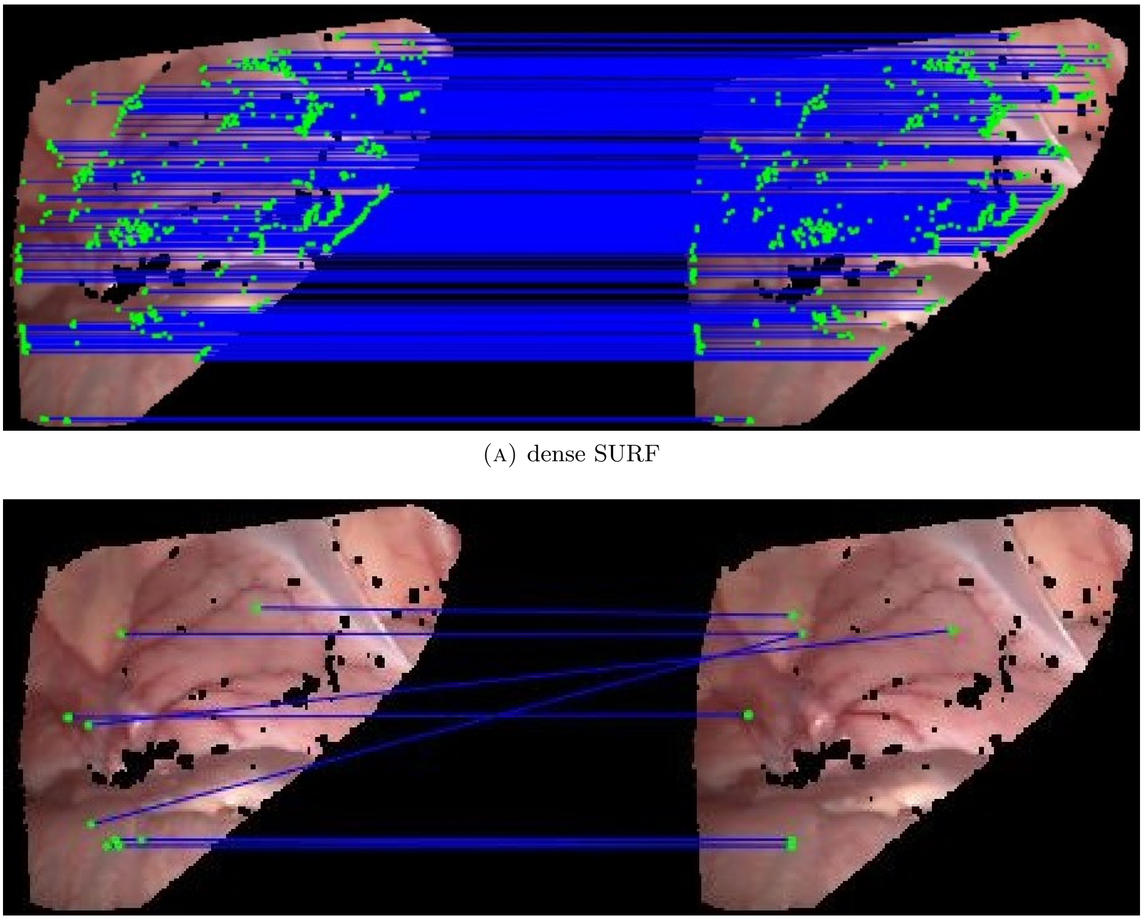
Dense SURF key points estimation. We employ the key points strategy proposed by Innmann et al. [13] to overcome the drift caused by a smooth surface and the error of the ED graph estimation. Dense SURF is applied to extract key points. The reason this chapter uses dense SURF is to deal with motion blur and low-quality images resulted from the fast movement of the scope, while the original SIFT or SURF cannot provide enough key points provided the same data. Dense SURF provides enough key points in the video of low quality or with a fast moving object. Besides, spatially scattered key points greatly enhance the stability of texture in the overlapping region. In some extreme situations, when fast movements occur in scope, no SIFT key points correspondences can be detected.

The grid sampling size of dense SURF is set to 3 pixels and a large number of corresponding points are obtained. In practice, it can be figured out that the extracted dense key points range from 100 to 1000 while conventional SIFT and SURF generate points from 0 to 200. After the dense SURF process, we refine key points and generate rigid rotation and translation by RANSAC which is a typical strategy adopted in implementing SLAM in MIS [16][17]. The threshold for filtering outliers is set as 2 mm similar to [17]. Fig. 3.8 indicates that dense SURF generates enough key points for registration. Fig. 3.9 shows that registration without key points makes the registration process converges locally which either results in disorder of texture or squeezes on the soft-tissue.

Accuracy validation. Similar to the template based approach, the accuracy is measured with the average Euclidean distance of the recovered points and the target depth (Eq. (3.17)). Fig. 3.10 shows the soft-tissue reconstruction of the proposed SLAM framework in different frames, using 5 in-vivo laparoscope datasets [91]. Results demonstrate that the soft-tissues are reconstructed incrementally with texture. The average distance of back-projection registration of the five scenarios are 0.19 mm (abdomen wall), 0.08mm (Liver), 0.21 mm (Abdomen (1)), 0.15mm (Abdomen (2)) and 0.14 mm (Abdomen (3)).

To further validate the accuracy, the robocentric template free SLAM is further tested on the synthetic datasets. Since the synthetic dataset do not provide colored 2D image from texture shape, the formulation Eq. (3.10) is simplified by ignoring the sparse registration term E_{corr}. Fig. 3.11 shows the final result of the simulation presented in the form of axial, coronal, sagittal and 3D maps (heart). The average error of the three models is: 0.46 mm (heart), 0.68 mm (liver), 0.82 mm (right kidney) respectively. The inferior accuracy to in-vivo datasets is partially attributed to lacking sparse key points registration.

Moreover, the proposed template free SLAM algorithm is also tested on two ex-vivo phantom based validation dataset from Hamlyn [91] (Fig. 3.12). As the phantom deforms periodically, this research processes the whole video and compare it with the ground truth generated from the CT scan. The average accuracies of the proposed method are 0.28 mm



(A) dense SURF

(B) SIFT

Figure 3.8: The comparison between the dense SURF and SIFT using stereo videos of abdomen wall. Results imply that dense SURF can generate more key points which are critical in soft-tissue matching while SIFT produce less or even no correspondences.

and 0.35 mm. The good result is mainly contributed by the abundant textures benefited both depth estimation and dense SURF key points extraction. The error from SIFT key points is due to the drift on the smooth surface, which is addressed in case of clothes [13].

The proposed method is also compared with VolumeDeform [13] by implementing a sample dataset, and the result is presented in Fig. 3.13. The source code of dynamic fusion and VolumeDeform are not available, one dataset published by [13] is used for the comparison. VolumeDeform claims less drift than DynamicFusion [12] approach while this research ensures less drift than VolumeDeform, due to the model-to-frame framework rather than the frame-to-frame framework in VolumeDeform. Thus, the result from this work has better texture.

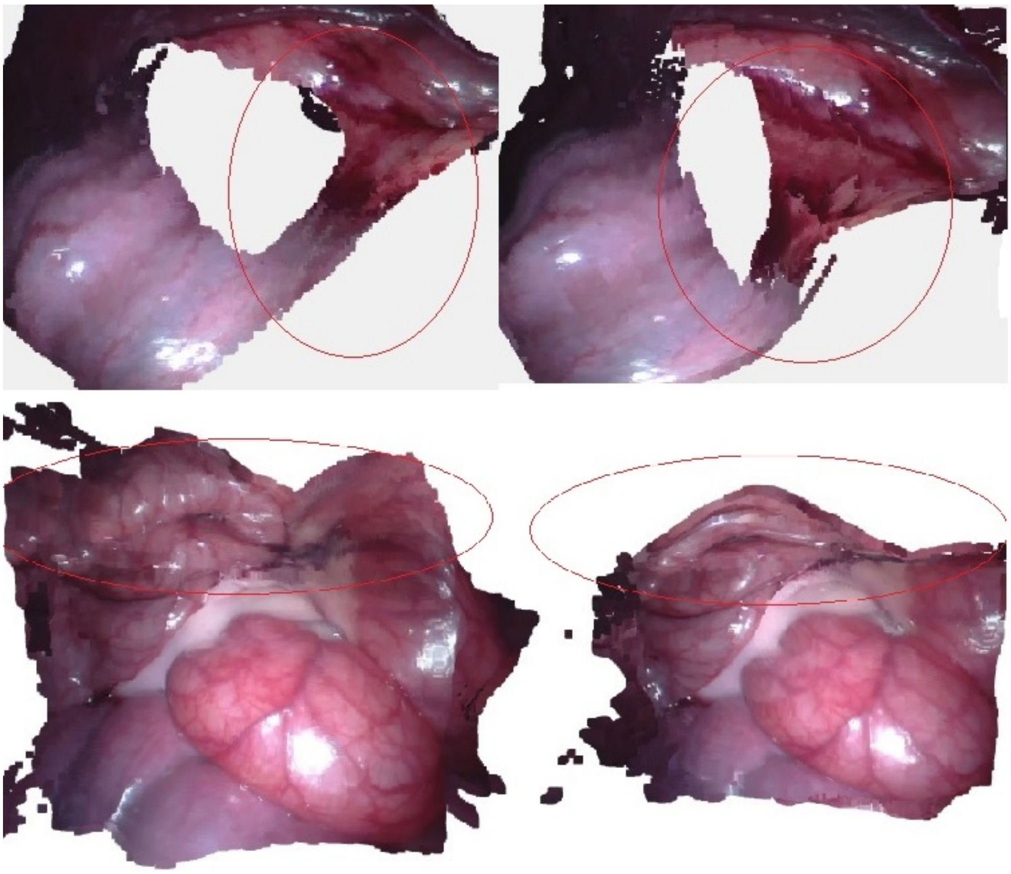
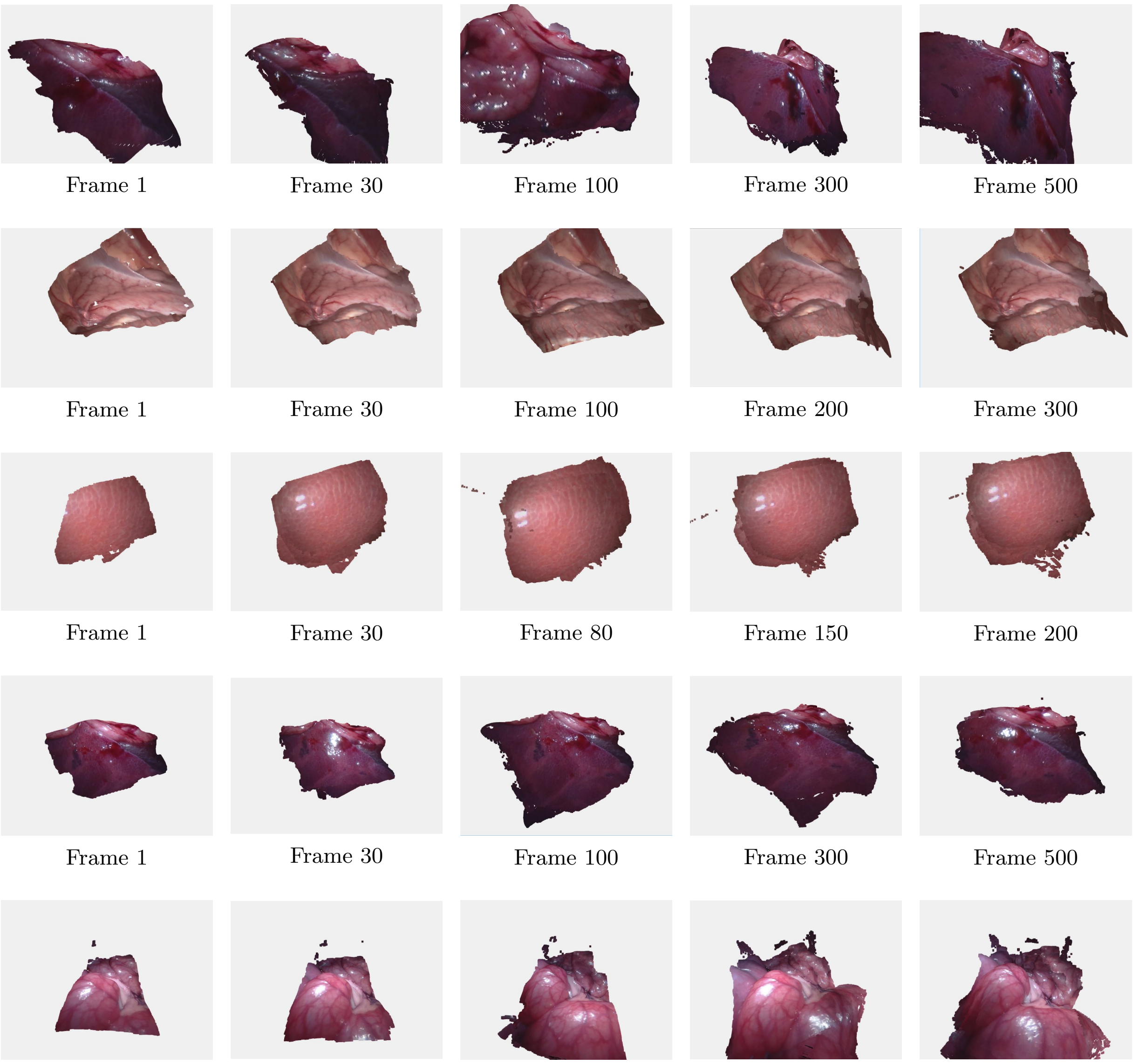


Figure 3.9: The comparison between pipeline with and without dense SURF constraint (Left is with constraint while right is without). Significant errors happen either in texture or in topologies without SURF constraint.

Discussion and limitations. Different from Newcombe et al. [12], Innmann et al. [13] and Dou et al. [14] which use 3D volume named TSDF as model management tool and extract mesh (structured 3D surface with vertices and triangle faces) for estimating ED graph, this research directly acquires dense point cloud as data management. The difference in data management affects this research pipeline fundamentally. First, previous approaches apply marching cube to extract mesh from the 3D volume at each frame due to the requirement of estimating the visible points in the ED graph. While, the point cloud data management proposed in this research does not require any marching cube points extraction process, and efficient real-time live model rendering can be easily achieved. Besides, data management in this research pipeline enables the model to move freely without the predefined range in volume based method. As a matter of fact, clinically it is annoying or even impossible for surgeons to predefine the volume range. In this way, surgeons can



Frame 1 Frame 30 Frame 100 Frame 300 Frame 500

Frame 1 Frame 30 Frame 100 Frame 200 Frame 300

Frame 1 Frame 30 Frame 80 Frame 150 Frame 200

Frame 1 Frame 30 Frame 100 Frame 300 Frame 500

Frame 1 Frame 30 Frame 100 Frame 300 Frame 500

Figure 3.10: Non-rigid reconstruction of different soft tissues using in-vivo datasets. Illustrated are the sequences of 3D reconstructions. The five videos are (from top to bottom): abdomen (1), abdomen wall, liver, abdomen (2) and abdomen (3).

perform the reconstruction at any time without pre-requisite range and grid size setting steps. Another important benefit is that with point cloud data management, the estimation of the ED graph at the current frame is directly from the previous frame, instead of from the initial model to the current frame as in DynamicFusion [12] and VolumeDeform [13]. A canonical model, or model in the initial frame, is not required at all in the proposed framework. Using the method proposed in this thesis, the latest observation can be registered consecutively to the previous model instead of the initial model, which results in a more accurate estimation of the ED graph. In this way, the proposed method relatively avoids drift issues in the model-to-frame matching in slow motion. After extracting dense key points, they are lifted into a 3D coordinate. In fact, Dou et al. [14] periodically resets the entire volume to handle large misalignments caused by deforming the initial model to the last state that cannot be recovered from the per-voxel refresh.

All results demonstrate that, with a correct ED graph, the proposed method can perform almost the same fusion process as TSDF, but do not require a predetermined volume. Weighted points based method offers a variety of benefits: (1) Points based data management and fusion free the geometry extent while still maintain the ability to beget a smooth fused surface. (2) By using the points based management, all the components in the proposed framework, e.g. visible points prediction, ED graph estimation and model update are unified in points. The process like conversion between volume and mesh is not necessary anymore. (3) The proposed approach enables the current live model updated from the model in the previous step instead of from the reference model which prevents drifts accumulated in the ED graph.

In the video on youtube [2], there are some illumination differences in the texture and the rapid fluctuations on the edges of the reconstructed model. The texture difference is due to the different angle of the light source in different image times. As this work tries to preserve the latest texture, the texture is updated directly instead of implementing the weighted average process. The rapid fluctuations on the edges result from the quick wave of depth generated from ELAS. The proposed pipeline deforms the reconstructed model to match the depth and fluctuations on the depth's edge forces model deform accordingly. This can be solved by developing a more robust depth estimation algorithm or filter the edges of the depth model.

While this research pipeline works well on the test datasets, the challenges facing reconstruction problems using a stereoscope should also be addressed. The first and most important challenge is the fast movement of the scope. The proposed algorithm fails to track motion when the scope moving fast. Similar to traditional SLAM approaches [16]

[2] https://youtu.be/QL1uUHJDZ1E

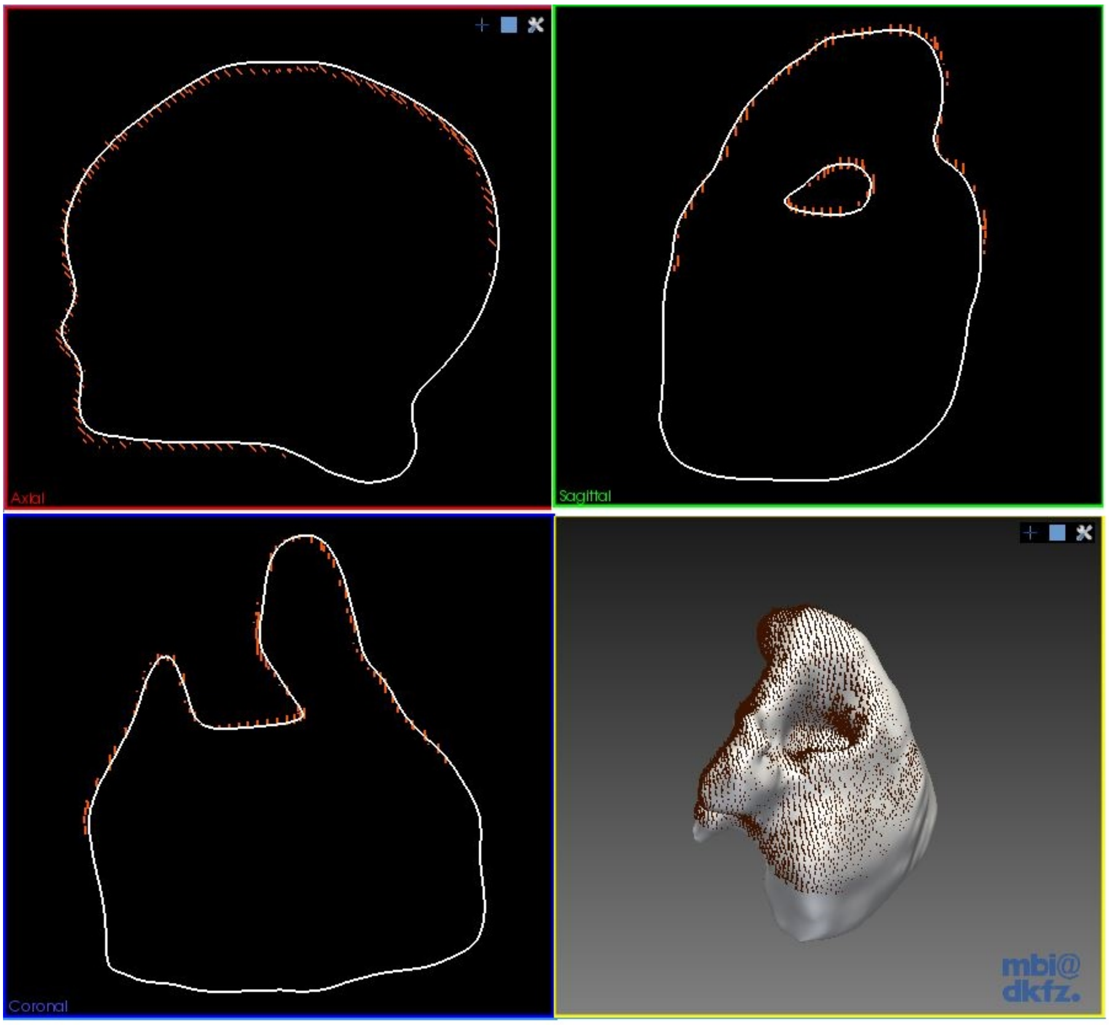


FIGURE 3.11: The Axial, Coronal, Sagittal and 3D views of the deformed model and ground truth at the last frame (heart). The red points denote the scan of the last frame.

[17], serious consequences of fast motion are the blurry images and relevant disorder of depths. These phenomenons happen especially when the latest constructed model deforms to match the depth with false edges suffering from image blurring. That's why this research visualizes periodic deformation like respiration and heartbeat clearly on central regions but shows obvious drifts on the edges. Fast motion is a challenging issue as the only data source is the blurry images. Another issue is the accuracy of texture. The laparoscope with a narrow field of view results in obvious drifts and gaps on texture especially in blurry images.

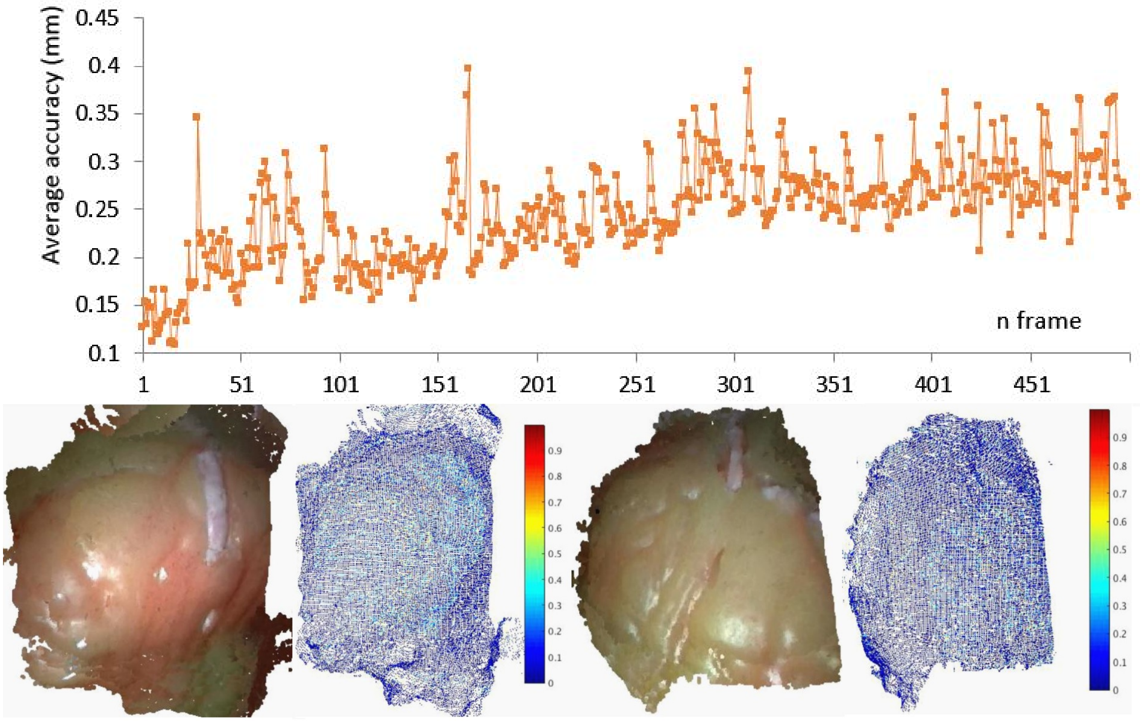


FIGURE 3.12: Ex-vivo validation with the two Hamlyn validation datasets: Silicon heart phantoms deforming with cardiac motion and associated CT scans. The upper figure is the time series of average error. The lower figures are the reconstructed geometry and corresponding error maps measured by distance to ground truth.

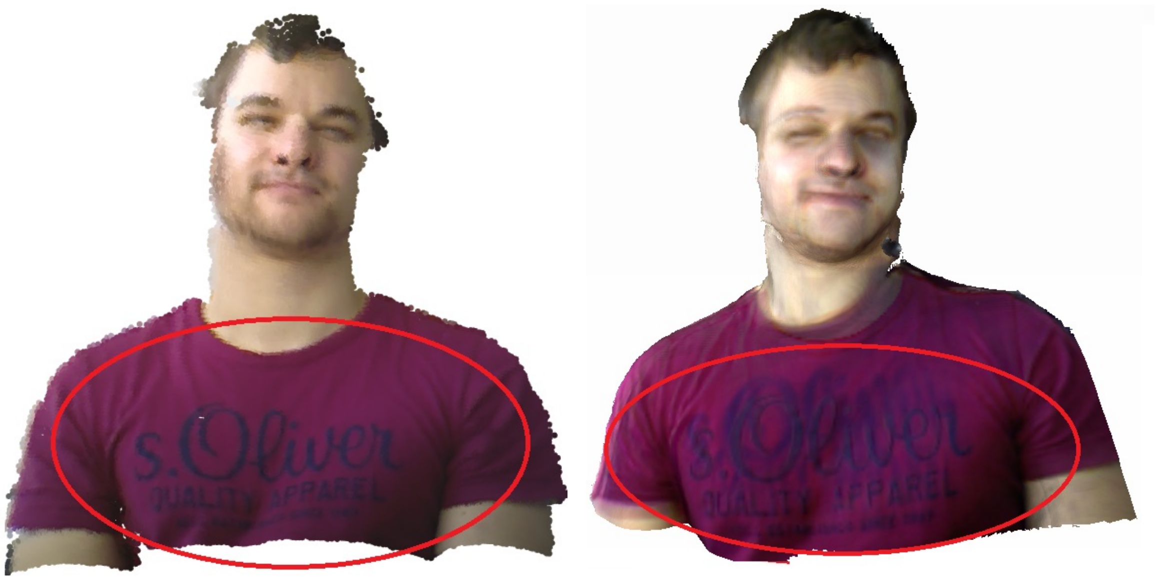


FIGURE 3.13: Comparison with VolumeDeform approach. Left is the result of the robocentric template free SLAM. The right is VolumeDeform's result. Note the difference in texture (letters on the T-shirt).

3.5 Chapter summary

In this chapter, aiming at describing deformation in SLAM of non-rigid environments, the ED deformation graph is introduced and tested on two cases: robocentric template based SLAM and robocentric template free SLAM. In the template based SLAM scenario, a deformation recovery framework for the 3D reconstruction of the deformable soft-tissue is proposed in the scenario of MIS based on the pre-operative CT data and real-time depth sensing. The DFF is proposed for robust, efficient and accurate optimization and the model-to-scan registration and model deformation can be solved simultaneously in the proposed framework. Simulation results using three public available soft-tissue models show that the deformations are recovered accurately using the proposed algorithm with very good convergence, which is promising for real-time implementation. Accuracy analysis shows that soft-tissue shape in the previous step can be efficiently deformed to fit the partially observed scan in the current step by using the proposed formulation. And the results from the simulated sequential deformation of three different soft-tissues demonstrate the potential clinical value for MIS. However, there are hardly any publicly commercialized direct medical apparatus like time of flight sensor or structured light sensor, which greatly limits the direct application of the template based SLAM approach. Stereoscope can be a solution to these direct 3D scopes.

The template free approach proposes a dynamic deformation recovery SLAM framework for reconstructing the 3D shape of deformable soft-tissues in the scenario of MIS based stereoscope. In contrast to conventional non-rigid scene reconstruction, the template free SLAM replaces the current volume based approach with point cloud and adjusts the fusion process for the purpose of the relatively large spatial requirement. Simulation and in-vivo experiments validate the feasibility of this research dynamic SLAM framework. Next chapter will focuses on exploring a more robust key points extraction algorithm for enhancing robustness in a situation when scope moves fast.

Chapter 4

MIS-SLAM: A complete robocentric SLAM system for MIS scenario

Based on ED deformation graph, this chapter aims at enabling large scale robocentric SLAM system in MIS. It solves the problem of system failure of scope fast movement as well as large scale system performance efficiency. Comparing with conventional SLAM, MIS brings shortcomings such as lack of field of view, poor localization of scope, fewer surrounding information and fast scope motion with regard to the surface. As mentioned in Chapter 3, the first and most important challenge to the pipeline is the fast movement of the scope. Fast motion not only makes visual odometry unstable but also causes blurry images and worse registrations. This issue has also been reported in [16] and [17].

Therefore, even Section 3.3 proposes a robocentric template free SLAM, a dynamic reconstruction system of deformable soft-tissue with the stereoscope with a warping field based on the ED graph, the robustness problem is still a big issue to be solved. Besides, more strategies need to be carried out with the help of GPGPU for real-time implementation.

Inspired by the researches Mahmoud et al. [39], Mahmoud et al. [40] and Turan et al. [41] who demonstrate the robustness of scope pose estimation from ORB-SLAM, ORB-SLAM may have the potential to be improved and coupled with dense deformable SLAM. This

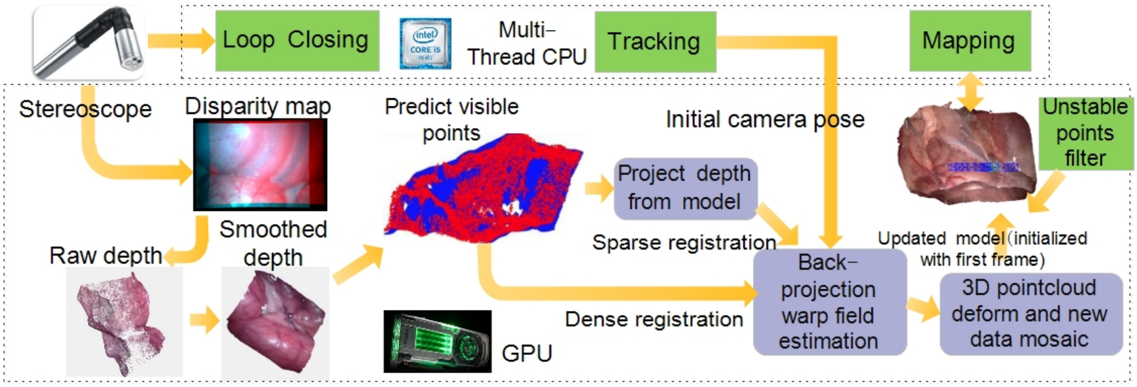


Figure 4.1: The framework of MIS-SLAM. CPU is responsible for ORB-SLAM, uploading features, rigid and start a visualization module. GPU processes depth estimation, registration, fusion and visualization.

chapter proposes MIS-SLAM based on the preliminary template free SLAM proposed in Section 3.3 with the following major improvements: (1) Propose a heterogeneous framework to make full use of both GPU (dense robocentric deformable SLAM) and CPU (ORB-SLAM) to recover the dense deformed 3D structure of the soft-tissues in MIS scenario. The computational power of the CPU is fully exploited to run an improved ORB-SLAM to provide complementary information to GPU modules. (2) Modules from GPU and CPU are deeply integrated to boost performance and enhance the efficiency. Sparse ORB features, as well as rigid transformation, are uploaded to GPU. (3) Upgrade former model point storage system and fusion management strategy to enhance large scale soft-tissue reconstructing. Comparing with TSDF widely used in computer vision community [12] [13] and [14], point cloud management in MIS-SLAM notably reduces memory as well as boosts the performance. (4) Real-time visualization is achieved on the GPU end. MIS-SLAM can process large scale surface reconstruction in just one desktop in real-time. Associate videos [1] are on youtube to fully appreciate the live capabilities of the system.

4.1 Overview of MIS-SLAM

The architecture can be divided as **initial tracking** and **deformable tracking and dense mapping**. The initial tracking is achieved with an improved ORB-SLAM algorithm on the CPU end. Deformable tracking and dense mapping is implemented on GPU end.

[1] MIS-SLAM: https://youtu.be/2pXokldQBWM

In the initial tracking step, ORB-SLAM is first launched on CPU; ORB features and global rigid transformation are uploaded from CPU to GPU global memory. This initial global rigid transformation significantly increases the robustness of the system.

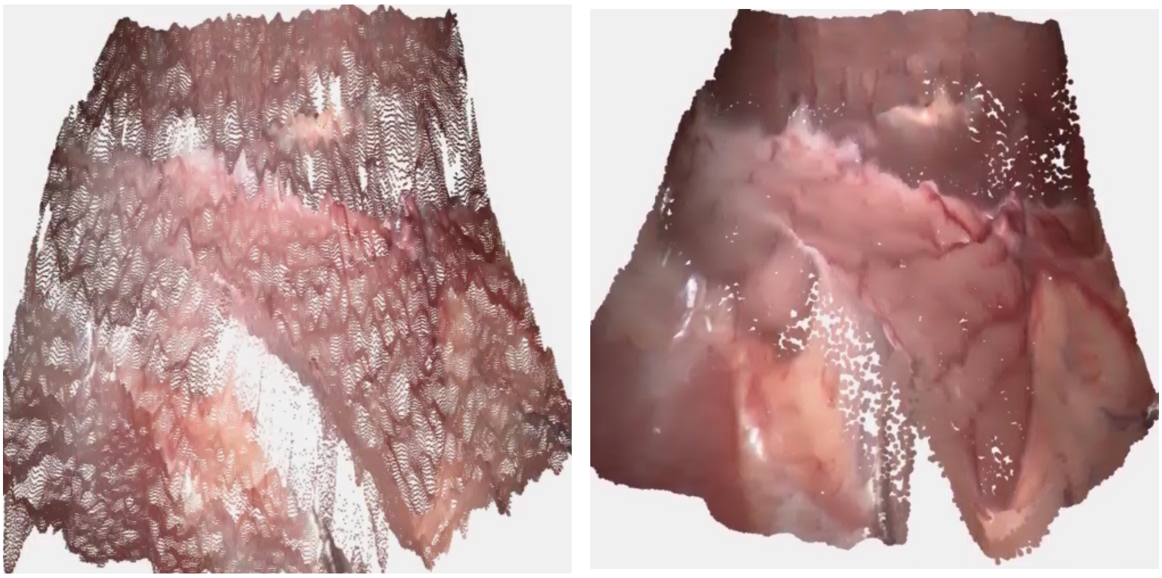
In the deformable tracking and dense mapping step, after receiving initial rigid transformation from CPU end, it first initializes the model with the first estimated depth. Each time when a new observation is acquired, the matched ORB features are uploaded to GPU. Potential visible points are extracted from the model and projected on 2D depth images. A registration process is performed to estimate optimal rigid transformation as well as the non-rigid warping field. The live model is then deformed to current shape according to this transformation and fused with the new observation. This research makes use of the feature called 'Graphic Interoperability' in Compute Unified Device Architecture (CUDA) to directly visualize model on GPU end. Fig. 4.1 demonstrates the pipeline of these processes.

Realizing the point cloud generated from stereo images are much less reliable than depth perception sensors, this research modifies and update the robocentric template free SLAM (Section 3.3) with more properties. Each point stores six domains: coordinate $\mathbf{v}_i$, normal $\mathbf{n}_i$, weight ω_i, color $\mathbf{C}_i$, time stamp t_i and a boolean variable stability $\mathbf{s}_i$. Visible points selection approach is updated to have better model to depth registration (Algorithm 1). This research adds t_i, $\mathbf{n}_i$ and $\mathbf{s}_i$ and introduces model filtering technique to have smooth model with less noisy points (Algorithm 2 and 3).

4.2 Depth estimation, sparse key correspondences and global rigid transformation

Similar to Section 3.3.2, this work adopts ELAS [89] as the depth estimation method. Originally designed to map large scale scenario in near real time, ELAS has also been proved to achieve a good result in surgical vision [28]. Therefore, this research applies ELAS as the module for providing initial colored depth for soft-tissues from stereo images. Fig. 4.2 shows the example of the original depth and smoothed depth.

The main issue in Section 3.3 is the inaccuracy of relative scope to soft-tissue pose leading to instability of the pipeline. The deformation graph based approach is a typical model-to-frame visual odometry process lacking additional mechanics to ensure global pose tracking robustness. Without initialization, dense mapping inevitably suffers from drift or lose track. To improve the robustness of the system, the idled CPU is fully exploited to run ORB-SLAM for providing good initial pose (equivalent to rigid soft-tissue transformation) for enhancing robustness. ORB-SLAM module provides the ORB features which are fully exploited on GPU. Comparing to SLAM system proposed in Section 3.3, this strategy saves computational powers on GPU: (1) Dense SURF extraction and matching step in original approach [19] is therefore not needed as matched ORB features are uploaded. (2) Visual Odometry and Random sample consensus (RANSAC) on GPU end in [19] is replaced with rigid transformation and ORB features from ORB-SLAM on CPU end.



Original depth

Smoothed depth

Figure 4.2: Examples of depth and smoothed depth.

4.3 Deformation field estimation

This chapter follows and extends the general ED deformation graph formulation described in Section 3.1. The first two constraints E_{rot} and E_{reg} follow strictly Eq. (3.5) and Eq. (3.7). E_{data} and E_{corr} follow the Eq. (3.11) and Eq. (3.13) with point to plane distances.

Two new terms, E_r and E_p, are added to ensure robustness of global rigid transformation. Overall, the objective function formulated is composed of six terms: Rotation, regularization, the point to plane distances between the visible points and the target scan, sparse key points correspondence and global rigid transformation (new terms) as:

$$\underset{\mathbf{R}_c,\mathbf{T}_c,\mathbf{A}_1,\mathbf{t}_1\ldots\mathbf{A}_m,\mathbf{t}_m}{\operatorname{argmin}} \quad w_{rot}E_{rot} + w_{reg}E_{reg} + w_{data}E_{data} + w_{corr}E_{corr} + w_rE_r + w_pE_p, \tag{4.1}$$

Similar to Section 3.1, this research follows ED graph [83] to constrain deformation graph from unreasonable deformation with two constraints **Rotation** and **Regularization**. All m nodes follow the two constraints.

Data Term. This research follows Algorithm 1 to find registrations of model points and minimize point to plane distance of all the registered points. For each model point $\mathbf{v}_i$, if it is registered to depth, it is assumed to be a visible point. In Algorithm 1, ϵ_{dv} and ϵ_{nv} are thresholds for extracting visible points based on distance and angle.

Similar to Section 3.3.2, after extracting registered visible points, this work adopts back-projection approach as a model-to-scan registration strategy that penalizes misalignment of the predicted visible points $\mathbf{v}_i$ $(i \in \{1, ..., N\})$ and current depth scan $\mathbb{D}$. As described in Fusion4D [14], back-projection and point to plane strategies make full use of the input depth image, so that the Jacobians can be calculated in regularized 2D space which leads to fast convergence and robustness to outliers.

Correspondence. Similar to Section 3.3, this work also utilizes RGB information for enhancing robustness. It first tracks frame-to-frame feature points and minimizes the Euclidean distance between pair-wise sparse key points generated from features described in Section 3.3.2 in the following form. In this chapter, Dense SURF (Section 3.3.2) is substituted with ORB features uploaded from ORB-SLAM.

Global rigid transformation. The new term is added with regard to previous formulation [19]. It is measured by the variations of rotation and transformation. First frame is fixed as the coordinate origin. This work uses Euclidean distance and Euler angles to define the difference between optimized global rigid transformation (rotation $\mathbf{R}_c^n$ and translation $\mathbf{T}_c^n$) and global rigid transformation (rotation $\tilde{\mathbf{R}}_c^n$ and translation $\tilde{\mathbf{T}}_c^n$) provided from ORB-SLAM. It is presented in the following form:

$$E_r = ||\tilde{\mathbf{R}}_c^n - \mathbf{R}_c^n|| \qquad E_p = ||\tilde{\mathbf{T}}_c^n - \mathbf{T}_c^n|| \tag{4.2}$$

Algorithm 1 is adopted to find visible point set $\mathbb{V}$ for optimization. This chapter follows previous strategy (Section 3.3.2) using Levenberg-Marquardt to solve the nonlinear optimization problem. The efficiency is almost the same as Section 3.3.2 because only 6 more variables (Global rotation and translation) are added.

Algorithm 1: Model points to depth image registration

```
Input: Model $\mathbb{P}^{n-1}$ in last frame
       Latest observed depth $\mathbb{D}^n$
       Distance threshold of two points $\epsilon_{dv}$
       Normal angle threshold of two normals $\epsilon_{nv}$
Output: Visible points set $\mathbb{V}^n$ regarding to depth $\mathbb{D}^n$
foreach Model point $\mathbf{v}_i$ do
    if $\mathbb{D}(P(\mathbf{v}_i)) \neq null$ then
        if $(\|\mathbf{v}_i - \Gamma(P(\mathbf{v}_i))\| < \epsilon_{dv}$ and $\mathbf{n}_i \cdot \mathrm{H}(P(\mathbf{v}_i)) > cos(\epsilon_{nv}))$
         then
            Add $\mathbf{v}_i$ to $\mathbb{V}^n$
        end
    end
end
```

4.4 Model update with new observation

Inspired by Keller et al. [49], new properties (normal, time step and stability) are introduced to point management. Model is fused with depth following Algorithm 2. Then Algorithm 3 is used to remove noisy model points.

The basic idea of Algorithm 2 is building three different groups of the point cloud. The original model is classified into registered (Group 1) and unregistered (Group 2) with regard to the depth image. Points in Group 1 are fused with depth image. After which pixel from the depth image that's not registered with model points are lifted and initialized as new observations (Group 3). All three groups are merged to compose the new model.

In Algorithm 3, 'stability $\mathbf{s}_i$' is applied to filter model points influenced by noisy depth. Unstable model point is defined as a point with low weight (only seen in few times) which has not been observed for several recent frames. This point is likely a noisy point resulting from inaccurate depth estimation. Algorithm 3 shows how to filter the noisy points.

Algorithm 2: Fusion of Point cloud with depth image

```
Input: Model $\mathbb{P}^{n-1}$ in last frame and current depth $\mathbb{D}^n$
       Distance and normal thresholds $\epsilon_{df}$ and $\epsilon_{nf}$
Output: Fused model set $\mathbb{P}^n$
Step 1: Register and fuse model with depth (Group 1), the rest model are
 unregistered points (Group 2)
foreach $\mathbf{v}_i \in \mathbb{P}^{n-1}$ do
    Deform $\mathbf{v}_i$ to $\tilde{\mathbf{v}}_i$
    if $\mathbb{D}(P(\tilde{\mathbf{v}}_i)) \neq null$ and
      $\|\tilde{\mathbf{v}}_i - \Gamma(P(\tilde{\mathbf{v}}_i))\| < \epsilon_{df}$ and
      $\mathbf{n}_i \cdot \mathrm{H}(P(\tilde{\mathbf{v}}_i)) > cos(\epsilon_{nf})$ then
        Fuse $\tilde{\mathbf{v}}_i$ following Eq. (4.3, 4.4, 4.5) and Eq. (4.6).
        Push fused $\tilde{\mathbf{v}}_i$ Group 1
    else
        Push $\tilde{\mathbf{v}}_i$ to Group 2
    end
end
Step 2: Add newly observed points (Group 3)
foreach $u_k \in \mathbb{D}^n$ do
    if $u_k$ is not fused in Step 1 then
        Lift $u_k$ into 3D space (position ($\mathbf{v}_i$), normal($\mathbf{n}_i$), color $\mathbf{C}_i$
        Initialize color, $\omega_i = 1$, time stamp $t_i = i + 1$, stability $\mathbf{s}_i$= False. and pushed
         into Group 3
    end
end
Step 3: Fuse different types of points
Merge Group 1 Group 2 Group 3 to new model $\mathbb{P}^n$.
```

For a single point $\mathbf{v}_i^n$ in nth step, fusion with new depth is achieved by:

$$\tilde{\mathbf{v}}_i^{n+1}|_z = \frac{\tilde{\mathbf{v}}_i^n|_z * \omega_i^n + \mathbb{D}^{n+1}(P(\tilde{\mathbf{v}}_i^n))}{\omega_i^n + 1} \tag{4.3}$$

$$\mathbf{C}_i^{n+1} = \frac{\mathbf{C}_i^n * \omega_i^n + \mathbb{C}^{n+1}(P(\tilde{\mathbf{v}}_i^n))}{\omega_i^n + 1} \tag{4.4}$$

Algorithm 3: Removing noisy unstable model points

```
Input: Fused model set $\mathbb{P}^n$
        Time and weight thresholds $\tau_{time}$ and $\tau_{weight}$
Output: Filtered model set $\mathbb{P}^{'n}$
        New node positions $\mathbf{g}$
foreach $\mathbf{v}_i \in \mathbb{P}^n$ do
    if $t_i < (i - \tau_{time})$ and $\omega_i < \tau_{weight}$ and $\mathbf{s}_i$= False then
        Delete $\mathbf{v}_i$
    else
        Stamp $t_i = i + 1$
        if $t_i \geq (i - \tau_{time})$ and $\omega_i \geq \tau_{weight}$ then
            $\mathbf{s}_i$= True
        end
    end
end
Regenerate new nodes $\mathbf{g}$ and initialize rotation $\mathbf{A}$ as identity matrix and translation $\mathbf{t}$
 as zero vector.
```

$$\tilde{\mathbf{n}}_i^{n+1} = \frac{\tilde{\mathbf{n}}_i^n \omega_i^n + \mathbb{N}^{n+1}(P(\tilde{\mathbf{v}}_i^n))}{\omega_i^n + 1} \tag{4.5}$$

$$\omega_i^{n+1} = min(\omega_i^n + 1, \omega_{max}) \tag{4.6}$$

where $\tilde{\mathbf{v}}_i^n|_z$ is the value of deformed point $\tilde{\mathbf{v}}_i^n$ in the z direction. $\tilde{\mathbf{n}}_i^n$ is the deformed normal of $\mathbf{n}_i^n$. $\mathbb{D}^n$, $\mathbb{C}^n$ and $\mathbb{N}^n$ are depth map, color map and normal map in step n respectively. ω_{max} is the maximum weight for each point. Different from rigid transformation where uncertainty of all the points in 3D space are considered as equal, in the case of non-rigid fusion, if a point is further away to the nodes of warping field, it is less likely to be the registered depth [12]. Therefore, this research practically measures this certainty by using the minimum distance from point to nodes and regularizes it with half of the unified node distance. Algorithm 2 and Eq. (4.3,4.4,4.5) and Eq. (4.6) show the details for point fusion.

The improved weighted points based method brings many benefits: Point based data management is free of extent limitation; with fusion based Algorithm 2 and noise point filter approach Algorithm 3, fused geometry can still keep its shape smooth while avoiding noisy input; the reconstructed geometry preserves more vivid texture and details.

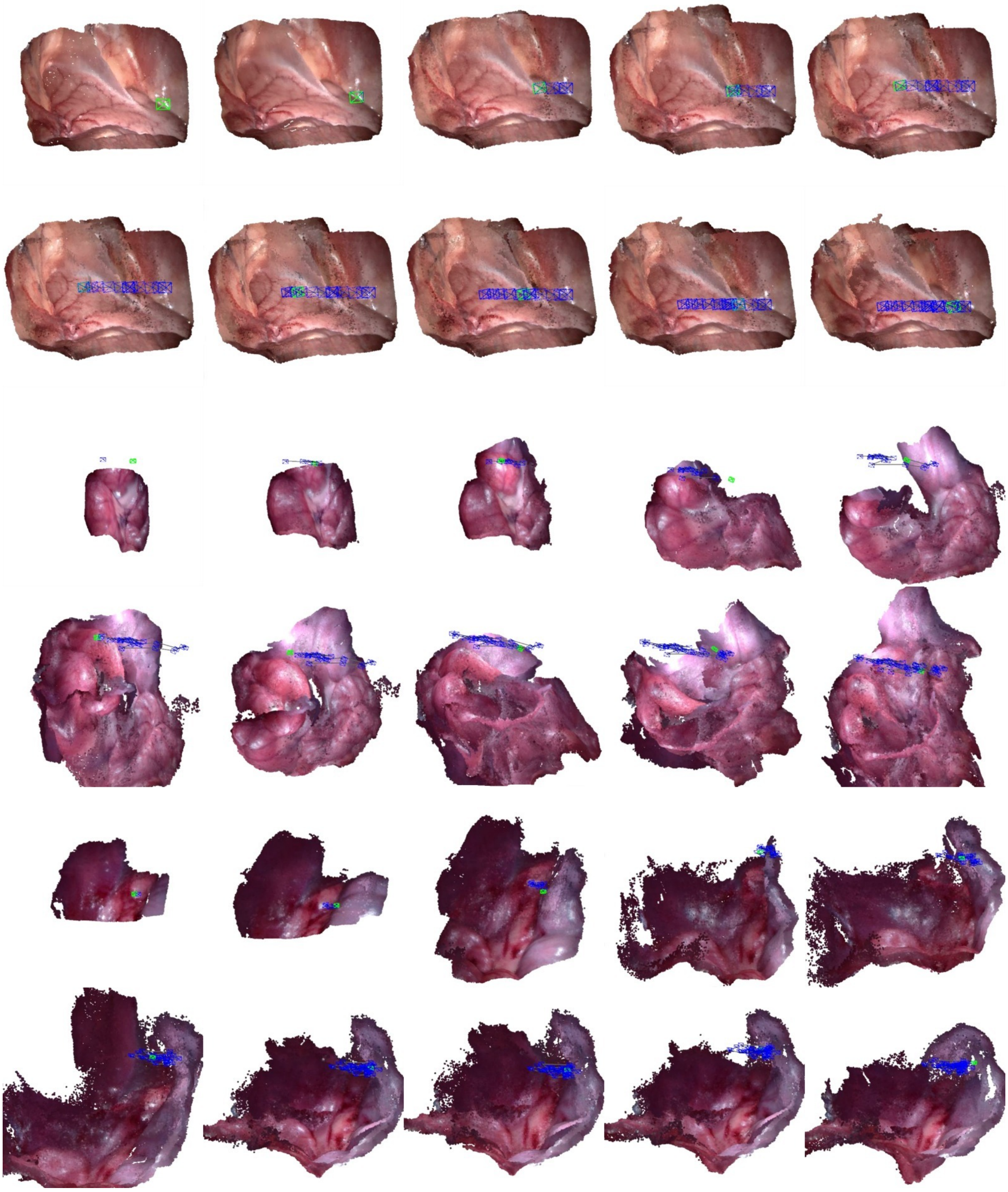


FIGURE 4.3: MIS-SLAM processes 3 in-vivo datasets. Figures present the whole constructed model at different frames. The three videos are (from top to bottom): Abdomen wall (1), abdomen (2) and abdomen example (3).

4.5 Results and discussion

This research first validates MIS-SLAM on publicly available in-vivo stereo video datasets provided by the Hamlyn Centre for Robotic Surgery [91]; deformations are caused by respiration and tissue-tool interactions. MIS-SLAM is also validated on ex-vivo phantoms and some simulations, and compared with ground truth. When testing on in-vivo validation, three videos with deformation and rigid scope movement are utilized. Other videos either have no deformation or no scope motion. Please note that no extra sensing data other than stereo videos from the scope is used in the proposed algorithm. The frame rate and image size of the in-vivo porcine dataset (model 1 in Fig. 4.3) are 30 frames per second and 640 × 480 while other two datasets are 25 frames per second and 720 × 288. Distance from scope to the surface of soft-tissue is between 40 to 70 mm. In last chapter (Section 3.3.2), due to poor quality of obtained images and some extremely fast movement of scope, videos tested on porcine with fast or abrupt motion cannot generate good results. In this chapter, however, the proposed MIS-SLAM can process large scale environment with much better robustness.

The open source ORB-SLAM is executed on desktop PC with Intel Core i7-4770K CPU @ 3.5 GHz and 8GB RAM. This chapter follows Mahmoud et al. [39] to tune the parameters and structures. The average tracking time is 15ms with 640x480 image resolution and 12 ms with 720x288 image resolution. As the frame rate of the three datasets is 25 or 30 frames per second, ORB-SLAM can achieve real-time tracking and sparse mapping. By parallelizing the proposed methods for GPGPU, MIS-SLAM algorithm is implemented in CUDA with the hardware 'Nvidia GeForce GTX TITAN X'.

For model 1, the point cloud density is set to 0.2 mm and node density is set to 4 mm. For model 2 and 3, the point cloud density is set to 1 mm and node density is set to 10 mm. Point cloud downsampling process is achieved by setting a fixed box to average points fill inside each 3D box. The parameters for optimization are chosen as $w_{rot} = 1000, w_{reg} = 10000, w_{data} = 1, w_{corr} = 10, w_{corr} = 1, w_r = 1000000, w_p = 1000$. Thresholds are set to extract predicted visible points with a point to plane distance ϵ_{dv} as 15 mm and angle threshold ϵ_{nv} as 10°. The accuracy is measured by subtracting projected model image and the observed depth image. The maximum weight is set to 20 and time stamp threshold is

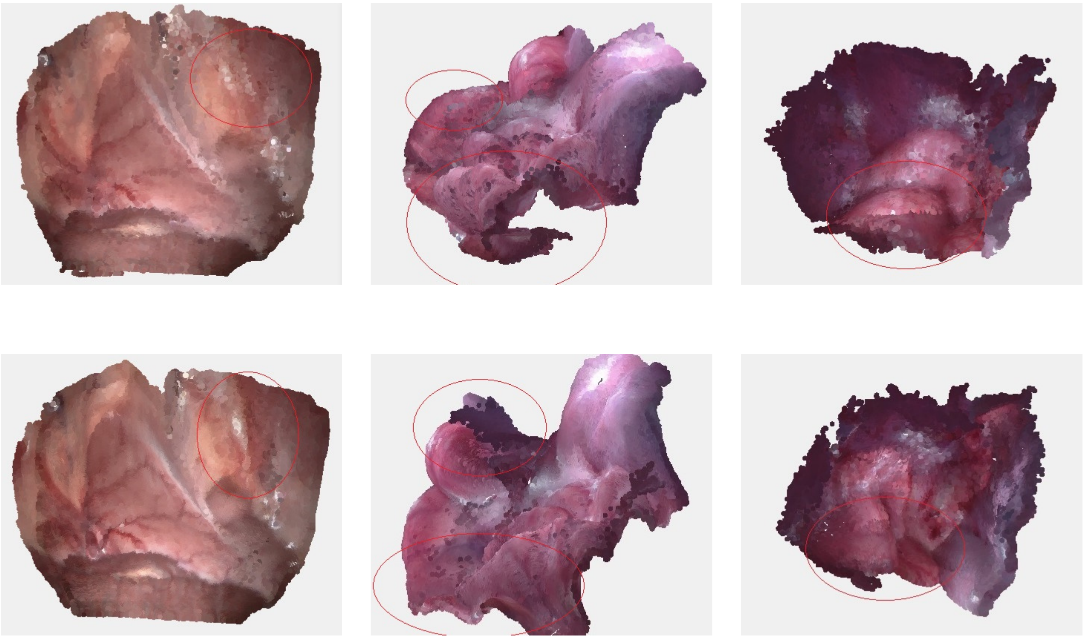


Figure 4.4: Comparisons between Section 3.3.2 (First row) and the proposed MIS-SLAM (Second row).

set to 10. Thresholds ϵ_{nf} and ϵ_{df} for point to depth registration is set as 10 degrees and 10 mm (20 mm for model 2/3). Truncated distance is set as 40 mm (60 mm for model 2/3). All the quantitative accuracy comparison is carried out with Eq. (3.17).

4.5.1 Robustness enhancement

The robustness of MIS-SLAM is significantly improved when global rigid transformation from ORB-SLAM is uploaded to GPU and employed as initial scope rigid transformation. Fig. 4.4 shows the comparison between previous work (Section 3.3) and the proposed method.

One challenge facing reconstruction using stereoscope is the fast movement of scope [19]. Configuration without global rigid transformation initialization fails to track motion when scope moves fast. Like traditional SLAM approaches, severe consequences of fast motion are the blurry images and relevant disorder of depths. These phenomenons happen especially when the latest constructed model deforms to match the depth with false edges suffering from image blurring. Localizing in fast motion is very challenging because the

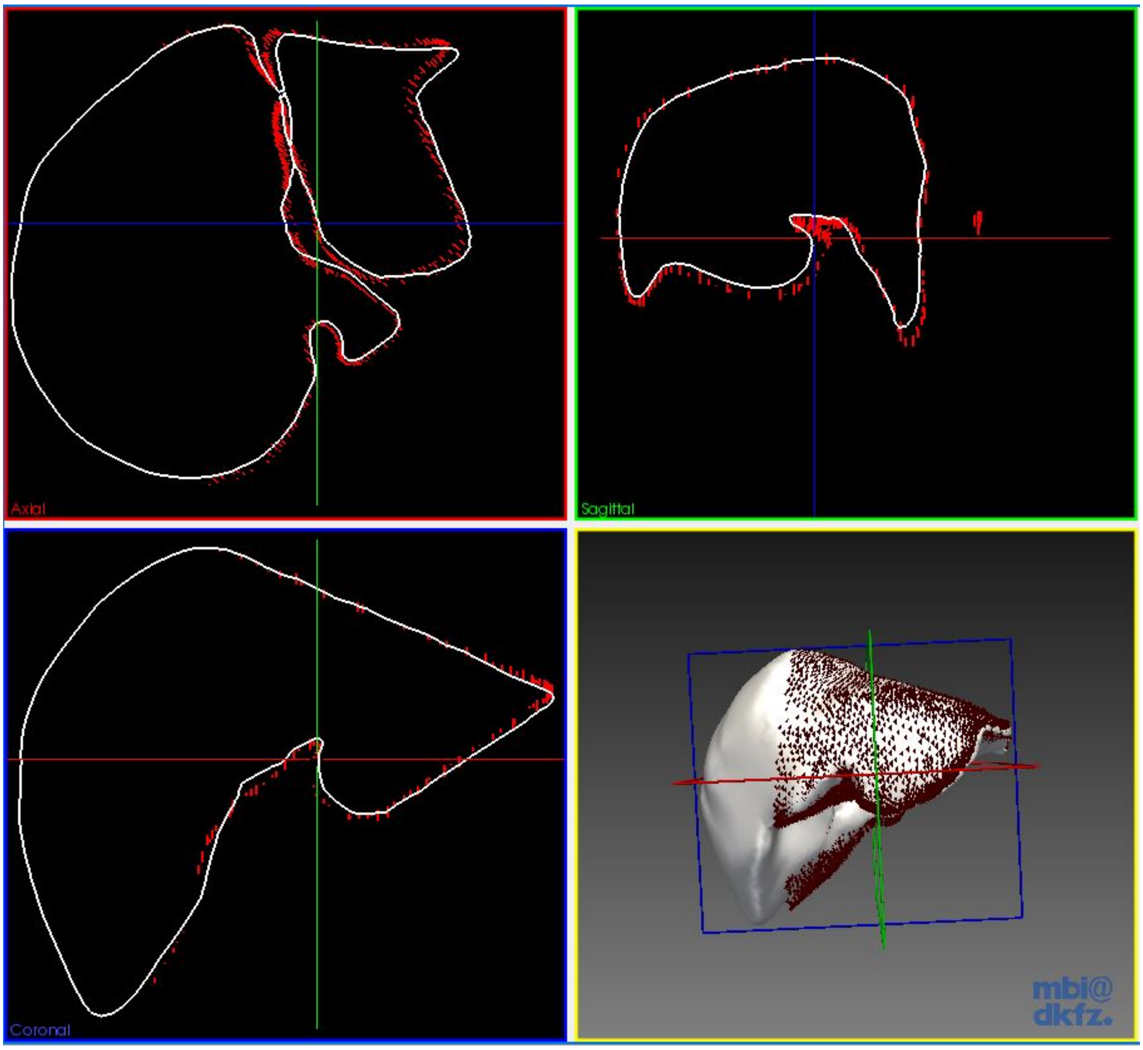


FIGURE 4.5: The Axial, Coronal, Sagittal and 3D views of the deformed model and ground truth at the last frame (liver). The red points denote the scan of the last frame.

only information for positioning is the blurry images. ORB-SLAM, however, is a robust feature based system even works in deformable surgery scenario [39] [40] [41]. Though based on prior to a stationary environment, it still relatively keeps the global pose. The supplementary video [2] clearly shows how initialization of global rigid transformation prevents the system from failing to track scope rigid transformation.

[2] MIS-SLAM: https://youtu.be/2pXokldQBWM

4.5.2 Deforming the model and fusing new depth

A threshold is employed to discard some frames which have a large error due to low-quality depth generated from blurry images. Different from previous research, with good initialization of depth image, MIS-SLAM is robust against losing track. Fig. 4.3 shows the results of soft-tissue reconstruction of MIS-SLAM in different frames, using 3 in-vivo laparoscope datasets [91]. The results demonstrate that the soft-tissues are reconstructed incrementally with texture.

We test the difference with average Euclidean distance of the recovered shape with the target dpeht defined in Eq. (3.17). The average distances of back-projection registration of the three simulation scenarios are 0.18 mm (1), 0.13 mm (2) and 0.12 mm (3). The test on datasets with ground truth (Hamlyn dataset 10/11) achieves 0.08 mm, 0.21 mm (Average errors).

4.5.3 GPU implementation and computational cost

The MIS-SLAM system is implemented with heterogeneous computing. The ORB-SLAM runs on CPU. The rest is executed on GPU. Initial global rigid transformation and ORB features are transferred to GPU for further optimization. This CPU to GPU data transferring does not require much bandwidth as the amount of data is small. CPU initializes OpenGL for visualization, but the proposed system utilizes the interoperability from Nvidia's CUDA to directly visualize the model in GPU end which saves a huge amount of data transferring. Because normally our GPU end is slower than CPU end, this research utilizes the first-in-last-out feature in the 'stack' data structure to ensure GPU always processes the latest data. Processing rate for each sample dataset is around 0.07s per frame. ORB-SLAM does feature matching on CPU end, saved computation is spent on visualization. Computation increases as the model grow and the number of nodes rises.

4.5.4 Validation using simulation and ex-vivo experiments

This work also validates the MIS-SLAM on simulation and ex-vivo experiment. In simulation validation process, three different soft-tissue models (heart, liver and right kidney) are downloaded from OpenHELP [85], which are segmented from a CT scan of a healthy, young male undergoing shock room diagnostics. The deformation of the soft-tissue is simulated by randomly exerting 2-3 mm movement on a point with respect to the status of the deformed model from the last frame [18]. This research randomly picks up points in the model as the accuracy is measured by averaging all the distances from the source points to target points. Fig. 4.5 shows the final result of the simulation presented in axial, coronal, sagittal and 3D maps figures. By initializing with rigid transformation, the overall accuracies are improved from 0.46 mm, 0.68 mm, 0.82 mm to 0.41 mm, 0.66 mm, 0.62 mm regarding to heart, liver and right kidney.

This research also tests MIS-SLAM on two ex-vivo phantom dataset from Hamlyn [91]. As the phantom deforms periodically, this research manages the whole process and compares it with the ground truth generated from CT scan. The average accuracies are 0.28 mm and 0.35 mm.

4.5.5 Limitations and discussions

One of the biggest problem in MIS-SLAM is texture blending. Results (Fig. 4.3 and video on youtube [3]) indicate that when scope moves, the brightness of visible region shows significant illumination differences from other invisible regions. Some tissues even show blurry textures. The texture blending procedure involves pixel selection and blending described in Fig. 1. If in perfect registration and identically fused, the reconstruction will only suffer from illuminations from different angles of light. This illumination problem causes a systematic difference between the two images. In MIS-SLAM, creating a clean, pleasing looking texture map in the non-rigid scenario is more difficult than a static scenario. There are many other challenges in MIS-SLAM: The number of nodes increases leading to slow optimization; the scope is very close to the tissue and the exposure differs much as it moves, resulting in visible seams in final model; image motion blurring is another problem due to the scope moves fast.

[3] MIS-SLAM: https://youtu.be/2pXokldQBWM

4.6 Chapter summary

This chapter proposes MIS-SLAM: a complete real-time large scale robocentric dense deformable SLAM system with stereoscope in MIS based on heterogeneous computing. It significantly improves the robustness by solving unstableness caused by the fast movement of scope and blurry images. Benefiting from robustness, MIS-SLAM is the first robocentric SLAM system achieving large scale scope localization and dense mapping in real-time. MIS-SLAM can potentially be useful for clinical augmented reality or virtual reality applications when the scope is moving relatively fast. Next chapter will focus on reducing the computational complexity when models grow and exploring an approach to balance textures from different illumination.

Chapter 5

Efficient two step optimization in ED based SLAM

The last chapter proposes MIS-SLAM, a real-time large scale robocentric dense deformable SLAM system in MIS based on heterogeneous computing. MIS-SLAM achieves large scale scope localizing and dense mapping in real-time, enabling localization and mapping in the medical scenario. What's more, MIS-SLAM proves that ED graph based deformation formulation is applicable in simulating the deformation of a soft-tissue. However, one major issue in ED graph is that when more geometry is observed, the number of nodes increases dramatically demanding heavy computations. This issue has not been addressed in the field of 3D human reconstruction because the target size and moving extend are predefined. In more general cases, however, as reported in Chapter 4, ED nodes optimization computation is close to $O(n^2)$. **This chapter aims at solving the problem the computation complexity in larger scale environment.** Taking ED nodes and visible vertices relations in optimization step into consideration, this chapter classifies the nodes into point relevant (PR) nodes and point irrelevant (PI) nodes and propose a two-level optimization strategy. Overall, this chapter extends ED graph research by converting the formulation into matrix form and reveal inherent sparsity in Jacobian of ED graph. The constant computation complexity of the lossy strategy should have great potential for applications in ED graph based large scale SLAM.

5.1 Efficient two step optimization

5.1.1 Matrix form of ED graph deformation

$$\underset{3m\times n}{\mathbf{M}} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_{i=\mathbb{N}(1,1)} * \big(v_1 - g_{i=\mathbb{N}(1,1)}\big) & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \omega_{i=\mathbb{N}(n,1)} * \big(v_n - g_{i=\mathbb{N}(n,1)}\big) \\ \omega_{i=\mathbb{N}(1,2)} * \big(v_1 - g_{i=\mathbb{N}(1,2)}\big) & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,2)} * \big(v_n - g_{i=\mathbb{N}(n,2)}\big) \\ \vdots & \vdots & \vdots & \vdots & \omega_{i=\mathbb{N}(n,3)} * \big(v_n - g_{i=\mathbb{N}(n,3)}\big) \\ \omega_{i=\mathbb{N}(1,3)} * \big(v_1 - g_{i=\mathbb{N}(1,3)}\big) & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_{i=\mathbb{N}(1,4)} * \big(v_1 - g_{i=\mathbb{N}(1,4)}\big) & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,4)} * \big(v_n - g_{i=\mathbb{N}(n,4)}\big) \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{5.1}$$

To fully exploit the structure, we rewrite point transformation in Eq. (3.2) into matrix form for the convenience of sparsity analysis. Let's consider a group of predefined key source points defined as $\mathbf{P} = [\mathbf{v}_1 ... \mathbf{v}_n]$ and $\tilde{\mathbf{P}} = [\tilde{\mathbf{v}}_1 ... \tilde{\mathbf{v}}_n]$ to be the key target points in the vector form. According to Eq. (3.2), each point $\mathbf{v}_i$ is deformed by its 4 neighboring nodes. Thus this chapter defines two matrix $\mathbf{M}$ (Eq. (5.1)) and $\mathbf{C}$ (Eq. (5.2)):

$$\underset{m\times n}{\mathbf{C}} = \begin{bmatrix} \vdots & & & & \\ \omega_{i=\mathbb{N}(1,1)} & \cdots & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,1)} \\ \omega_{i=\mathbb{N}(1,2)} & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,2)} \\ \vdots & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,3)} \\ \omega_{i=\mathbb{N}(1,3)} & \cdots & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ \omega_{i=\mathbb{N}(1,4)} & \cdots & \cdots & \cdots & \omega_{i=\mathbb{N}(n,4)} \\ \vdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{5.2}$$

In matrices $\mathbf{M}$ and $\mathbf{C}$, note that non-zero elements are not aligned. Each column only has 4 non-zero elements (neighboring nodes). The sum of each column in matrix $\mathbf{C}$ is 1 due to the location of each element is dependent on the topology of points to nodes. In Eq. (3.2), a source point $\mathbf{v}_i$ is transformed by its 4 neighboring nodes making 4 non-zero elements every column in $\mathbf{M}$ and $\mathbf{C}$. The sum of all weight $\omega_j(\mathbf{v}_i)$ is 1. Note that different source points have different topology, thus the location of non-zero elements are not aligned well in each column. The parameters of ED nodes $\mathbf{A}_i$ and $\mathbf{t}_i$ are arranged in the following form:

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{A}_1 & \cdots & \mathbf{A}_m \end{pmatrix} \tag{5.3}$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{t}_1 + \mathbf{g}_1 & \cdots & \mathbf{t}_m + \mathbf{g}_m \end{pmatrix}. \tag{5.4}$$

With regard to general data term E_{data} defined in Eq. (3.8), to solve geometrical model to frame registration, ‘back-projection’ formulation is proposed as a substitution to iterative closest point (ICP). They make full use of 2D depth image for fast convergence. Readers may refer to [14] [12] [83] [20]. For simplicity, this work introduces the basic source and target key point pairs described by Sumner et al. [83]. Let $\mathbf{v}_i$ and $\tilde{\mathbf{v}}_i$ be the arbitrary key source point and key target points. Normally predefined in interactive phase, these key points define how model is to be deformed like bending the head or stretching the arm. The goal is to minimize the distance of deformed point set to target point set:

$$E_{data} = \sum_{i=1}^{n} ||\mathbf{R}_c \sum_{j \in \mathbb{N}(j)} \omega_j(\mathbf{v}_i)[\mathbf{A}_j(\mathbf{v}_i - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j] + \mathbf{T}_c - \tilde{\mathbf{v}}_i||^2. \tag{5.5}$$

Then Eq. (5.5) takes the following form:

$$E_{data} = ||\mathbf{R}_{\mathrm{c}} \cdot [\mathbf{\Lambda} \cdot \mathbf{M} + \mathbf{T} \cdot \mathbf{C}] + \mathbf{T}_{\mathrm{c}} \otimes \mathbf{1} - \tilde{\mathbf{P}}||_F^2. \tag{5.6}$$

where $\otimes$ is the kronecker product. $\mathbf{1}$ is $1 \times n$ vector of ones. And $|| \cdot ||_F^2$ is the Frobenius norm. Eq. (5.6) can be written compactly in the following form according to conclusions drawn from last section:

$$E_{data} = ||\mathbf{R}_{\mathrm{c}} \cdot \begin{pmatrix} \mathbf{\Lambda} & \mathbf{T} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{M} \\ \mathbf{C} \end{pmatrix} + \mathbf{T}_{\mathrm{c}} \otimes \mathbf{1} - \tilde{\mathbf{P}}||_F^2. \tag{5.7}$$

This work defines $\mathbf{\Pi} = [\mathbf{M}^T \ \mathbf{C}^T]$ and $\mathbf{\Phi} = [\mathbf{\Lambda} \ \mathbf{T}]^T$. Therefore, Eq. (5.7) can be transformed to following formulation:

$$E_{data} = ||\mathbf{R}_{\mathrm{c}}[\mathbf{\Pi\Phi}]^T + \mathbf{T}_{\mathrm{c}} \otimes \mathbf{1} - \tilde{\mathbf{P}}||_F^2. \tag{5.8}$$

The property of Jacobian of E_{data} is determined by $\mathbf{\Pi}$.

5.1.2 Sparsity in ED graph formulation

It is natural to solve Eq. (5.8) in a batch. As the number of vertices increases, the Jacobian relating to state $\mathbf{\Phi}$ increase dramatically. Luckily, this work explores the structure of $\mathbf{\Pi}$ because only part of nodes are related to model points matching to observation. Fig. 5.1 indicates that the size of the depth image (blue points) is constant due to the limited field of view of the camera. The model keeps expanding while the target depth remains in small size. A typical ED node and target depth relationship is illustrated in Fig. 5.1(d); 2/3 of the nodes are not within range of target depth resulting no contribution to E_{data}. Fig. 5.2 shows a typical Jacobian of the cost function. In E_{data} block, shadow region indicates nodes connected to points (PR nodes) while zero block shows the nodes (PI nodes) free of any connected points. In this chapter, we make full use of the sparsity of nodes in zero blocks.

The same sparsity also applies to Eq. (5.1) and Eq. (5.2). By rearranging matrix $\mathbf{\Pi}$ from Fig. 5.2(a) to Fig. 5.2(b), this research achieves a new Jacobian with zero block. Using this new matrix, Eq. (5.8) is rewritten to following form:

$$\begin{aligned} E_{data} &= ||\mathbf{R}_{\mathrm{c}}[\mathbf{\Pi\Phi}]^T + \mathbf{T}_{\mathrm{c}} \otimes \mathbf{1} - \tilde{\mathbf{P}}||_F^2 \\ &= ||\mathbf{R}_{\mathrm{c}}[\begin{pmatrix} \mathbf{\Pi}' & \mathbb{O} \end{pmatrix} \begin{pmatrix} \mathbf{\Phi}_1 \\ \mathbf{\Phi}_2 \end{pmatrix}]^T + \mathbf{T}_{\mathrm{c}} \otimes \mathbf{1} - \tilde{\mathbf{P}}||_F^2, \end{aligned} \tag{5.9}$$

where $\mathbf{\Phi}_1$ is the $\mathbf{A}_j$ and $\mathbf{t}_j$ of PR node set and $\mathbf{\Phi}_2$ is the $\mathbf{A}_j$ and $\mathbf{t}_j$ of PI node set. $\mathbf{\Pi}'$ is the subset of $\mathbf{\Pi}$ relating to PR nodes (shadow region in Fig. 5.1).

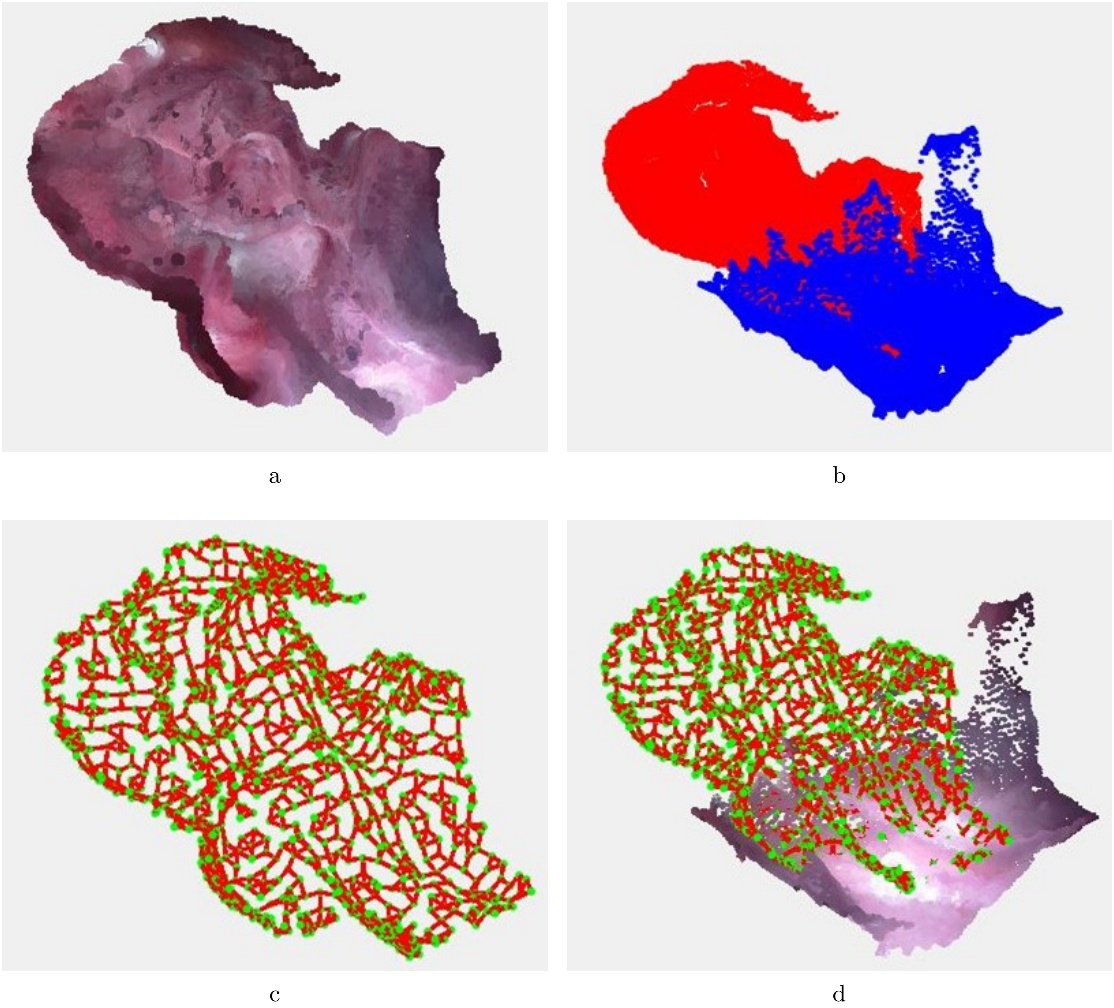


FIGURE 5.1: Illustrated are the spatial relations of the visible points and the node graph. (a) is the latest reconstruction. (b) shows both the model (red) and the target depth (blue). (c) is the ED nodes and their edges. (d) presents the ED nodes and the target depth.

5.1.3 Lossy two-level optimization

Explained in Section 5.1.2, the size of the PR nodes is almost constant due to the limited size of depth image in a scenario where the map keeps expanding. For instance, the point cloud generated from Hamlyn dataset [91] (grabbed from monitor) is only $320 \times 240 = 76800$ at most. **As the model grows, the total number of nodes in the ED graph is increasing but the number of PR nodes is almost constant**. Taking advantage

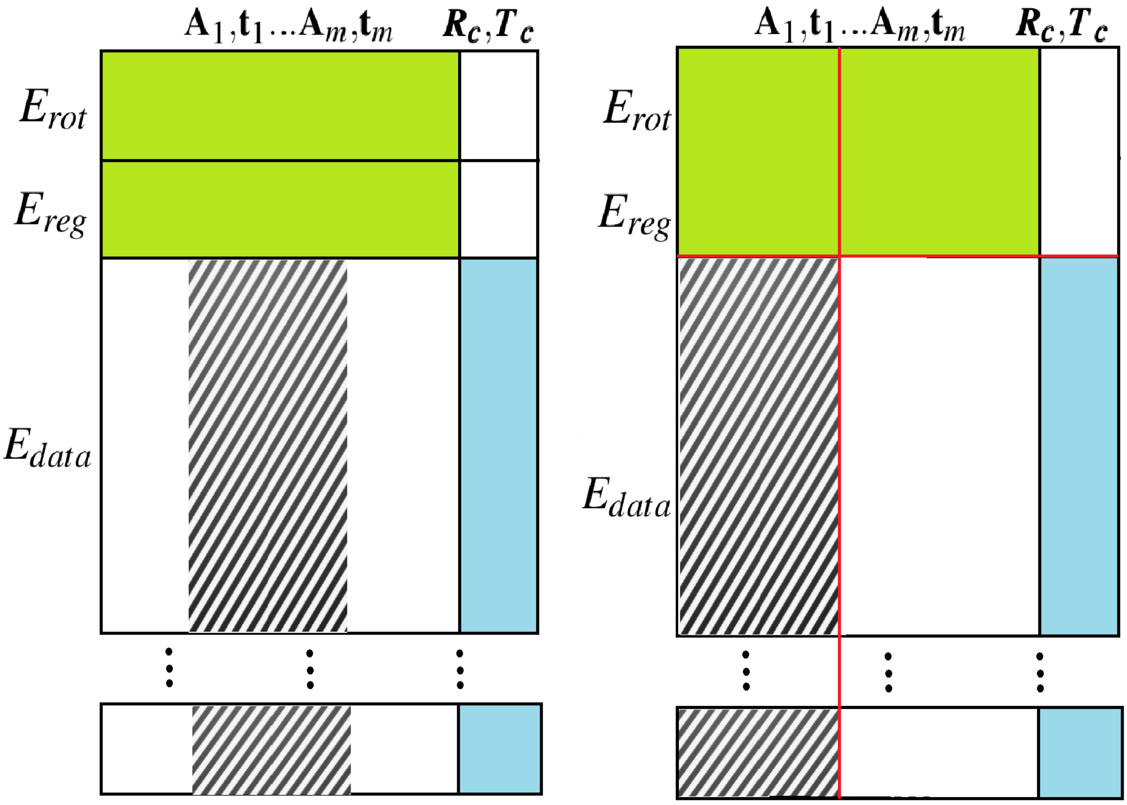

$\mathbf{A}_1,\mathbf{t}_1...\mathbf{A}_m,\mathbf{t}_m$
$\boldsymbol{R_c},\boldsymbol{T_c}$
E_{rot}
E_{reg}
E_{data}
$\mathbf{A}_1,\mathbf{t}_1...\mathbf{A}_m,\mathbf{t}_m$
$\boldsymbol{R_c},\boldsymbol{T_c}$
E_{rot}
E_{reg}
E_{data}

FIGURE 5.2: (a) is an example of Jacobian. Empty block means the element in this block are zero. (b) is re-ordered Jacobian.

of Eq. (5.9), the optimization can be divided into two levels: the optimization of PR nodes $\mathbf{\Phi}_1$ and the optimization of the rest PI nodes $\mathbf{\Phi}_2$.

Therefore, this work first optimizes ($\mathbf{\Phi}_1$, $\mathbf{R}_c$ and $\mathbf{T}_c$) by fixing $\mathbf{\Phi}_2$ in (**Level I**) optimization, to obtain an estimation of the three parameters $\mathbf{\Phi}_1$, $\mathbf{R}_c$ and $\mathbf{T}_c$. Then the value of the parameters obtained from **level I**, will be fixed in **Level II**, together with the two soft constraints E_{rot} and E_{reg} to optimize the parameter Φ_2. This chapter explicitly enforces the idea with the following formulation:

$$\underset{\mathbf{R}_c,\mathbf{T}_c,\mathbf{\Phi_1}}{\operatorname{argmin}} \ \omega_{rot}\tilde{E}_{rot} + \omega_{reg}\tilde{E}_{reg} + \omega_{data}E_{data} \tag{5.10}$$

$$\underset{\mathbf{\Phi}_2}{\operatorname{argmin}}\ \omega_{rot}E_{rot} + \omega_{reg}E_{reg} \tag{5.11}$$

Eq. (5.10) and Eq. (5.11) are the **Level I** and **Level II** energy functions, where $\tilde{E}_{rot}$ and $\tilde{E}_{reg}$ are the subsets of energy function of E_{rot} and E_{reg} containing Φ_1. In other words, the size of Eq. (5.10) is only related to the size of PR nodes. Therefore, the computational complexity in optimizing **Level I** is reduced from $O(n^2)$ to constant $O(1)$ thanks to the constant size of $\mathbf{\Phi}_1$. Undoubtedly, optimizing **Level II** is still $O(n^2)$, but considering the scale of data term E_{data} is far larger than the rest, the computational cost in **Level II** is much smaller. Note that the new strategy keeps time consuming step **Level I** constant while **Level II** still $O(n^2)$. **But the size of Level II is almost negligible comparing to Level I**.

5.1.4 Connection with marginalization and information loss

This section draws the connection of the proposed two-level optimization method with an exact marginalization based method. The analysis will show that the information loss is very low, illustrating the feasibility of the decoupled optimization Eq. (5.10) and Eq. (5.11).

When generating Eq. (5.9), the Jacobian shown in Fig. 5.2 is re-ordered by classifying $[\mathbf{A}_1, \mathbf{t}_1 ... \mathbf{A}_m, \mathbf{t}_m]$ into PR nodes set $\mathbf{\Phi}_1$ and PI nodes set $\mathbf{\Phi}_2$. The state in cost function $[\mathbf{R}_c, \mathbf{T}_c, \mathbf{A}_1, \mathbf{t}_1 ... \mathbf{A}_m, \mathbf{t}_m]$ are classified as PR nodes with global pose $\mathbf{x}_c = (\mathbf{R}_c, \mathbf{T}_c, \mathbf{A}_1, \mathbf{t}_1 ... \mathbf{A}_k, \mathbf{t}_k)$ and PI nodes $\mathbf{x}_f = (\mathbf{A}_{k+1}, \mathbf{t}_{k+1} ... \mathbf{A}_m, \mathbf{t}_m)$. Fig. 5.2 shows the Jacobian in the new order. The first two term are combined due to their sparsity because E_{rot} and E_{reg} are node-wise and are unrelated to source points. The only full block in Fig. 5.2 is E_{data} with regard to $\mathbf{\Phi_1}$ (shadow region), specifically $\frac{\partial \mathbf{J}_2}{\partial \mathbf{x}_c}$. Let us write down the Jacobian and Hessian as,

$$\mathcal{J} = \begin{bmatrix} \frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_c} & \frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_f} \\ \frac{\partial \mathbf{J}_2}{\partial \mathbf{x}_c} & \mathbb{O} \end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix} \mathbf{J}_{1c} & \mathbf{J}_{1f} \\ \mathbf{J}_{2c} & \mathbb{O} \end{bmatrix} \tag{5.12}$$

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{J}_{1c}^T\mathbf{J}_{1c} + \mathbf{J}_{2c}^T\mathbf{J}_{2c} & \mathbf{J}_{1c}^T\mathbf{J}_{1f} \\ \mathbf{J}_{1f}^T\mathbf{J}_{1c} & \mathbf{J}_{1f}^T\mathbf{J}_{1f} \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\Lambda}_{cc} & \boldsymbol{\Lambda}_{cf} \\ \boldsymbol{\Lambda}_{cf}^T & \boldsymbol{\Lambda}_{ff} \end{bmatrix} \tag{5.13}$$

Obviously, the density of $\frac{\partial \mathcal{J}_2}{\partial \mathbf{x}_c}$ makes $\boldsymbol{\Lambda}_{cc}$ the only dense block among Hessian $\boldsymbol{\mathcal{H}}$. Taking this advantage we use marginalization technique [54] from classic rigid SLAM and separate the optimization in following form:

$$\begin{bmatrix} \boldsymbol{\Lambda}_{cc} & \boldsymbol{\Lambda}_{cf} \\ \boldsymbol{\Lambda}_{cf}^T & \boldsymbol{\Lambda}_{ff} \end{bmatrix} \begin{bmatrix} \mathbf{x}_c \\ \mathbf{x}_f \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{1c}^T\mathbf{F} + \mathbf{J}_{2c}^T\mathbf{F} \\ \mathbf{J}_{1f}^T\mathbf{F} \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{y}_c \\ \mathbf{y}_f \end{bmatrix} \tag{5.14}$$

$$\begin{bmatrix} \boldsymbol{\Lambda}_{cc} - \boldsymbol{\Lambda}_{cf}\boldsymbol{\Lambda}_{ff}^{-1}\boldsymbol{\Lambda}_{cf}^T & \mathbb{O} \\ \boldsymbol{\Lambda}_{cf}^T & \boldsymbol{\Lambda}_{ff} \end{bmatrix} \begin{bmatrix} \mathbf{x}_c \\ \mathbf{x}_f \end{bmatrix} = \begin{bmatrix} \mathbf{y}_c - \boldsymbol{\Lambda}_{cf}\boldsymbol{\Lambda}_{ff}^{-1}\mathbf{y}_f \\ \mathbf{y}_f \end{bmatrix} \tag{5.15}$$

After enforcing Schur complement, this work successfully achieves only solving $\mathbf{x}_c$ independent of $\mathbf{x}_f$. The computation of $\mathbf{x}_c$ (including $\mathbf{R}_\mathrm{c}$, $\mathbf{T}_\mathrm{c}$ and $\boldsymbol{\Phi}_1$) is constant (explained in Section 5.1.3). After solving $\mathbf{x}_c$, the optimization of $\mathbf{x}_f$ is in tiny computation as the $\boldsymbol{\Lambda}_{ff} = (\frac{\mathbf{J}_1}{\partial \mathbf{x}_f})^T(\frac{\mathbf{J}_1}{\partial \mathbf{x}_f})$ and is only related to E_{rot} and E_{reg}. The sparsity of Hessian and small number of nodes ($n \gg m$) makes the time of solving $\mathbf{x}_f$ much less.

Fig. 5.3 is the Jacobian of E_{rot} and E_{reg} relating to all nodes. The first term E_{rot} is the sum error of affine transformation (Eq. (3.5)) making the Jacobian strictly diagonal. The second term E_{reg} defines the transformation error within node and its neighbors (Eq. (3.7)). The major part $\mathbf{A}_j(\mathbf{g}_k - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j$ is also within one node j except the very last $-(\mathbf{g}_k + \mathbf{t}_k)$. The last variable $\mathbf{t}_k$ makes the Jacobian not strictly diagonal (refer to Fig. 5.3(b)). This works comes up with an idea that by ignoring Jacobian of $\mathbf{t}_k$, **Level I** and **Level II** energy function are separable. Therefore, after reordering two diagonal Jacobians, the combination $[\frac{\partial \mathcal{J}_1}{\partial \mathbf{x}_c}\frac{\partial \mathcal{J}_1}{\partial \mathbf{x}_f}]$ is diagonal and $(\frac{\partial \mathcal{J}_1}{\partial \mathbf{x}_c})^T(\frac{\partial \mathcal{J}_1}{\partial \mathbf{x}_f}) = 0$. Eq. (5.14) is transformed to Eq. (5.15) solved directly without strict marginalization.

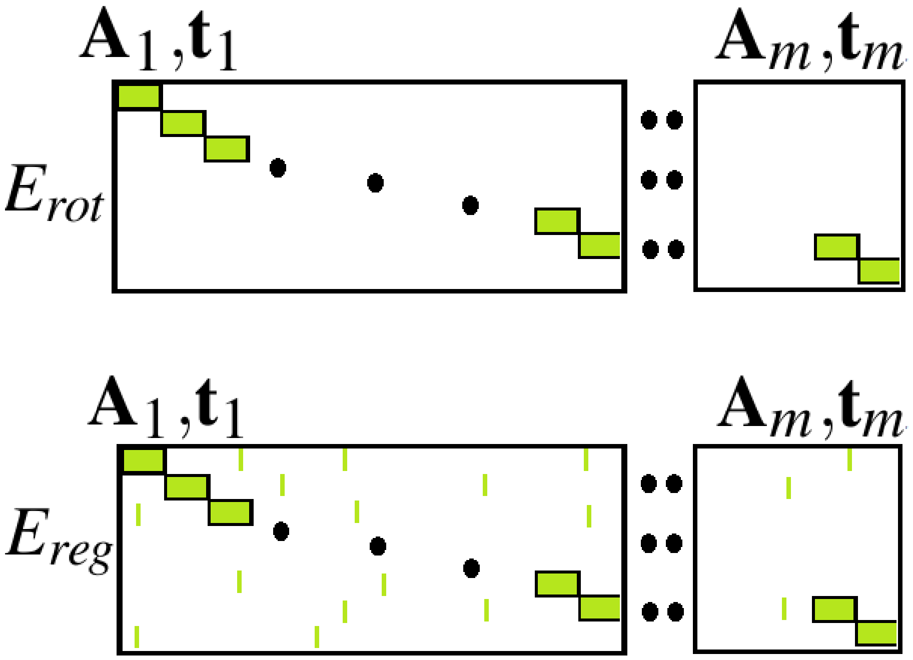

$\mathbf{A}_1,\mathbf{t}_1$
$\mathbf{A}_m,\mathbf{t}_m$
E_{rot}
$\mathbf{A}_1,\mathbf{t}_1$
$\mathbf{A}_m,\mathbf{t}_m$
E_{reg}

FIGURE 5.3: (a) is Jacobian of E_{rot} with regard to all nodes. (b) is Jacobian of E_{reg} with regard to all nodes.

Fig. 5.4 visualizes the feasibility of the lossy decoupled optimization approach in geometry. PR nodes (green) are the only nodes that are connected to visible points and contribute to E_{data}. All PI nodes (purple) merely share edges with PR nodes and are passively deformed according to the behaviors of PR nodes. Equivalently, the inter-nodes relations in the Jacobian of E_{reg} (Fig. 5.3(b)) shows these connections (Fig. 5.3(b)). In view of this, the proposed lossy decoupled optimization approach first optimizes PR nodes and then estimates PI nodes.

In conclusion, solving energy function Eq. (3.4) by first ignoring Jacobian of $\mathbf{t}_k$ is equivalent to the proposed decoupled optimization in Eq. (5.10) and Eq. (5.11).

The information loss of the proposed approach is relatively low. Fig. 5.4 illustrates the connection between PR nodes and PI nodes is weak on the boundary, in contrast with the dense connections among PR nodes. Fig. 5.5 demonstrates how the connection between PR and PI nodes are removed in **Level I** optimization. Correspondingly, the connection between $\frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_c}$ and $\frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_f}$ are removed resulting in $\mathbf{\Lambda}_{cf} = (\frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_c})^T(\frac{\partial \mathbf{J}_1}{\partial \mathbf{x}_f}) = \mathbb{O}$. The

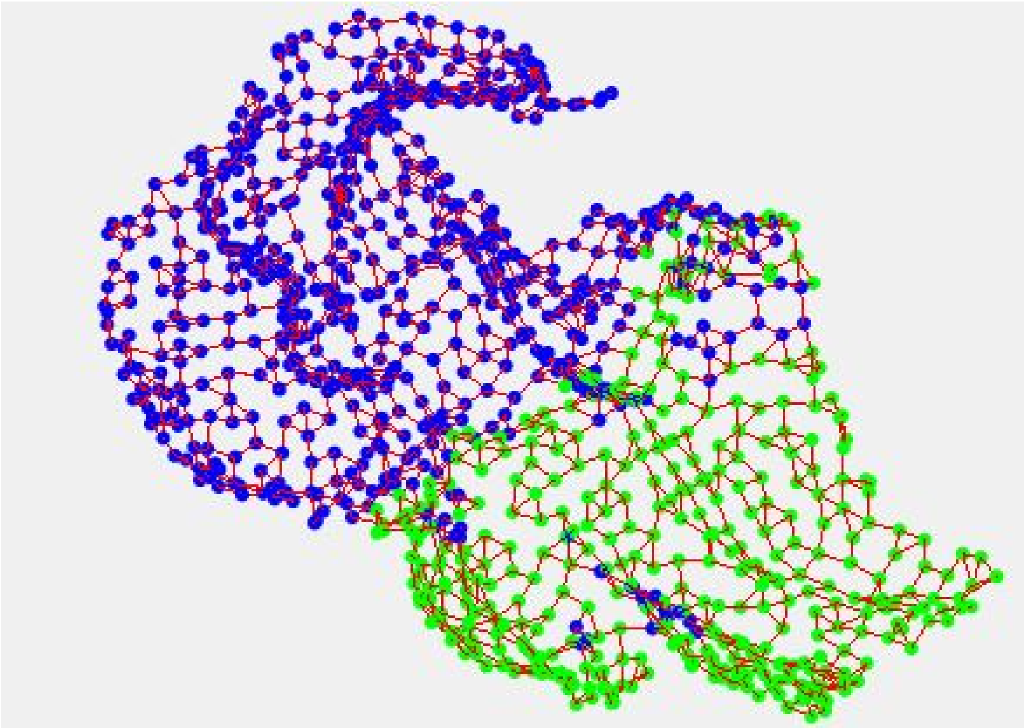


Figure 5.4: Two types of nodes and edges. Green nodes are the PR nodes and purple nodes are PI nodes.

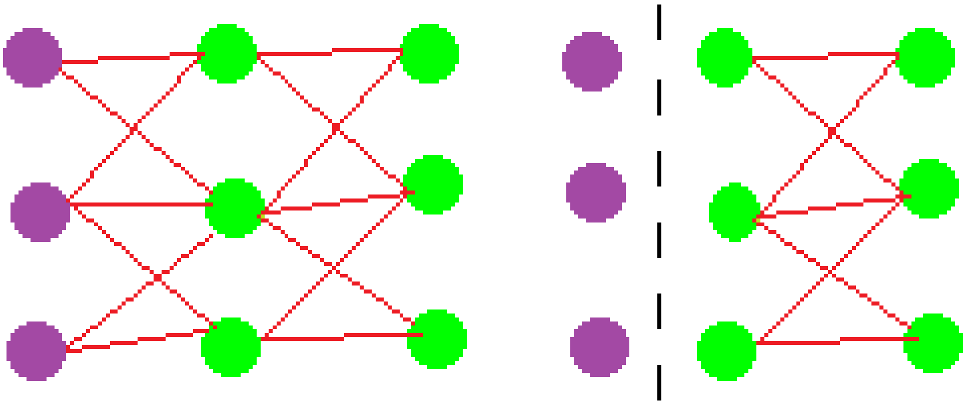


Figure 5.5: Connections of PI (purple) and PR (green) nodes. Left is the full connection while the PI and PR connections are cut in **Level I** optimization

information between two PR nodes is strong while that among the PI nodes is weak. The weak information is neglected in **Level I**, attributing to relatively low information loss in optimization process.

5.2 Results and discussion

The goal is to have both qualitative assessments as well as quantitative comparisons between the original ED graph optimization and the proposed 2 level optimization method. For qualitative comparison, this chapter shows the sacrificed accuracy has few impacts on the final reconstructed map. A dolphin model is downloaded from Turbosquid (`https://www.turbosquid.com`) for qualitative comparison. With regard to quantitative test in SLAM, both methods are compared on a tiny synthetic dataset and datasets from the Hamlyn Centre for Robotic Surgery [91], where this work chooses three in-vivo stereo videos with deformation and rigid scope movement. Other videos either have no deformation or scope motion. The frame rate and size of the in-vivo porcine dataset (model 1) are 30 frames per second and 640×480 while the other two is 25 frames per second and 720×288. Distance from camera to surface of soft-tissue is between 40 to 70 mm. The experiments are conducted on the same hardware and software environment of MIS-SLAM (Chapter 4). The module of state estimation in MIS-SLAM is modified to **Level I** optimization described in this thesis. Note that an iterative solver, i.e. the preconditioned conjugate gradient method, is employed to solve the resulting linear systems, as it provides a way of parallel computing on GPU. Section 5.2.1 shows the qualitative comparisons while the rest shows time complexity as well as the accuracy.

5.2.1 Qualitative ED deformation comparisons

Fig. 5.6 demonstrates the comparisons of ED graph deformed dolphin (middle) and the result with the proposed approach (right). This comparison is not aiming as proof of the superiority of the proposed method over the original ED graph method. Sumner et al. [83] has already claimed real-time implementation on CPU as well as very nice results. Aiming at speeding up deformable SLAM application, the qualitative result of the proposed lossy

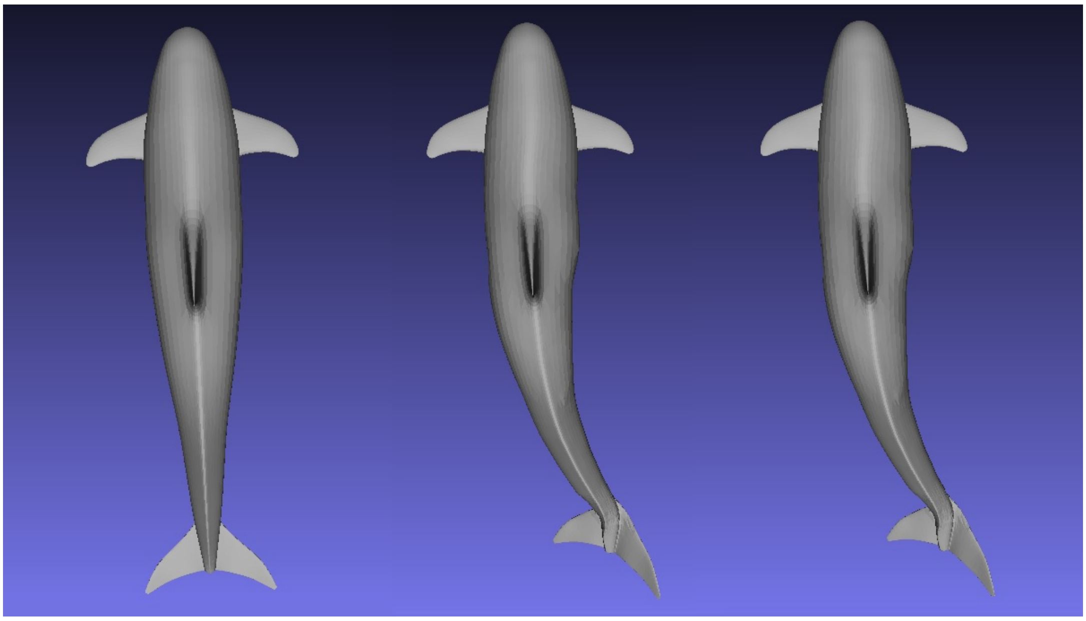


FIGURE 5.6: Qualitative comparisons of the proposed strategy and original ED based deformation. The first shape is the original dolphin mesh. It shows the result of deformed shape (the last) along with the result of classical ED deformation (middle).

decoupled approach is not comparable to the batch estimation of ED. However, this chapter aims at illustrating that the proposed method can achieve a similar result as ED graph and the difference is not visible to the naked eye or difficult to make out. Fig. 5.6 confirms that the deformed shapes performed by the proposed approach do actually have a similar result. The result looks very close to the original ED graph partially due to the simple topology of dolphin. Other complex models like a human will result in a visible but not very apparent difference. Although the decoupled optimization normally works well, the results can be much worse than ED when the nodes are too sparse. The deformation is dependent on PR nodes involved and the insufficiency of PR nodes (or nodes in conventional one-step ED) causes the wrong deformation that the expected target is not reached.

5.2.2 Time complexity comparisons

Fig. 5.7 and Fig. 5.8 show a tiny one-step toy simulation and the result. The tractable time consumption remains small because the PR node does not change.

This chapter compares the original MIS-SLAM (Section 4) with the improved version. Fig. 5.9 illustrates the running time for three Hamlyn datasets (model 6, 20 and 21). In all

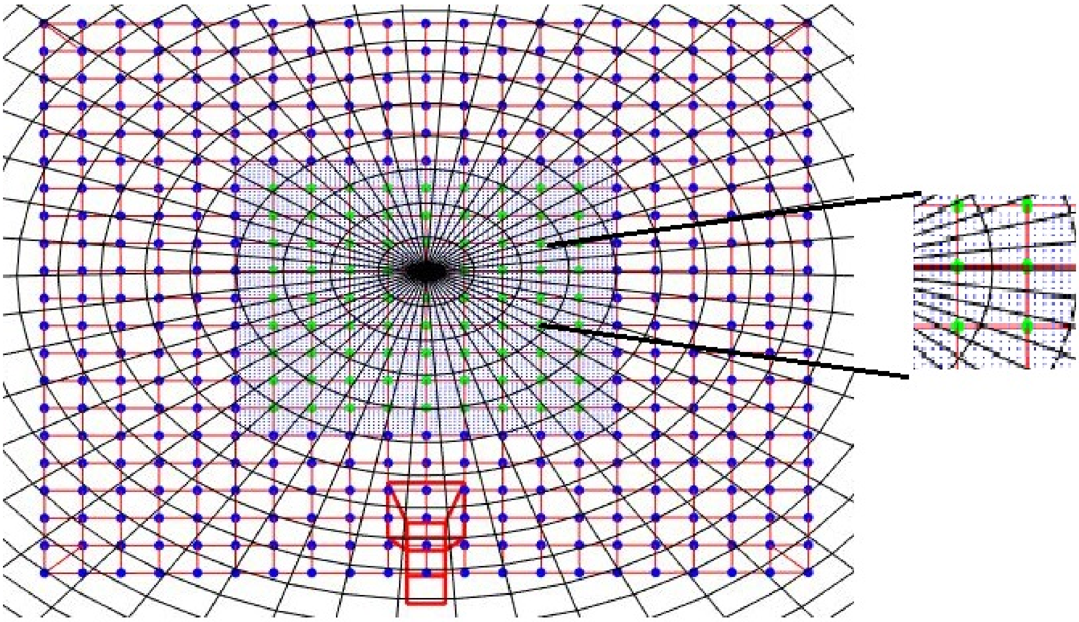


Figure 5.7: A toy model. One step optimization step from plain visible surface model (tiny dots) to warped surface (grid). PI nodes are in blue and PR nodes are in green. Camera remains static.

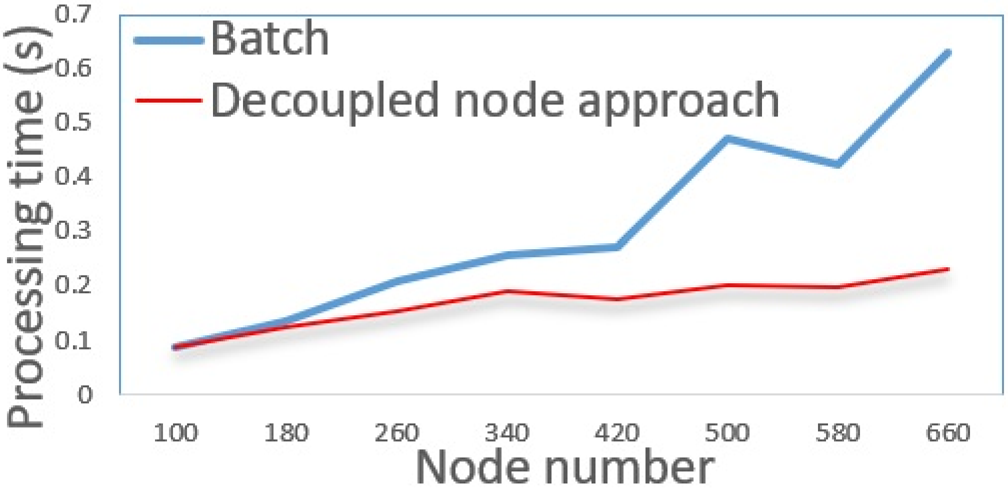


Figure 5.8: Test results relating to Fig. 5.7.

scenarios decoupled optimization yields better efficiency than batch processing especially in case of long term process (the last dataset in Fig. 5.9). In the first few steps, the robot is steady and ED graph is not expanding significantly. This attributes to the similar processing time in the first few hundred steps. When the robot moves, the ED graph expands intensively and processing time increases abruptly in state optimization. In view of this, by limiting the size of the node graph, decoupled optimization keeps time consumption stable due to the constant PR node scale_.

From the demo video on youtube [1], readers will find the range of movement in model 6 is much smaller than model 20 and 21. That's the main reason the proposed method does not contribute greatly to model 6. However, as the environment gets larger, the proposed approach keeps much lower time consumption.

Fig. 5.10 also shows the number of decoupled nodes. The algorithm significantly keeps the **Level I** tractable. Note that this chapter does not present the time for **Level I** and **Level II** separately because the time consumption of **Level II** is only a few percent of **Level I**. Normally the size of energy function for **Level I** is 15000 to 40000 while there are only 500 to 1500 for **Level II** optimization. Based on the test on GPU, the time consumption of a typical **Level I** is 0.1 second while **Level II** is around 0.003 second.

5.2.3 Accuracy comparisons

The lossy decoupled optimization strategy inevitably attributes the loss of accuracy. Section 5.2.1 shows the quality of the deformed map is well preserved in ED deformation process. Moreover, this work compares the optimization performance of the lossy formulation and the original one on the same parameters and weights of terms in SLAM application (MIS-SLAM). Different from arbitrary key points matching in ED deformation formulation (Eq. (5.5)), in SLAM application the E_{data} term is in the form of model-to-depth scan matching like point to plane ICP. For quantitative validation, this work measures the point-to-point distance of deforming map and target scan. For direct validation to ground truth, three synthetic datasets (heart, right kidney and stomach) are generated by deliberately deforming models from CT scanned phantom. This work compares the error

[1] https://youtu.be/7b7giRibvRI

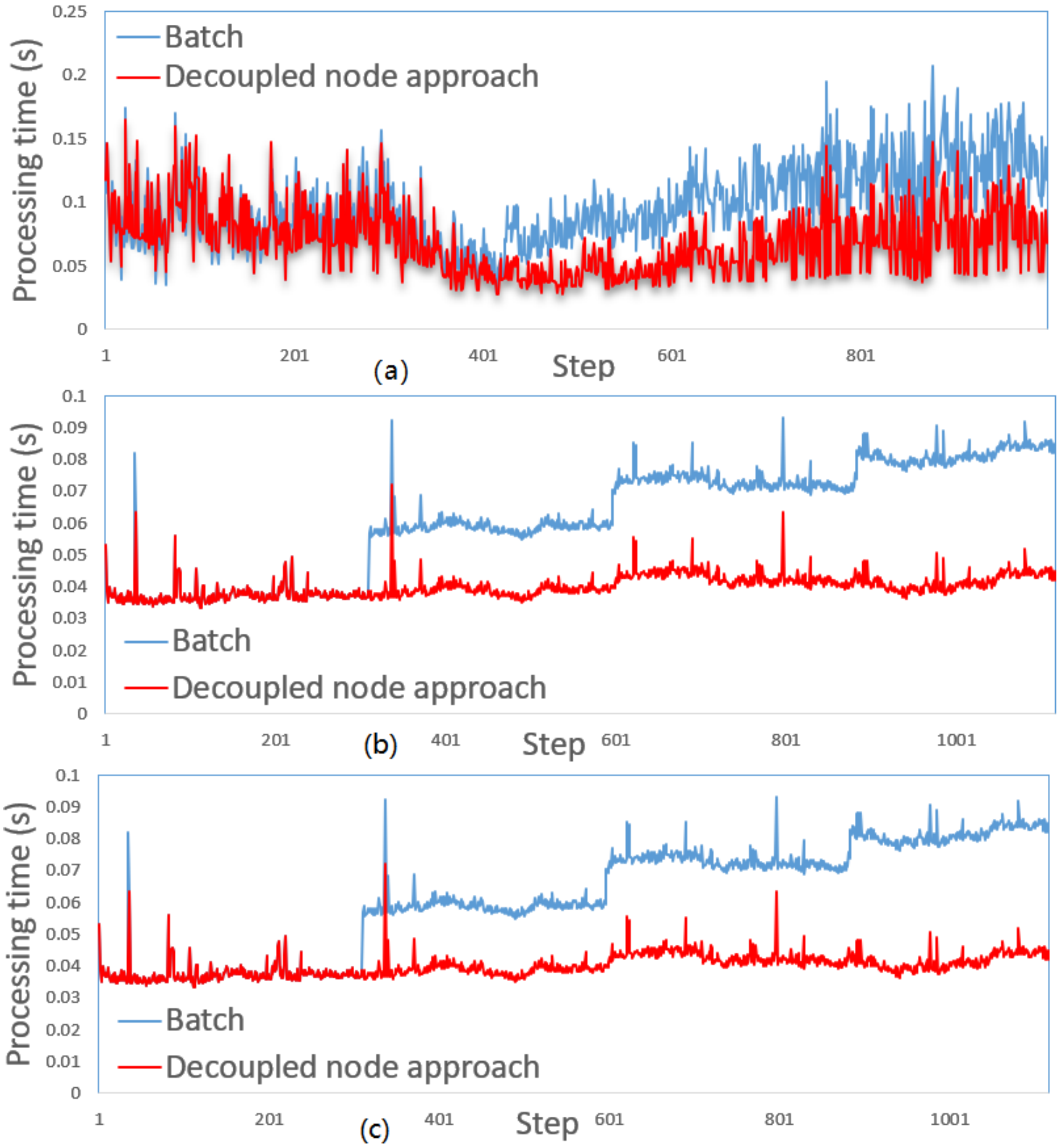


FIGURE 5.9: Processing time comparsions of model 6 (a), 20 (b) and 21 (c) in Hamlyn dataset. Blue lines are the batch optimization and red lines are the nodes decoupled optimization. **Level I** and **Level II** cannot be shown separately due to time consumption of **Level II** is extremely low.

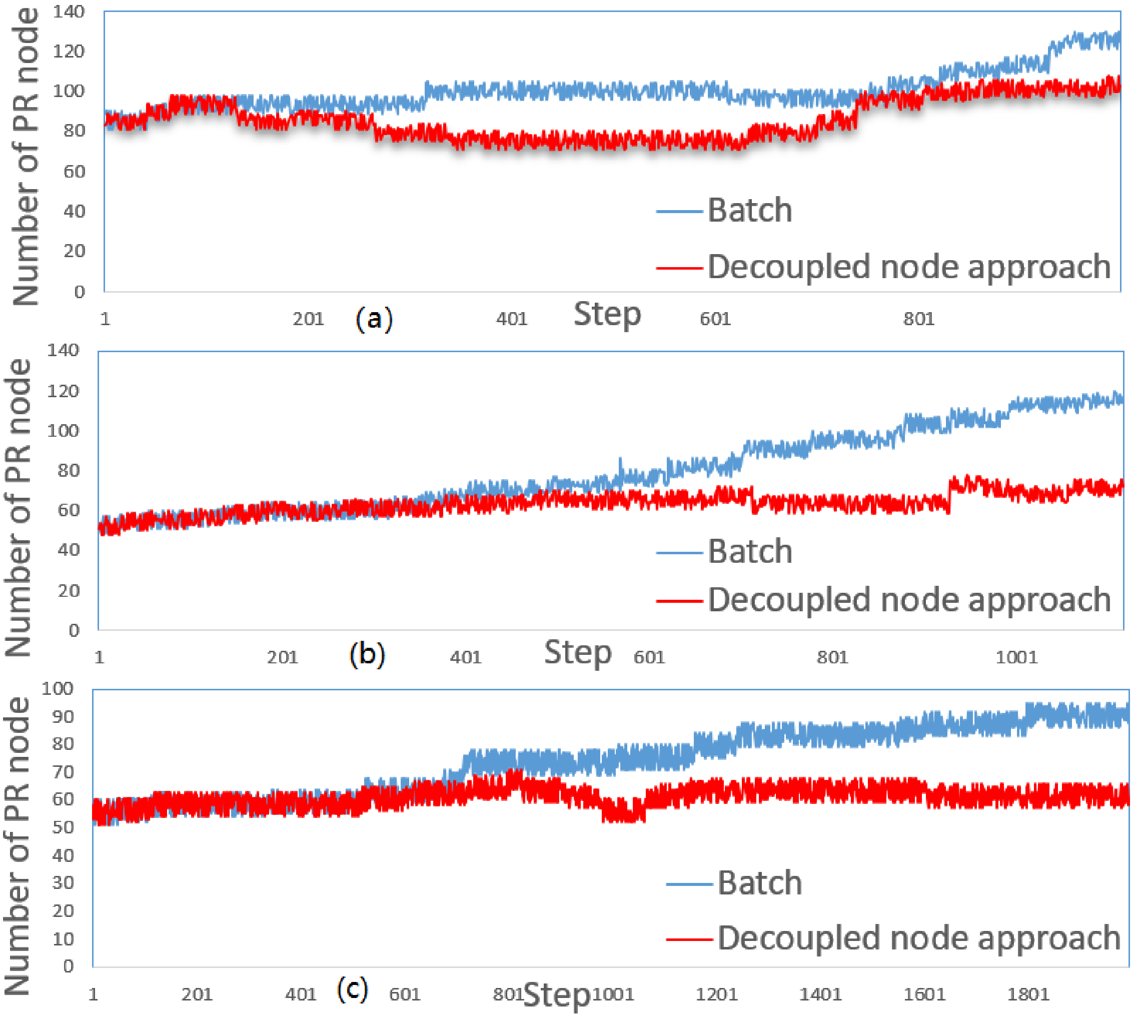


FIGURE 5.10: Optimizing nodes comparisons in first level computation of model 6 (a), 20 (b) and 21 (c) in Hamlyn dataset. The red lines are the result of the decoupled optimization strategy while the blue lines are the original batch strategy.

from the reconstruction to the ground truth. As a compliment, the three laparoscopy datasets from Hamlyn are tested, but only the back-projection error in each iteration is available since there's no ground truth. In the batch approach, the average distance of back-projection registration of the three simulation scenarios is 0.18 mm (model 6), 0.13 mm (model 20) and 0.12 mm (model 21). While dataset with ground truth (Hamlyn dataset 10/11) achieves 0.08mm, 0.21 mm (Average errors). With the proposed decoupled optimization approach, this chapter achieves 0.31 mm (model 6), 0.26 mm (model 20), 0.22 mm (model 21) and 0.14 mm, 0.29 mm errors. In the video on youtube [2], there is no big difference in terms of structure and texture.

[2] https://youtu.be/7b7giRibvRI

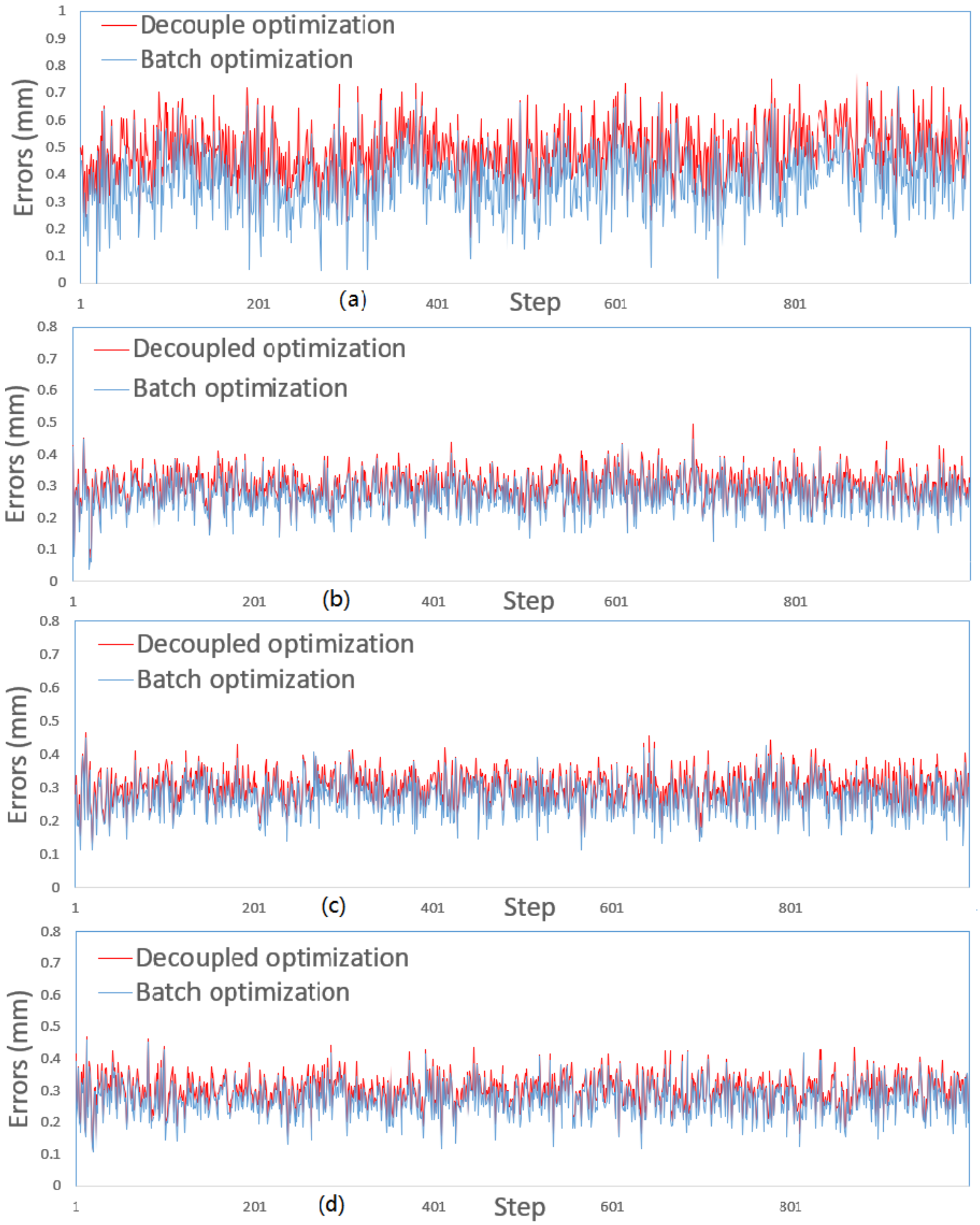


Figure 5.11: Mean average error of 3D reconstruction (mm). From the first to last are Hamlyn dataset with ground truth, the synthetic heart, synthetic left kidney and synthetic stomach. Please refer to video for the synthetic data.

Fig. 5.11 illustrates the average error of model to ground truth. On top of ground truth in Hamlyn dataset, this work generates several synthetic datasets with ground truth. Please refer to the video for more details of synthetic data.

5.3 Chapter summary

This chapter proposes a novel two-level deformation node decoupling approach that supports faster computation and reduces computational complexity from $O(n^2)$ to near $O(1)$. The decoupled optimization structure achieves faster computation in scenario of expanding environment. The proposed strategy sacrifice a small amount of accuracy in exchange for near-constant processing speed. The constant computation complexity of the lossy strategy should have great potential for applications in ED graph based SLAM applications in unbounded map scenario. Nevertheless, chapter 3, 4 and 5 all focus on 3D shape reconstruction in robocentric scenario. Next chapter will move toward estimating camera pose in worldcentric scenario.

Chapter 6

A time series SLAM algorithm for deformable environment

Chapter 3, 4 and 5 discuss ED graph based dense robocentric deformable SLAM, which measures the accuracy by the shape and texture of reconstructed geometries, and camera pose estimation is off topic. The reason is that the motion of the camera and an entirely deforming soft-tissue is intuitively non-separable. While there are numerous robocentric SLAM implementations in the deformable scenario, no analysis is reported on the separability, or defined as observability in classical control theory, of camera motion (transformation) and environment deformation (non-rigid deformation) in deformable SLAM. To theoretically revealing the relation, in this chapter, we extensively discuss the observability in the ED graph based SLAM. In the field of control, observability of system is defined of the ability to fully and uniquely recover the system state, from a finite number of observations of its outputs and the knowledge of its control inputs [61]. When ED graph is applied in non-static SLAM, the results of 3D reconstructions are intuitive, and many works ignore the accuracy of camera pose. Although the reconstructed geometry from ED graph based mapping is appealing, the motion is a mixed result of pose tracking and environment deformation estimating. Thus, our previous robocentric formulation is a bypass to separation. This chapter focuses on estimating camera pose and deforming environment. Particularly, the question is 'Is global pose of camera observable in an environment unique?'. If the answer is no, then 'How can we enable observability of pose in a deformable environment?'. There are four major contributions in this chapter: (1) A counterexample is provided when analyzing ED graph based visual SLAM system in the deformable environment. We clearly demonstrate that the global pose of the camera can be embedded into environment deformation formulation which is not separable. (2) We theoretically prove the above conclusion by analyzing the rank of the Fisher information matrix (FIM). (3) We propose an innovative back-end SLAM system with time series assumption which can efficiently calculate accurate pose as well as the deforming environment. (4) A prove is also provided to validate that the time series formulation is observable. The proposed time series method is inspired by a Fourier Transformation. We introduce a priori that theoretically deformation is a mixture of base shapes. Typical deformations this method are suitable to handle include heartbeat, breathe, periodic body movement. Other deformation can also, to some extent, be approximated by several historical basis shapes with rigid movement. The proposed time series method explicitly enforces correct observability constraints to overcome camera pose mixing with non-rigid deformation field. The result is compared with conventional static SLAM and ED formulation.

6.1 Observability analysis of ED based SLAM

Last three chapters 3, 4 and 5 as well as previous researches [12–14, 93] demonstrate the effectiveness of ED graph describing the deformation of environment. The three chapters only show the robocentric deformable SLAM and focus on the 3D reconstruction of the soft tissue. The goal of this section is to analyze the observability by presenting the basic form and the corresponding matrix formulation of ED graph. Based on the matrix formulation, observability of ED graph formulation is analyzed with an example as well as theoretical prove. Surprisingly, the rotation and translation of the camera can be precisely mixed with ED graph. Based on the analysis, we conclude that the global pose and local deformation cannot be accurately estimated unless prior environment motion information is available. This chapter is the cornerstone for a new time series based SLAM algorithm for better localization of the camera assuming that feature behaves in a mixture of historical trajectory.

Aiming at proving the ED based formulation is not observable when we turn to 'world-centric' instead of previously discussed 'robocentric', we follow the classical observability definition [94] by showing the camera pose with ED graph formulation is not solvable, that is the existence of multiple optimal solutions for the ED formulation. When there is no adequate information to uniquely obtain the solution, we call the problem is unobservable [94]. Intuitively, the definition demonstrates the close connection within two intertwined variables, showing the underlying reason for unobservability. Moreover, we also follow another classical way of testing system observability [61][54] by presenting a theoretical prove based on information matrix analysis.

6.1.1 Qualitative analysis of ED based SLAM formulation

By combining Eq. (5.1), Eq. (7), Eq. (5.3), Eq. (5.4), we upgrade single point transformation Eq. (3.2) to multiple points (model) transformation matrix formulation:

$$E_{data} = \mathbf{R}_c \cdot [\mathbf{\Lambda} \cdot \mathbf{M} + \mathbf{T} \cdot \mathbf{C}] + \mathbf{T}_c \otimes \mathbf{1} - \hat{\mathbf{P}} \tag{6.1}$$

In SLAM problem formulation, the state vector is denoted as $X_i = [\mathbf{R}_{ci}, \mathbf{T}_{ci}, \mathbf{\Lambda}_i, \mathbf{T}_i]$ in i-th step.

Camera to target measurement model: A typical SLAM observation model is a range and bearing model. In practice, there are several different measurements due to different sensors. Back-projection presentation is the most widely adopted observation model in RGB-D and stereo SLAM [19] [20]. It is a modified version of ICP taking advantage of regularized 2D depth observation, but in essence it is point to point pairing. For simplicity, we employ basic observation model, that is feature positions are directly observed by camera.

We propose a counter-example to illustrate Eq. (3.2 is not observable. According to the definition [94], if there exists multiple optimal solutions $[\mathbf{R}_c, \mathbf{T}_c, \mathbf{\Lambda}, \mathbf{T}]$ in one step transitional process (Eq. (6.1)), it is at least partially not observable. Multiple optimal solutions lead to low-rank of the FIM. The study of parameter observability examines whether the information provided by the available measurements is sufficient for estimating parameters without ambiguity [61]. In other words, multiple solutions to the problem can be found attributed to insufficient information. Therefore, unobservability can be proved if multiple solutions to Eq. (6.1) are found, meaning global pose and non-rigid deformation formulation combined is not observable at the same time.

Here we show there are infinite solutions to Eq. (6.1). For the optimal solution [$\hat{\mathbf{R}}_c$ $\hat{\mathbf{T}}_c$ $\hat{\mathbf{\Lambda}}$ $\mathbf{T}$], we define an arbitrary rotation matrix $\Delta\mathbf{R}$. For a set of point cloud transformation (from $\mathbf{P}$ to $\hat{\mathbf{P}}$), Eq. (6.1) with the state vector [$\hat{\mathbf{R}}_c$ $\hat{\mathbf{T}}_c$ $\hat{\mathbf{\Lambda}}$ $\mathbf{T}$] can be reformulated into following form:

$$\begin{aligned}\hat{\mathbf{P}} &= \hat{\mathbf{R}}_c[\hat{\mathbf{\Lambda}}\mathbf{M} + \mathbf{T}\mathbf{C}] + \hat{\mathbf{T}}_c \otimes \mathbf{1} \\ &= \hat{\mathbf{R}}_c\Delta\mathbf{R}\Delta\mathbf{R}^T[\hat{\mathbf{\Lambda}}\mathbf{M} + \hat{\mathbf{T}}\mathbf{C}] + \hat{\mathbf{T}}_c \otimes \mathbf{1} \\ &= \hat{\mathbf{R}}_c\Delta\mathbf{R}[\Delta\mathbf{R}^T\hat{\mathbf{\Lambda}}\mathbf{M} + \Delta\mathbf{R}^T\hat{\mathbf{T}}\mathbf{C}] + \hat{\mathbf{T}}_c \otimes \mathbf{1}\end{aligned} \tag{6.2}$$

Therefore, there is a new solution [$\hat{\mathbf{R}}_c\Delta\mathbf{R}$, $\hat{\mathbf{T}}_c$, $\Delta\mathbf{R}^T\hat{\mathbf{\Lambda}}$, $\Delta\mathbf{R}^T\mathbf{T}$]. Considering $\Delta\mathbf{R}$ is arbitrary, it's obvious that the incremental camera rotation $\mathbf{R}_c$ can be offset by rotating the affine transformations matrix $\mathbf{\Lambda}$ in the opposite direction. For the **Rotation** constraint, $\hat{\mathbf{\Lambda}}$ is transformed by a rotation matrix which means E_{rot} is unchanged. For the **Regularization** constraint:

$$E_{reg} = ||\Delta\mathbf{R}^T[\hat{\mathbf{\Lambda}}\mathbf{M}' + \mathbf{T}\mathbf{C} + \mathbf{T}]||_F^2, \tag{6.3}$$

where $\mathbf{M}'$ is similar to $\mathbf{M}$ with $\mathbf{v}_i$ $(i = 1, ..., n)$ substituted by $\mathbf{g}_i$ $(i = 1, ..., m)$. And $|| \cdot ||_F^2$ is the Frobenius norm. E_{red} is a rotation of previous vector and the vector norm remains unchanged. In all, the new solutions [$\hat{\mathbf{R}}_c\Delta\mathbf{R}$, $\hat{\mathbf{T}}_c$, $\Delta\mathbf{R}^T\hat{\mathbf{\Lambda}}$, $\Delta\mathbf{R}^T\mathbf{T}$] are also the optimal solutions to objective function Eq. (3.4) in addition to [$\hat{\mathbf{R}}_c$ $\hat{\mathbf{T}}_c$ $\hat{\mathbf{\Lambda}}$ $\mathbf{T}$].

Similarly, for any arbitrary $\Delta\mathbf{T}$, we can find additional solutions satisfying Eq. (3.4). Note that $\Delta\mathbf{T} \otimes \mathbf{1} = \mathbf{R}_c\Delta\mathbf{T} \otimes \mathbf{1}\mathbf{C}$ due to the fact that the column sum of $\mathbf{C}$ is always to 1 (sum of weight). Thus we rewrite Eq. (6.1) to:

$$\begin{aligned}\hat{\mathbf{P}} &= \hat{\mathbf{R}}_c[\hat{\mathbf{\Lambda}}\mathbf{M} + \mathbf{TC}] + (\hat{\mathbf{T}}_c + \Delta\mathbf{T} - \Delta\mathbf{T}) \otimes \mathbf{1} \\ &= \hat{\mathbf{R}}_c[\hat{\mathbf{\Lambda}}\mathbf{M} + \mathbf{TC} + \Delta\mathbf{T} \otimes (\mathbf{1C})] + (\hat{\mathbf{T}}_c - \Delta\mathbf{T}) \otimes \mathbf{1} \\ &= \hat{\mathbf{R}}_c[\hat{\mathbf{\Lambda}}\mathbf{M} + (\mathbf{T} + \Delta\mathbf{T} \otimes 1)\mathbf{C}] + (\hat{\mathbf{T}}_c - \Delta\mathbf{T}) \otimes 1\end{aligned} \tag{6.4}$$

Accordingly, we have other solutions $[\hat{\mathbf{R}}_c,\ \hat{\mathbf{T}}_c - \Delta\mathbf{T},\ \hat{\mathbf{\Lambda}},\ \mathbf{T} + \Delta\mathbf{T} \otimes 1]$ to Eq. (3.4). For the **Rotation** constraint, $\hat{\mathbf{\Lambda}}$ remains independent to $\Delta\mathbf{T}$. For the **Regularization** constraint:

$$\begin{aligned}E_{reg} = &\sum_{j=1}^{m} \sum_{k \in \mathbb{N}(j)} \alpha_{jk} ||\mathbf{A}_j(\mathbf{g}_k - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j + \Delta\mathbf{T} - \\ &(\mathbf{g}_k + \mathbf{t}_k + \Delta\mathbf{T})||^2 = \sum_{j=1}^{m} \sum_{k \in \mathbb{N}(j)} \alpha_{jk} ||\mathbf{A}_j(\mathbf{g}_k - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j - (\mathbf{g}_k + \mathbf{t}_k)||^2\end{aligned} \tag{6.5}$$

Therefore, E_{reg} remains the same for new solution $[\hat{\mathbf{R}}_c\Delta\mathbf{R},\ \hat{\mathbf{T}}_c,\ \Delta\mathbf{R}^T\hat{\mathbf{\Lambda}},\ \Delta\mathbf{R}^T\mathbf{T}]$.

Remark 6.1. Prove is provided to show there are infinite number of optimal solutions to the energy function Eq. (3.4). The global rotation matrix $\mathbf{R}_c$ or translation matrix $\mathbf{T}_c$ are entangled with ED parameters $[\mathbf{\Lambda}, \mathbf{T}]$.

6.1.2 Prove of unobservability in ED based SLAM formulation

After a qualitative analysis, we provide observability analysis based on full FIM analysis. Based on the discussion above, the unobservable lies in the E_{data} defined in Eq. (6.1) with pairs $[\mathbf{R}_c, \mathbf{\Lambda}]$ and $[\mathbf{T}_c, \mathbf{T}]$, and is unrelated to Eq. (3.5) and Eq. (3.7). Since global transformation parameters $\mathbf{R}_c$ and $\mathbf{T}_c$ are irrelevant to E_{rot} and E_{reg}, the observability of these two terms are not affected. It's easy to prove that the partial FIM with regard to E_{rot} and E_{reg}, is full rank. Therefore, we only focus on the simplified case shown in Fig. 6.1 with regard to Eq. (6.1) to analyze the observability. The conclusion of this node and one step camera movement can be generalized to multiple steps with a larger ED graph. Similarly, we prove that the low rank is located in information matrix with regard to $[\mathbf{R}_c, \mathbf{\Lambda}]$ and $[\mathbf{T}_c, \mathbf{T}]$. For simplification, we consider the residual of a single point $\mathbf{p}$ deformed by m nodes to $\hat{\mathbf{p}}$:

$$E_{data}^{'} = \mathbf{R}_c[\mathbf{\Lambda}\mathbf{M} + \mathbf{TC}] + \mathbf{T}_c \otimes 1 - \hat{\mathbf{p}} \tag{6.6}$$

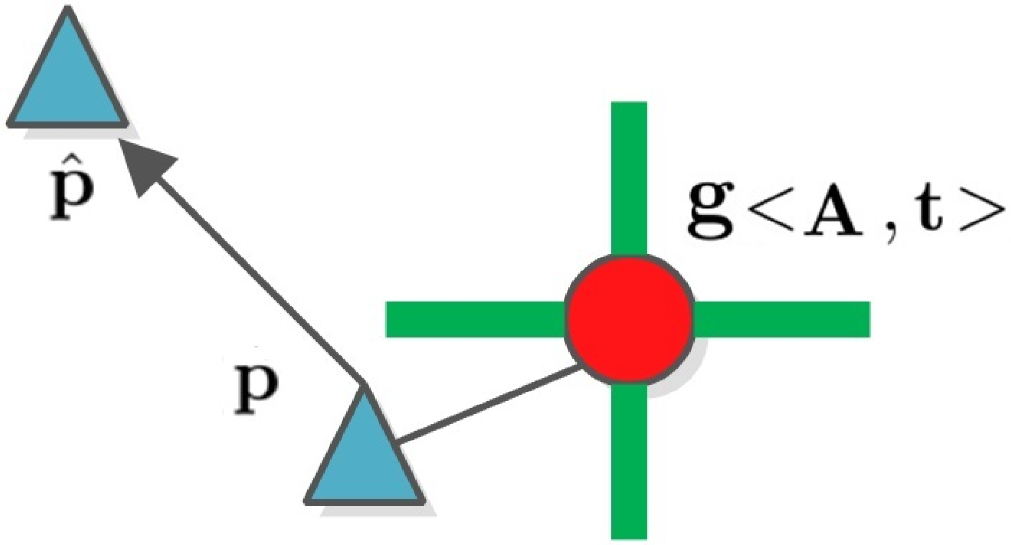


FIGURE 6.1: One step camera movement. Camera moves from $\mathbf{p}$ to $\hat{\mathbf{p}}$. The movement is a mixture of camera transformation and deformation by ED node $\mathbf{g}$. The red line are the connecting edge from node $\mathbf{g}$ to other nodes.

where the state are $[\mathbf{\Lambda}_m, \mathbf{T}_m, \mathbf{R}_c, \mathbf{T}_c]$. We vectorize the $\mathbf{\Lambda}_m$ and $\mathbf{T}_m$ and rewrite them into current form $[\overline{\mathbf{\Lambda}}, \overline{\mathbf{T}}, \mathbf{R}_c, \mathbf{T}_c]$. Lie algebra is applied to optimize rotation matrix $\mathbf{R}_c$. For the convenience, we mark following variables:

$$\underset{3\times 9m}{\widehat{\mathbf{R}}_\mathbf{c}} = \begin{bmatrix} \mathbf{R}_c & \cdots & \mathbf{R}_c \end{bmatrix} \tag{6.7}$$

$$\underset{3\times 3m}{\widetilde{\mathbf{R}}_\mathbf{c}} = \begin{bmatrix} \mathbf{R}_c & \cdots & \mathbf{R}_c \end{bmatrix} \tag{6.8}$$

$$\underset{3\times 3m}{\widehat{\mathbf{C}}} = \begin{bmatrix} \mathbf{C} & \cdots & \mathbf{C} \\ \vdots & & \vdots \\ \mathbf{C} & \cdots & \mathbf{C} \end{bmatrix} \tag{6.9}$$

$$\underset{3\times 9m}{\widehat{\mathbf{M}}} = \begin{bmatrix} \mathbf{M} & \cdots & \mathbf{M} \\ \vdots & & \vdots \\ \mathbf{M} & \cdots & \mathbf{M} \end{bmatrix} \tag{6.10}$$

$$\underset{3\times 3}{\mathbf{S}} = skew(\mathbf{\Lambda} \cdot \mathbf{M} + \mathbf{T} \cdot \mathbf{C}) \tag{6.11}$$

where $skew(\cdot)$ is the skew symmetric operator. The Jacobian of Eq. (6.6) with regard to $[\overline{\mathbf{\Lambda}}, \overline{\mathbf{T}}, \mathbf{R}_c, \mathbf{T}_c]$ is:

$$\mathbf{J} = \left(\; \widehat{\mathbf{R_c}} \odot \widehat{\mathbf{M}} \quad \widetilde{\mathbf{R_c}} \odot \widehat{\mathbf{M}} \quad -\mathbf{R}_c\mathbf{S} \quad \mathbf{I} \; \right) \tag{6.12}$$

where $\mathbf{I}$ is a 3 by 3 identity matrix. $\odot$ represents Hadamard product. Before estimating information matrix, we first mark the following matrix:

$$\underset{9m \times 9m}{\mathbf{A_1}} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \tag{6.13}$$

$$\underset{9m \times 3m}{\mathbf{A_2}} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \tag{6.14}$$

$$\underset{3m \times 3m}{\mathbf{A_3}} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \tag{6.15}$$

$$\underset{9m \times 3}{\mathbf{A_4}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{6.16}$$

$$\underset{3m \times 3}{\mathbf{A_5}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{6.17}$$

Based on all the definitions, the Hessian matrix $\boldsymbol{\mathcal{H}}_{ed}$ can be presented in the following form:

$$\boldsymbol{\mathcal{H}}_{ed} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \mathbf{H}_3 \\ \mathbf{H}_4 \end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix} \mathbf{A_1} \odot (\widehat{\mathbf{M}}^T\widehat{\mathbf{M}}) & \mathbf{A_2} \odot (\widehat{\mathbf{M}}^T\widehat{\mathbf{C}}) & -\widehat{\mathbf{M}}^T \odot (\mathbf{A_4}\mathbf{S}) & \widehat{\mathbf{R}}_{\mathbf{c}}^T \odot \widehat{\mathbf{M}}^T \\ \mathbf{A_2} \odot (\widehat{\mathbf{C}}^T\widehat{\mathbf{M}}) & \mathbf{A_3} \odot (\widehat{\mathbf{C}}^T\widehat{\mathbf{C}}) & -\mathbf{A_5} \odot (\widehat{\mathbf{C}}^T\mathbf{S}) & \widetilde{\mathbf{R}}_{\mathbf{c}}^T \odot \widehat{\mathbf{M}}^T \\ -(\mathbf{S}^T\mathbf{A_4}^T) \odot \widehat{\mathbf{M}} & -(\mathbf{S}^T\mathbf{A_5}^T) \odot \widehat{\mathbf{C}} & \mathbf{S}^T\mathbf{S} & -\mathbf{S}^T\mathbf{R}_c^T \\ \widehat{\mathbf{R}}_{\mathbf{c}} \odot \widehat{\mathbf{M}} & \widetilde{\mathbf{R}}_{\mathbf{c}} \odot \widehat{\mathbf{M}} & -\mathbf{R}_c\mathbf{S} & \mathrm{I} \end{bmatrix} \tag{6.18}$$

For the sub matrix $\mathbf{H}_1$ and $\mathbf{H}_2$ within the Hessian matrix $\boldsymbol{\mathcal{H}}_{ed}$, we split them into the group of every 3 lines. For example, $\mathbf{H}_1(i)$ is the group i ranging from line $3 * (i - 1) + 1$ to $3i$. By analyzing Hessian matrix $\boldsymbol{\mathcal{H}}_{ed}$, we discover the following law:

$$\mathbf{H}_1(i) = \dot{\mathbf{M}} \odot [-(\mathbf{S}^T)^{-1}\mathbf{H}_3] \tag{6.19}$$

$$\mathbf{H}_2(i) = \dot{\mathbf{C}} \odot [\mathbf{R}_c^T\mathbf{H}_4] \tag{6.20}$$

where $\dot{\mathbf{M}}$ and $\dot{\mathbf{C}}$ are defined in the following form:

$$\underset{3\times18m}{\dot{\mathbf{M}}} = \begin{bmatrix} \widehat{\mathbf{M}} & \widehat{\mathbf{M}} & \widehat{\mathbf{M}} \end{bmatrix} \tag{6.21}$$

$$\underset{3\times18m}{\dot{\mathbf{C}}} = \begin{bmatrix} \widehat{\mathbf{C}} & \cdots & \widehat{\mathbf{C}} \end{bmatrix} \tag{6.22}$$

Obviously, this one point transformation scenario can be extended to multiple points. Eq. (6.19) and Eq. (6.20) indicate that the global rotation $\mathbf{R}_c$ matrix and translation vector $\mathbf{T}_c$ can be embedded into local affine deformation matrix $\boldsymbol{\Lambda}]$ and $\mathbf{T}$ respectively. This conclusion also validates the qualitative conclusions (Remark 6.1) we draw in Section 6.1.1.

6.2 Priori based SLAM formulation

Section 6.1 shows the inner-connection between the relative transformation and the non-rigid deformation formulations. The two pairs, $(\mathbf{R}_c, \mathbf{\Lambda})$ and $(\mathbf{T}_c, \mathbf{T})$, are intertwined. Thus, with the camera to feature observation only, both global rotation and translation cannot be uniquely determined in conventional ED formulation on condition that no new information is provided. Robocentric SLAM is one efficient way to avoid the unobservability. Otherwise, there are an infinite number of solutions to the camera poses if camera to feature observations is the only source of input. **This conclusion can be generalized to other deformation formulation like FEM or structure-from-template because the degree-of-freedom of deformation enables model motion just with deliberately adjusted movement of model vertices.** With regard to this, the goal is to propose a prior to separate and determine the relative transformation from the deforming non-rigid tissue. Noteworthily, static SLAM algorithm with thresholds based feature classification strategy [39–41] also comes with prior, assuming the static features are separable and can be verified with thresholds. In this work, however, we still assume the whole soft-tissue is deforming. Experiments in Section 6.3 demonstrate that information matrix is full rank and estimated parameters are unique with the proposed priori.

In the field of nonrigid structure from motion, features are granted more freedom under the base shape constraints [95] [96]. Instead of one single static position in pose estimation, features are formulated with 3D locations in each frame. To prevent the irregular movement of the 3D shapes, base shapes [97], base trajectories [74] or base shape-trajectory [75] strategies are introduced to limit the degree of freedom of the soft shapes. They assume that the movement is a mixture of predefined bases, although these predefined bases are also unknown for the observation. After enforcing the bases, the freedom of the deformation is constrained and the rigidity of deformation can be controlled by the number of all the bases.

Taking advantage of this, we propose that deformation of the feature can be approximated by base historical shapes and the residuals of base shapes approximation are the camera movement. Theoretically, if provided with an infinite number of base shapes, deformation of features, as well as the camera, can be accurately estimated. In practice, comparing with traditional static SLAM or ED based SLAM, a limited number of base shapes can still generate good camera pose due to observability preserved. This is especially true in complex periodic deformation scenario where deformation is caused by breathing and heartbeat; the current deformed shape can be inferred from previous shapes.

Based on the proposed prior, a new feature motion formulation is introduced in the conventional back-end static SLAM formulation. In our study, the primary feature motion measurement is based on the idea that current structure $\mathbf{f}^{n+1}$ can be linearly fitted by its historical shapes $\mathbf{f}^{n}$ ... $\mathbf{f}^{n-t}$ where t is the processing window. A coefficient vector $\mathbf{c} = [\delta_1\ ,...,\ \delta_t]$ is introduced to describe the relations of these feature movements. The matrix $\mathbf{B}$ $(3N \times F)$ is the combination of all valid features. N is the number of features and F is the number of steps. Note that some elements in $\mathbf{B}$ is invalid because the viewing angle of the camera makes it unable to observe all features at all steps.

$$\mathbf{B} = \begin{bmatrix} \mathbf{f}_1^1 & \mathbf{f}_1^2 & \cdots & \cdots & \mathbf{f}_1^F \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{f}_N^1 & \mathbf{f}_N^2 & \cdots & \cdots & \mathbf{f}_N^F \end{bmatrix} \tag{6.23}$$

The term 'validity' of feature $\mathbf{f}_i^j$ in $\mathbf{B}$ refers to (1) feature i is observed by camera in step j and (2) feature i is observed in the period window t; in other words $[\mathbf{f}_i^j...\mathbf{f} + \mathbf{t}_i^j]$ are all observed by camera. The validity ensure building correlations in a consecutive movement of feature.

The proposed formulation is based on conventional back-end static SLAM. We first introduce static SLAM here. In 3D scenario where one camera freely moves with N static features, the state to be estimated is denoted as:

$$\mathbf{X} = [\mathbf{R}\ \mathbf{p}\ \mathbf{f}_1\ \cdots\ \mathbf{f}_N], \tag{6.24}$$

The general camera motion model from step n to $n+1$ without noise is described as:

Table 6.1: Pose and feature errors in Monte Carlo simulations.

	Deformable SLAM (m)	Least Square (m)	ED node based VO (m)
Simulation 1			
Camera Position X(m)	0.942	2.538	8.743
Camera Position Y(m)	0.526	1.012	3.197
Camera Heading (Rad)	0.005	0.009	0.014
Simulation 2			
Camera Position X(m)	0.119	0.277	2.098
Camera Position Y(m)	0.138	0.498	3.38
Camera Heading (Rad)	0.002	0.002	0.009

$$\begin{aligned}\mathbf{p}^{n+1} &= \mathbf{p}^{n} + \mathbf{R}^{n}\mathbf{v}^{n} \\ \mathbf{R}^{n+1} &= \mathbf{R}^{n}\omega^{n}\end{aligned} \tag{6.25}$$

where $\mathbf{v}^n$ is the linear translation of one step movement. $\omega^n \in \mathbb{SO}(3)$ is the rotation matrix describing orientation variation.

The static SLAM formation is modified by applying the time series method. When depicting feature motion, we are bereft of an analogy of conventional feature movement, so our implementation is to build a relationship of a given feature in consecutive movement. The formulation maneuvers to constrain feature motion to a mixture of historical movement. The constraint of feature motion model is expressed by building linear relations within a window of feature locations. The main advantage of linearly modeling the feature locations over historic base shape modeling is that it can initialize feature locations with the rigid assumption (using conventional visual odometry) and avoids base feature recognition. In non-rigid structure from motion, base shapes are essential to describe deformation. Moreover, base shapes require different window sizes for modeling which poses a heavy computational burden. The proposed linear constraint, however, is flexible and

Table 6.2: Pose and feature errors of heart, stomach and lung.

	Deformable SLAM	static SLAM	ED node based VO
Heart scenario			
Camera Position X(unit)	0.149	2.006	8.743
Camera Position Y(unit)	0.085	0.951	3.197
Camera Heading (Rad)	0.001	0.001	0.010
Stomach scenario			
Camera Position X(unit)	2.263	7.004	2.098
Camera Position Y(unit)	2.566	6.894	3.380
Camera Heading (Rad)	0.006	0.008	0.009
Lung scenario			
Camera Position X(unit)	2.009	7.596	2.098
Camera Position Y(unit)	0.706	3.304	3.380
Camera Heading (Rad)	0.002	0.003	0.009

straightforward to complex mixed deformation. In addition to the camera motion model Eq. (6.25), the proposed linear feature motion is modeled as:

$$\mathbf{f}_i^{n+1} = \delta_1 \cdot \mathbf{f}_i^{n} + \delta_2 \cdot \mathbf{f}_i^{n-1} + \cdots + \delta_t \cdot \mathbf{f}_i^{n-t} \tag{6.26}$$

6.2.1 Prediction modelling

We modify the conventional state to $[\mathbf{R}^1, \mathbf{p}^1, ..., \mathbf{R}^n, \mathbf{p}^n, \mathbf{B}, \mathbf{c}]$.

$$\underset{\mathbf{R}^1,\mathbf{p}^1,...,\mathbf{R}^n,\mathbf{p}^n,\mathbf{B},\mathbf{c}}{\operatorname{argmin}} \quad E_{obs} + E_f + E_{ini} \tag{6.27}$$

Eq. (6.27) is the energy function for a visual deformable SLAM. E_{obs} is the sum error of camera to feature observations:

$$E_{obs} = \sum_{i=1}^{N} \sum_{j=1}^{F} [\mathcal{F}(\mathbf{R}^j, \mathbf{p}^j, \mathbf{f}_i^j) - \mathbf{m}_i^j]^2, \tag{6.28}$$

where $\mathbf{m}_i^j$ is the observation from camera to location of feature i in step j. $\mathcal{F}(\cdot)$ encodes the estimated observation from camera pose to feature position.

E_f denotes the error between current feature and its estimation from historical locations following Eq. (6.26):

$$E_f = \sum_{i=1}^{N} \sum_{j=1}^{t} (\mathbf{f}_i^{j+1} - \delta_1 \cdot \mathbf{f}_i^j - \delta_2 \cdot \mathbf{f}_i^{j-1} \quad - \cdots - \delta_t \cdot \mathbf{f}_i^{j-t})^2 \tag{6.29}$$

$$E_{ini} = \sum_{i=1}^{t} [\mathbf{p}^i - \mathbf{p}^0]^2 + \sum_{i=1}^{t} [\mathbf{R}^i \ominus \mathbf{R}^0]^2 \tag{6.30}$$

E_{ini} is to ensure the initial camera pose keeps static in the period size t. The notation $\ominus$ is called inverse retraction in differentiable geometry [98] and it is designed as a smooth mapping such that $\mathbf{R} = \mathbb{R} \ominus \mathbf{0}$. Similar to conventional static SLAM problem where the first pose need to be fixed [99], in our formulation the first period of poses should be fixed likewise.

Due to the field of view of the camera, some of them features may not be seen when the environment deforms. Fig. 6.2 shows one example of features not seen in some steps. Our approach is capable of processing this situation. If one feature is not fully observed any sliding window like the example shows, we will ignore this feature.

6.2.2 Observability analysis

In this section, we examine the parameter observability properties the proposed deformable SLAM formulation, which, for the time being, is considered as a parameter estimation problem. We will prove that the coefficient matrix $\mathbf{c}$ is not observable but the camera pose, as well as feature motions, are observable. This is a very satisfying result because

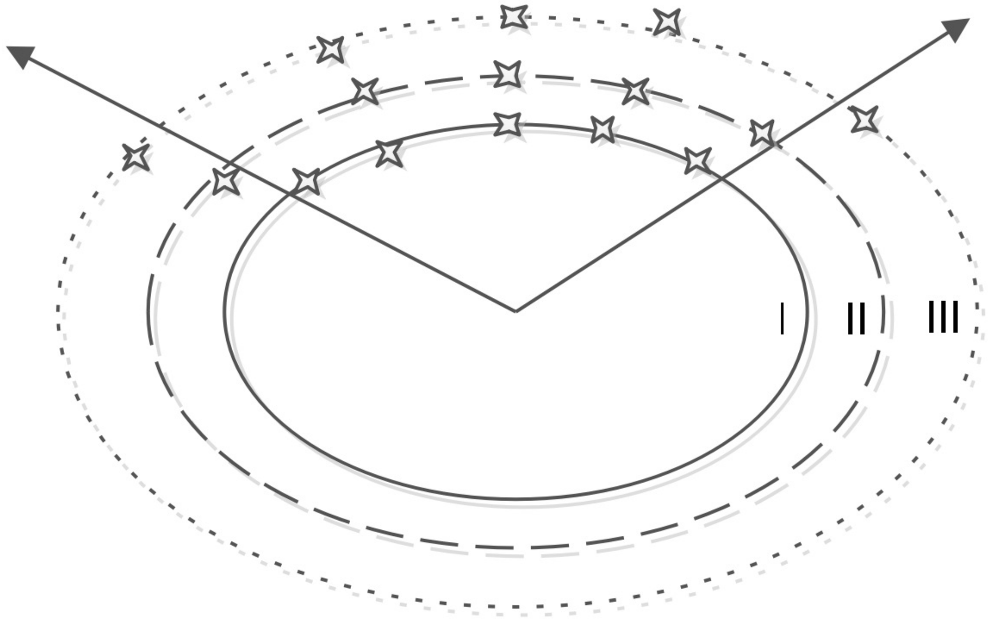


Figure 6.2: A typical feature deforming example. The ellipse deforms periodically depicted in 'I, II and III'. The region within arrows are the visible region. The leftmost feature is not observed in phase 'II' and 'III'. The rightmost feature is not observed in phase 'I'.

coefficient matrix $\mathbf{c}$ is only an auxiliary variable and is not physically explainable in a real scenario. Camera pose as well as feature motions, however, are physical processes and needs to be accurately estimated.

We adduce examples to prove coefficient matrix $\mathbf{c}$ is not observable. Taking into account the flexibility of presenting multiple period motions in Eq. (6.26), it will inevitably result in multiple solutions of feature motion combination. When features are static, the current shape of the environment will be passed to the next formulation which means all $\mathbf{c} = [1, 0\ , ...,\ 0]$. When there's only one periodic movement, the shape of the environment will be the same shape in history $\mathbf{c} = [0,\ , ...0, 1, 0, ...,\ 0]$. In a more general scenario, multiple periodic movements will lead to a full $\mathbf{c}$. We would like to emphasize that: The positions of features are not solvable. A simple example is when period is 2 but window is 4, this will be presented by $\mathbf{c} = [0, 1, 0, 0]$ or $\mathbf{c} = [0, 0.5, 0, 0.5]$.

In addition to qualitative analysis of observability, we gain a better understanding of the

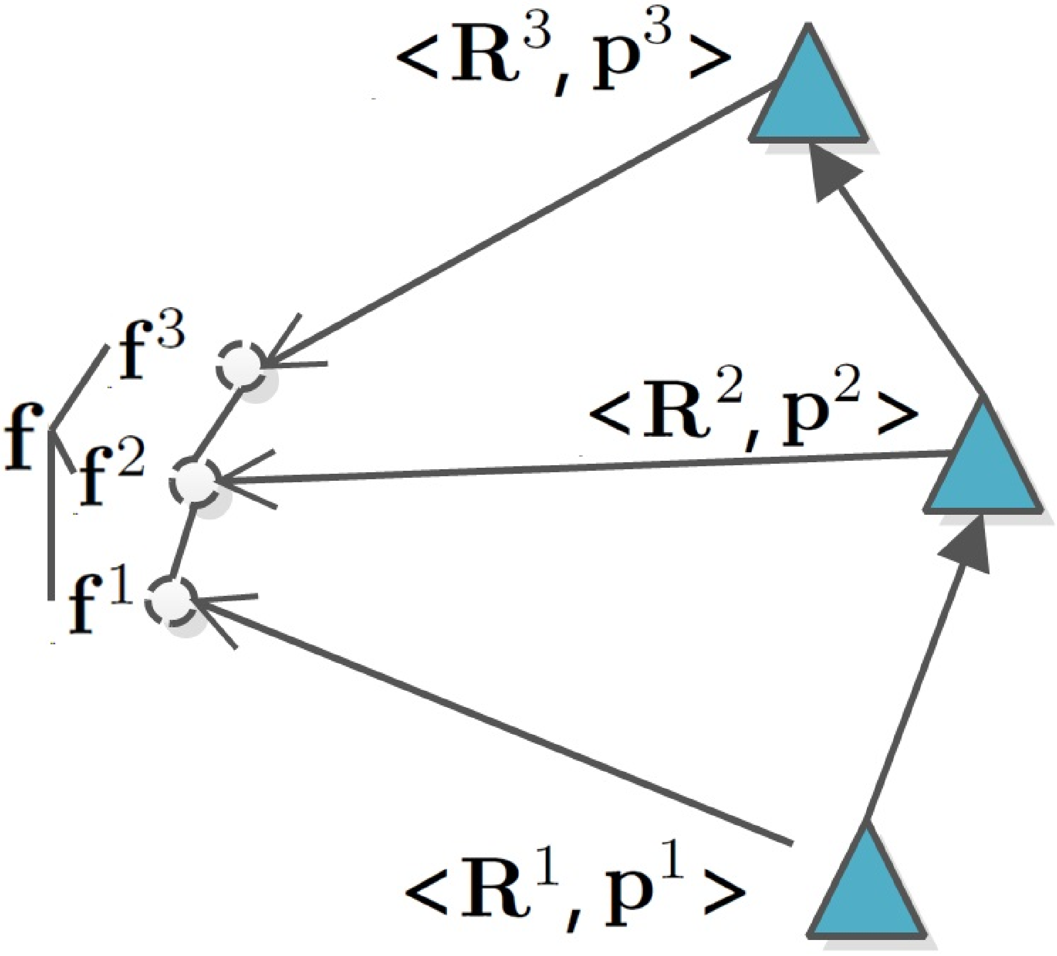


FIGURE 6.3: A simple example of 2 steps camera movement. Different from SLAM in rigid scenario, the feature $\mathbf{f}$ deforms in the space in position $\mathbf{f}^1$, $\mathbf{f}^2$ and $\mathbf{f}^3$.

formulation by proving with definition of observability. The study of parameter observability examines whether the information provided by the available measurements is sufficient for estimating the parameters without ambiguity; when parameter observability holds, the FIM is full rank and invertible [61]. From the stated example we have obtained the idea that the unobservable part lies in mismatch of real period and predefined window size. Therefore, we first prove this with simple scenario and extended to our conclusion. Consider the scenario shown in Fig. 6.3, one camera moves in three steps with orientation $\mathbf{R}^1$, $\mathbf{R}^2$, $\mathbf{R}^3$ and position $\mathbf{p}^1$, $\mathbf{p}^2$, $\mathbf{p}^3$. And it always observe one deforming feature with position $\mathbf{f}^1$, $\mathbf{f}^2$, $\mathbf{f}^3$. The observation is $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$. Window size t is set to 2. The residuals should be:

$$\mathrm{F}_{obj} = \begin{bmatrix} \mathbf{R}^1 \cdot (\mathbf{f}^1 - \mathbf{p}^1) - \mathbf{z}_1 \\ \mathbf{R}^2 \cdot (\mathbf{f}^2 - \mathbf{p}^2) - \mathbf{z}_2 \\ \mathbf{R}^3 \cdot (\mathbf{f}^3 - \mathbf{p}^3) - \mathbf{z}_3 \\ \mathbf{f}^3 - \delta_1 \cdot \mathbf{f}^1 - \delta_2 \cdot \mathbf{f}^2 \\ \mathbf{R}^1 \ominus \mathbf{I}_3 \\ \mathbf{R}^2 \ominus \mathbf{I}_3 \\ \mathbf{p}^1 \\ \mathbf{p}^2 \end{bmatrix} \tag{6.31}$$

Since $\ominus$ defines the distance of in the space of $SO(3)$, the first two orientation $\mathbf{R}^1$ and $\mathbf{R}^2$ is fixed (close to 3 identity matrix 1_3). The corresponding Jacobian of the toy model Eq. (6.31) is shown in Eq. (6.32). $S(\cdot)$ is the skew symmetric formulation. Therefore, the corresponding FIM matrix is Eq. (6.33). With regard to this scenario, after Gaussian elimination, Matrix H is full rank if the feature is moving ($\mathbf{f}^1$, $\mathbf{f}^2$ and $\mathbf{f}^3$ are not equivalent). However, considering the last 5×5 block of the matrix H, when feature is stable, all feature poses $\mathbf{f}^1$, $\mathbf{f}^2$ and $\mathbf{f}^3$ are equivalent and coefficients δ_1 and δ_2 are the same. In this case, H loses one rank thanks to the last two lines of matrix H. Thus, the only contribution to low rank lies in the last two lines of matrix H corresponding to variable coefficients δ_1 and δ_2 and is irrelevant to the number of features and number of steps. On the basis of these analysis we concluded that in general scenario, the low rank of Hessian is contributed by coefficients in the case of all features are stable.

$$\mathcal{J} = \begin{bmatrix} \Psi_1 & -\mathbf{R}^1 & & & & & \mathbf{R}^1 & & & \\ & & \Psi_2 & -\mathbf{R}^2 & & & & \mathbf{R}^2 & & \\ & & & & \Psi_3 & -\mathbf{R}^3 & & & \mathbf{R}^3 & \\ & & & & & -\delta_1 \cdot \mathbf{I}_3 & -\delta_2 \cdot \mathbf{I}_3 & \mathbf{I}_3 & -\mathbf{f}^1 & -\mathbf{f}^2 \\ -\mathbf{R}^1 & \mathbf{I}_3 & & & & & & & & \\ & & -\mathbf{R}^2 & \mathbf{I}_3 & & & & & & \end{bmatrix} \tag{6.32}$$

$$\Psi_i = -\mathbf{R}^i \cdot S(\mathbf{f}^i - \mathbf{p}^i)$$

$$
\boldsymbol{\mathcal{H}} = \begin{bmatrix}
2\mathbf{I}_3 & \mathbf{S}_1 & & & & & -\mathbf{S}_1 & & & & \\
2\mathbf{I}_3 & \mathbf{S}_1 & & & & & -\mathbf{S}_1 & & & & \\
\mathbf{S}_1 & 2\mathbf{I}_3 & & & & & -\mathbf{I}_3 & & & & \\
& & 2\mathbf{I}_3 & \mathbf{S}_2 & & & & -\mathbf{S}_2 & & & \\
& & \mathbf{S}_2 & 2\mathbf{I}_3 & & & & -\mathbf{I}_3 & & & \\
& & & & 2\mathbf{I}_3 & \mathbf{S}_3 & & & -\mathbf{S}_3 & & \\
& & & & \mathbf{S}_3 & 2\mathbf{I}_3 & & & -\mathbf{I}_3 & & \\
-\mathbf{S}_1 & -\mathbf{I}_3 & & & & & {\delta_1}^2\mathbf{I}_3 & \delta_1\delta_2\mathbf{I}_3 & -\delta_1\mathbf{I}_3 & \delta_1\mathbf{f}^1 & \delta_1\mathbf{f}^2 \\
& & -\mathbf{S}_2 & -\mathbf{I}_3 & & & \delta_1\delta_2\mathbf{I}_3 & {\delta_2}^2\mathbf{I}_3 & -\delta_2\mathbf{I}_3 & \delta_2\mathbf{f}^1 & \delta_2\mathbf{f}^2 \\
& & & & -\mathbf{S}_3 & -\mathbf{I}_3 & -\delta_1\mathbf{I}_3 & -\delta_2\mathbf{I}_3 & \mathbf{I}_3 & -\mathbf{f}^1 & -\mathbf{f}^2 \\
& & & & & & \delta_1{\mathbf{f}^1}^T & \delta_2{\mathbf{f}^1}^T & -{\mathbf{f}^1}^T\mathbf{I}_3 & {\mathbf{f}^1}^T * \mathbf{f}^1 & {\mathbf{f}^1}^T * \mathbf{f}^2 \\
& & & & & & \delta_1{\mathbf{f}^2}^T & \delta_2{\mathbf{f}^2}^T & -{\mathbf{f}^2}^T\mathbf{I}_3 & {\mathbf{f}^2}^T * \mathbf{f}^1 & {\mathbf{f}^2}^T * \mathbf{f}^2
\end{bmatrix}
$$

$$\mathbf{S}_i = S(\mathbf{f}^i - \mathbf{p}^i) \qquad (6.33)$$

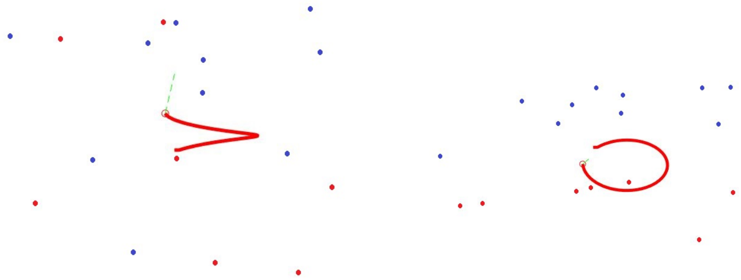
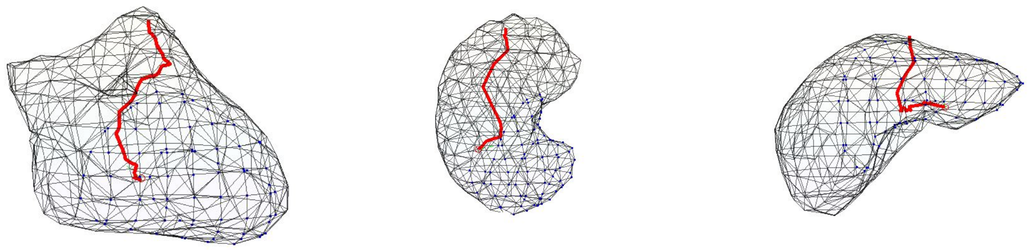


FIGURE 6.4: The two figures is an example of Monte Carlo simulation. Display area is illustrated from different directions for visualization.

6.3 Results and discussion

The last three chapters measure the accuracy by texture and shape of the reconstructed geometry. Pose accuracy is not discussed because they are robocentric formulation. This chapter enables accuracy comparisons of poses as well as features.



A deformable heart

Trajectory in a deformable stomach

Trajectory in a deformable liver

Figure 6.5: (a), (b) and (c) shows the camera moves randomly inside a deformable organ (Heart, stomach and liver). Red curves are the trajectories. Blue dots are the positions of the features and the attached quiver is the corresponding moving direction of each feature. Quiver only shows one step. Please refer to our video for the whole process.

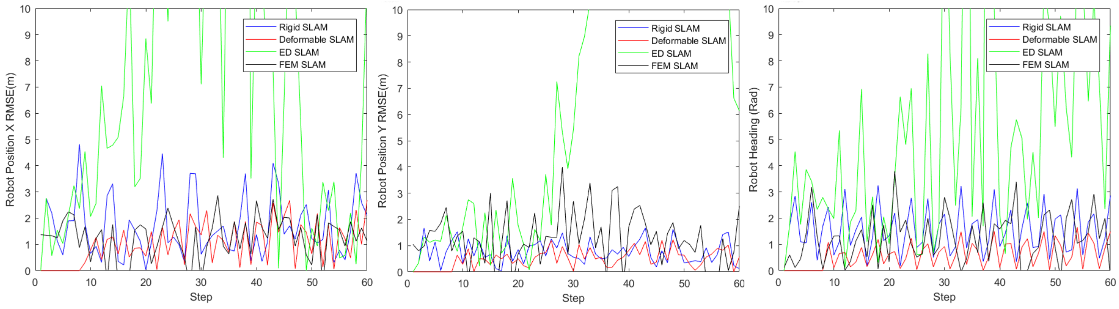


Figure 6.6: Estimation errors of static SLAM, FEM, ED graph and the proposed time-series SLAM. Row 1 to 3 are the tests on scenarios of heart, liver and left lung. Column 1 to 3 are the RMSE of camera position X, camera position Y and camera heading.

6.3.1 Monte Carlo simulations

In order to validate the proposed priori based approach as well as prove the unobservable camera pose in the deformable scenario, we conduct a series of Monte Carlo simulations under various conditions like different period of deformation, different movement of robots and different visibility of camera to feature observations. Fig. 6.4 is the typical 3D camera movements with 20 deforming features and 60 steps. The observation is defined as a feature position in camera coordinate which is a very common scenario of either stereoscope, lidar, RGB-D and stereo camera sensors. We adopt a deformation generator to simulate mixed kinematic deformations (different period and amplitude) of the features. The simulation size ranges 500×500 (mm). The camera moves in a predefined trajectory with 20 features deforming in a randomly mixed periodic way imitating soft-tissue movement. The viewing

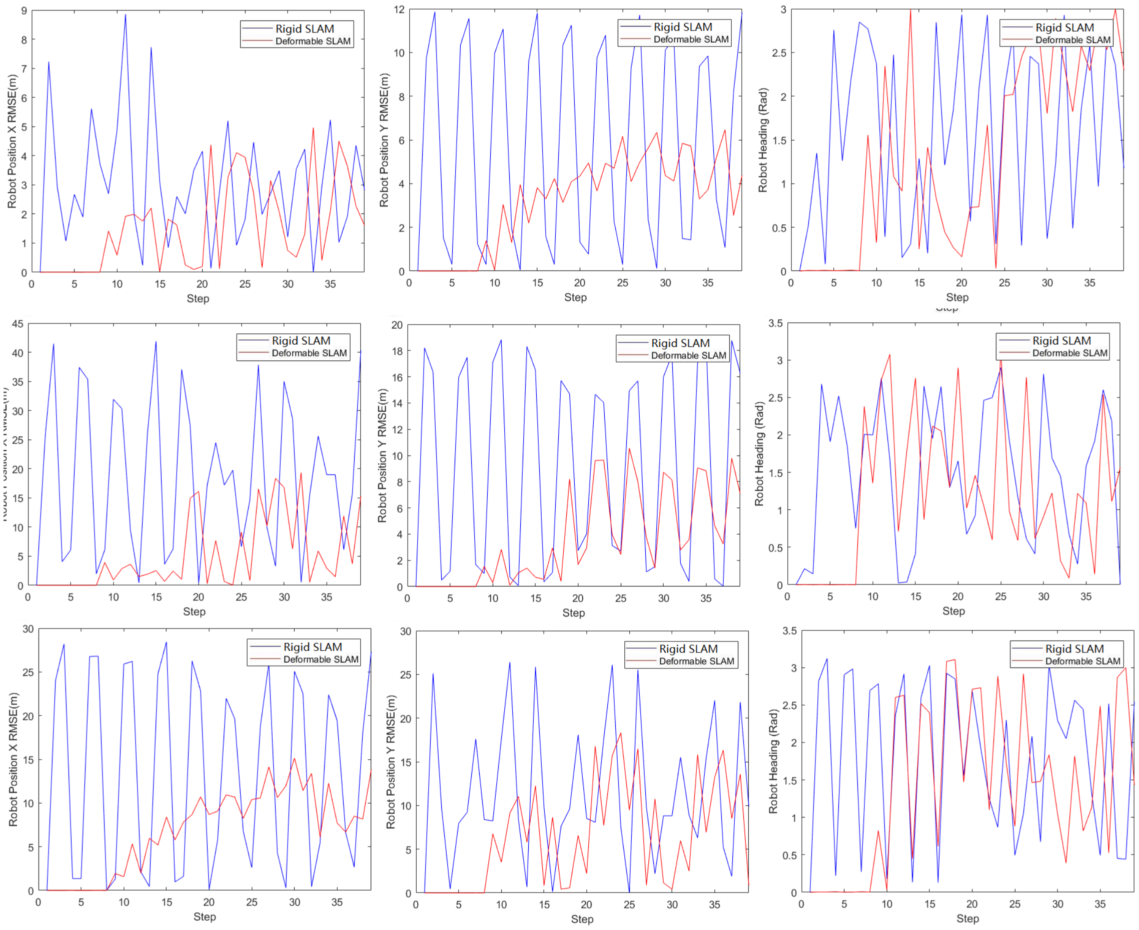


FIGURE 6.7: Estimation errors of static SLAM, FEM, ED graph and the proposed time-series SLAM. Row 1 to 3 are the tests on scenarios of heart, liver and left lung. Column 1 to 3 are the RMSE of camera position X, camera position Y and camera heading.

angle is also randomly chosen ranging from 30° to 90°. Noises are imposed on camera to feature observation ranging from 1 to 5 mm. In this test, we just focus on adverse camera-feature scenarios with a random motion to demonstrate the localization and tracking capability of the proposed estimation algorithm and ignore optimal camera path planning.

We conduct 50 Monte Carlo simulations and compare the proposed deformable SLAM, static SLAM approach and ED node based method. Note that different from the proposed method and static SLAM, ED graph based method is robocentric formulation estimating the rigid rotation and translation of the soft-tissue, which in turn can be regarded as the pose of the camera. Thus it serves as visual odometry making it inherently less accurate than the other two methods. Table 6.3 shows the comparisons. On the basis of these results, we concluded that the proposed deformable formulation outperforms static SLAM and ED based approach.

The results are compared by root mean squared errors (RMSE) which quantify the estimation accuracy. Fig. 6.6 is a typical Monte Carlo simulation showing the RMSE overtime.

6.3.2 SLAM in deformable soft-tissues

The proposed prior based deformable SLAM is also validated on ex-vivo experiments. In the simulation validation step, three different soft-tissue models (heart, liver and lung), which are segmented from a CT scan of a phantom, deform over time. The 3D deforming data are projected into 2D space and we simulate a camera moving inside each soft-tissues. The viewing angle of the camera is 60°. Fig. 6.4 shows the trajectory of the moving camera as well as the feature positions. The initial state of the feature and camera pose are estimated with traditional visual odometry. Fig. 6.7 presents the results of the three trajectories in the form of RMSE.

We also test the dataset on Hamlyn dataset 11 and 12. The camera remains stable (Fig. 6.8) observing two deforming soft-tissues. We track some key points and project them into 2D features to test if the estimated camera pose is stable. Results demonstrate that our algorithm achieves better camera pose (Average error 1.352 mm) than conventional SLAM (Average error 5.473 mm).

These results imply that the proposed priori attributes to outperforming conventional approaches, which results from the fact that many datasets conform to the mixture of historical shapes.

6.3.3 Observability test

To gain more insight into observability in the proposed SLAM system, we examine the parameter observability properties by testing the Hessian matrix of all the tests. The study of parameter observability is to analyze if a unique solution of the problem can be found; when parameter observability holds, the FIM is invertible [61]. We can gain insights about

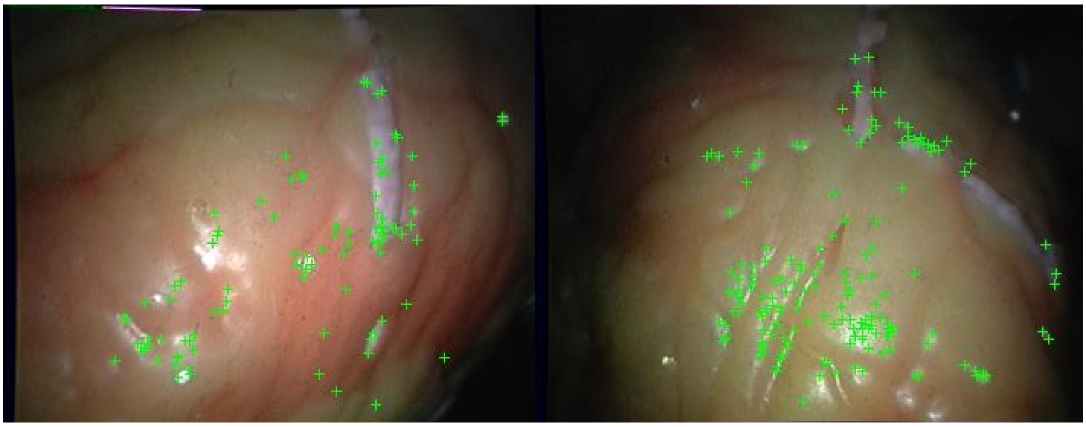


Figure 6.8: Ground truth dataset from Hamlyn center.

the null space due to the fact that FIM encapsulates all the information available. Section 6.2 shows that the null space of the proposed method lies in the deformable parameters **c**. After testing on all datasets, we find that the Hessian (FIM) has a nullspace of the size of **c**. We also test that Hessian becomes full rank when **c** is fixed. Therefore, even though **c** is not fully observable, the camera pose and feature positions are still unique. This test validate our theoretical analysis in Section 6.2.2.

Table 6.3: Feature estimation accuracies (m) in three models. All the simulation noises (invariances) are set to be 0.1 m.

	Heart	Stomach	Lung
Estimation error	1.242	2.120	3.197

6.4 Chapter summary

In this chapter, on the basis of previous robocentric SLAM scenario, our research extends the knowledge of observability analysis into deformable SLAM environment. We perform parameter observability analysis on ED parameterization and prove that in the case of no prior, the global pose is not separable from ED based deformation parameterization. Proofs of the existence of multiple solutions are provided for the ED based deformation formulation. The null space in both ED based formulation and static SLAM formulation makes the pose estimation not accurate. Based on our discussion, camera pose and the deforming environment in SLAM problems are entangled and cannot be estimated without priors.

To solve this, a new time series priori based algorithm is introduced for localizing camera as well as estimating the deformable environment, when robots operate in a dynamic scenario. We prove that the priori is enough to avoid ambiguity of rigid and non-rigid motions of the camera and the environment. The proposed algorithm is validated extensively on Monte Carlo simulations and medical datasets. It significantly outperforms conventional static SLAM formulation as well as ED formulation especially in a scenario with large and mixture of periodic deformations.

Chapter 7

Conclusions and future work

We study the problem of 3D non-rigid SLAM in MIS. This thesis starts with introducing ED graph to describe the free form soft surface deformation. For validating the effectiveness of ED graph, two deformable SLAM systems, robocentric template based SLAM and robocentric template free SLAM are proposed. Both systems demonstrate ED graph is capable to fully describe general deformation of soft-tissues. The template free SLAM is capable of incrementally rebuilding and deforming the soft surfaces in a surgical scenario with slow-moving stereo scope. Furthermore, to solve the robustness of robocentric template free SLAM, this thesis propose MIS-SLAM, the modified version of robocentric template free SLAM with modifications on: (1) A heterogeneous framework with GPU (dense robocentric deformable SLAM) and CPU (ORB-SLAM) (2) Highly integrated CPU and GPU modules for fast processing (3) An improved model point storage system and fusion management strategy (4) Real-time visualization. We show MIS-SLAM can process 3D deformable model reconstruction with relatively fast moving the stereoscope. To solve the problem in MIS-SLAM that the computational complexity in larger scale environment is $O(n^2)$, we classify the nodes into PR nodes and PI nodes and propose a two-level optimization strategy. In sacrifice small amount of accuracy, it successfully keeps the processing time to close $O(n^1)$.

Different from the three chapters (3, 4 and 5) describing dense robocentric deformable SLAM, the last chapter moves the theoretical works further toward conventional SLAM without assuming static scope as in robocentric scenario. Theoretical prove of unobservability of robot pose and ED graph is discussed and proved. This means the global rigid transformation and local non-rigid deformation of the robot is mixed. Thus prior is needed for robot pose separation. In this thesis, a prior based time series formulation is proposed and shows significant improvement than conventional rigid SLAM formulation, ED graph based SLAM and FEM based SLAM. A prove of observability is also provided on the time series approach.

There are some future directions that are natural extensions of this work. For the sake of clarity, we itemize them as follows:

(I) For template free SLAM approach, similar to the proposed stereoscope scenario, future work will also explore the feasibility of applying this research method on depth generated from a monocular scope with approaches like SfS [29] or deep neural network based depth recovery from monocular images [100].

(II) One shortcoming of MIS-SLAM is that it lacks dense loop closure which can close loop the whole soft-tissue. Future works will be how to design a better close loop module. ORB-SLAM uses sparse features to relocate camera based on the assumption that no relative motion exists in the environment. In the surgical vision, however, the deforming scenario makes the assumption invalid. More, hardware like EM sensors may be integrated for better scope pose initialization.

(III) For two-level ED optimization method, the node marginalization strategy in this thesis, however, is straightforward and arbitrary which only classify nodes based on the node-vertex connectivity. It's reasonable because different from the pose graph, ED graph is paralleled in GPU and time consumption requirement is more strict. But it remains to be of great interest to test if more complicated techniques like pose graph pruning method like Kullback–Liebler divergence minimization outperforms the proposed work while remains tiny consumption in GPU environment.

(IV) The time series approach proposed in this thesis requires heavy computation because of the large objective matrix, future work may exploit the structure of time series connection for reducing time and memory consumption.

Bibliography

[1] NDI. Electromagnetic tracking systems. `https://www.ndigital.com/products/electromagnetic-tracking-systems/`, 2019.

[2] Mingxing Hu, Graeme Penney, and Others. 3D reconstruction of internal organ surfaces for minimal invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 68–77. Springer, 2007.

[3] Peter Mountney and Guang-Zhong Yang. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1184–1187. IEEE, 2009.

[4] ETHICON. What is minimally invasive surgery. `http://www.ethicon.com/patients/learn-more/minimally-invasive-surgery`, 2017.

[5] Bingxiong Lin, Yu Sun, Xiaoning Qian, et al. Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2015.

[6] Danail Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 479–486. Springer, 2012.

[7] Nazim Haouchine, Jeremie Dequidt, Marie-Odile Berger, and Stephane Cotin. Monocular 3D reconstruction and augmentation of elastic surfaces with self-occlusion

handling. *IEEE transactions on visualization and computer graphics*, 21(12):1363–1376, 2015.

[8] Abed Malti, Adrien Bartoli, and Toby Collins. Template-based conformal shape-from-motion from registered laparoscopic images. In *MIUA*, volume 1, page 6, 2011.

[9] Xiaofei Du, Neil Clancy, et al. Robust surface tracking combining features, intensity and illumination compensation. *International journal of computer assisted radiology and surgery*, 10(12):1915–1926, 2015.

[10] Richard A Newcombe, Shahram Izadi, et al. KinectFusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[11] Michael Zollhöfer, Matthias Nießner, et al. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (TOG)*, 33(4):156, 2014.

[12] Richard A Newcombe, Dieter Fox, and Steven M Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[13] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016.

[14] Mingsong Dou, Sameh Khamis, et al. Fusion4d: real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016.

[15] Lena Maier-Hein, Peter Mountney, Adrien Bartoli, Haytham Elhawary, D Elson, Anja Groch, Andreas Kolb, Marcos Rodrigues, J Sorger, Stefanie Speidel, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical image analysis*, 17(8):974–996, 2013.

[16] Oscar G Grasa, Javier Civera, and JMM Montiel. EKF monocular SLAM with relocalization for laparoscopic sequences. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4816–4821. IEEE, 2011.

[17] Bingxiong Lin, Adrian Johnson, Xiaoning Qian, Jaime Sanchez, and Yu Sun. Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery. In *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pages 35–44. Springer, 2013.

[18] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. 3D shape recovery of deformable soft-tissue with computed tomography and depth scan. In *Proc Australasian Conf. Robot. Autom.*, pages 117–126, 2016.

[19] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters*, 3(1):155–162, 2018.

[20] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. MIS-SLAM: Real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing. *IEEE Robotics and Automation Letters*, 3(4):4068–4075, 2018.

[21] Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4): 14–24, 2010.

[22] Danail Stoyanov. Surgical vision. *Annals of biomedical engineering*, 40(2):332–345, 2012.

[23] William Lau, Nicholas Ramey, Jason Corso, Nitish Thakor, and Gregory Hager. Stereo-based endoscopic tracking of cardiac surface deformation. *Medical image computing and computer-assisted intervention–MICCAI 2004*, pages 494–501, 2004.

[24] Jedrzej Kowalczuk, Avishai Meyer, Jay Carlson, Eric T Psota, Shelby Buettner, Lance C Pérez, Shane M Farritor, and Dmitry Oleynikov. Real-time three-dimensional soft tissue reconstruction for laparoscopic surgery. *Surgical endoscopy*, 26(12):3413–3417, 2012.

[25] Johannes Totz, Stephen Thompson, Danail Stoyanov, Kurinchi Gurusamy, Brian R Davidson, David J Hawkes, and Matthew J Clarkson. Fast semi-dense surface reconstruction from stereoscopic video in laparoscopic surgery. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 206–215. Springer, 2014.

[26] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010.

[27] Miguel Lourenço, Danail Stoyanov, and Joao P Barreto. Visual odometry in stereo endoscopy by using pearl to handle partial scene deformation. In *Workshop on Augmented Environments for Computer-Assisted Interventions*, pages 33–40. Springer, 2014.

[28] Lin Zhang, Menglong Ye, Petros Giataganas, Michael Hughes, and Guang-Zhong Yang. Autonomous scanning for endomicroscopic mosaicing and 3d fusion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3587–3593. IEEE, 2017.

[29] Toby Collins and Adrien Bartoli. Towards live monocular 3d laparoscopy using shading and specularity information. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 11–21. Springer, 2012.

[30] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. Ieee, 2004.

[31] AD Milne, DG Chess, JA Johnson, and GJW King. Accuracy of an electromagnetic tracking device: a study of the optimal operating range and metal interference. *Journal of biomechanics*, 29(6):791–793, 1996.

[32] CGM Meskers, HM Vermeulen, JH De Groot, FCT Van der Helm, and PM Rozing. 3D shoulder position measurements using a six-degree-of-freedom electromagnetic tracking device. *Clinical biomechanics*, 13(4-5):280–292, 1998.

[33] Alfred M Franz, Tamas Haidegger, Wolfgang Birkfellner, Kevin Cleary, Terry M Peters, and Lena Maier-Hein. Electromagnetic tracking in medicine—a review of

technology, validation, and applications. *IEEE transactions on medical imaging*, 33 (8):1702–1725, 2014.

[34] Paschalis Panteleris and Antonis A Argyros. Vision-based SLAM and moving objects tracking for the perceptual support of a smart walker platform. In *European Conference on Computer Vision*, pages 407–423. Springer, 2014.

[35] Mingxing Hu, Graeme Penney, Daniel Rueckert, Philip Edwards, Fernando Bello, Roberto Casula, Michael Figl, and David Hawkes. Non-rigid reconstruction of the beating heart surface for minimally invasive cardiac surgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, pages 34–42, 2009.

[36] Toby Collins, Benoît Compte, and Adrien Bartoli. Deformable Shape-From-Motion in laparoscopy using a rigid sliding window. In *MIUA*, pages 173–178, 2011.

[37] Peter Mountney and Guang-Zhong Yang. Motion compensated SLAM for image guided surgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 496–504, 2010.

[38] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.

[39] Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and JMM Montiel. ORBSLAM-based endoscope tracking and 3D reconstruction. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 72–83. Springer, 2016.

[40] Nader Mahmoud, Alexandre Hostettler, Toby Collins, Luc Soler, Christophe Doignon, and JMM Montiel. SLAM based quasi dense reconstruction for minimally invasive surgery scenes. *arXiv preprint arXiv:1705.09107*, 2017.

[41] Mehmet Turan, Yasin Almalioglu, Helder Araujo, Ender Konukoglu, and Metin Sitti. A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots. *International journal of intelligent robotics and applications*, 1(4):399–409, 2017.

[42] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian Jun Zhang. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer methods and programs in biomedicine*, 158:135–146, 2018.

[43] Andres Marmol, Artur Banach, and Thierry Peynot. Dense-ArthroSLAM: dense intra-articular 3D reconstruction with robust localization prior for arthroscopy. *IEEE Robotics and Automation Letters*, 2019.

[44] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[45] Petros Giataganas, Christos Bergeles, Philip Pratt, Michael Hughes, Ara Darzi, and Guang-Zhong Yang. Intraoperative 3d fusion of microscopic and endoscopic images in transanal endoscopic microsurgery. In *The Hamlyn Symposium on Medical Robotics*, page 35, 2014.

[46] David B Kirk and W Hwu Wen-Mei. *Programming massively parallel processors: a hands-on approach*. Morgan kaufmann, 2016.

[47] Kai M Wurm, Armin Hornung, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: A probabilistic, flexible, and compact 3d map representation for robotic systems. In *Proc. of the ICRA 2010 workshop on best practice in 3D perception and modeling for mobile manipulation*, volume 2, 2010.

[48] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

[49] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *3D Vision-3DV 2013, 2013 International Conference on*, pages 1–8. IEEE, 2013.

[50] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. ElasticFusion: Dense SLAM without a pose graph. Robotics: Science and Systems, 2015.

[51] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635. IEEE, 2017.

[52] Matthias Innmann, Michael Zollhöfer, et al. VolumeDeform: Real-time volumetric non-rigid reconstruction - eccv 2016. `https://www.youtube.com/watch?v=khthUS7KVY4`, 2016.

[53] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, 2007.

[54] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

[55] Nicholas Carlevaris-Bianco, Michael Kaess, and Ryan M Eustice. Generic node removal for factor-graph SLAM. *IEEE Transactions on Robotics*, 30(6):1371–1385, 2014.

[56] Kevin Eckenhoff, Liam Paull, and Guoquan Huang. Decoupled, consistent node removal and edge sparsification for graph-based SLAM. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3275–3282. IEEE, 2016.

[57] Joan Vallvé, Joan Solà, and Juan Andrade-Cetto. Graph SLAM sparsification with populated topologies using factor descent optimization. *IEEE Robotics and Automation Letters*, 3(2):1322–1329, 2018.

[58] Olgierd Cecil Zienkiewicz, Robert Leroy Taylor, Perumal Nithiarasu, and JZ Zhu. *The finite element method*, volume 3. McGraw-hill London, 1977.

[59] Antonio Agudo, Begona Calvo, and JMM Montiel. FEM models to code non-rigid EKF monocular SLAM. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1586–1593. IEEE, 2011.

[60] Antonio Agudo, Begona Calvo, and JMM Montiel. 3D reconstruction of non-rigid surfaces in real-time using wedge elements. In *European Conference on Computer Vision*, pages 113–122. Springer, 2012.

[61] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[62] Dat Tien Ngo, Jonas Östlund, and Pascal Fua. Template-based monocular 3D shape recovery using laplacian meshes. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):172–187, 2016.

[63] Jose Lamarca and J.M.M. Montiel. Camera tracking for SLAM in deformable maps. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[64] Antoine Petit and Stéphane Cotin. Environment-aware non-rigid registration in surgery using physics-based simulation. In *Asian Conference on Computer Vision (ACCV) Workshops*, 2018.

[65] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3D shape from image streams. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 690–696. IEEE, 2000.

[66] Jing Xiao, Jin-xiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. In *European conference on computer vision*, pages 573–587. Springer, 2004.

[67] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1534–1541. IEEE, 2009.

[68] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-time combined 2D+ 3D active appearance models. In *CVPR (2)*, pages 535–542, 2004.

[69] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107 (2):101–122, 2014.

[70] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):878–892, 2008.

[71] Adrien Bartoli, Vincent Gay-Bellile, Umberto Castellani, Julien Peyras, Søren Olsen, and Patrick Sayd. Coarse-to-fine low-rank structure-from-motion. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[72] Minsik Lee, Chong-Ho Choi, and Songhwai Oh. A procrustean markov process for non-rigid structure recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1550–1557, 2014.

[73] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009.

[74] Jack Valmadre and Simon Lucey. General trajectory prior for non-rigid reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1394–1401. IEEE, 2012.

[75] Tomas Simon, Jack Valmadre, Iain Matthews, and Yaser Sheikh. Separable spatiotemporal priors for convex reconstruction of time-varying 3D point clouds. In *European Conference on Computer Vision*, pages 204–219. Springer, 2014.

[76] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In *Advances in Neural Information Processing Systems*, pages 55–63, 2014.

[77] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3D reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1542–1549, 2014.

[78] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, 71:428–443, 2017.

[79] Yu Gu, Fei Wang, Yanan Chen, and Xuan Wang. Monocular 3D reconstruction of multiple non-rigid objects by union of non-linear spatial-temporal subspaces. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, page 14. ACM, 2018.

[80] Antonio Agudo, Francesc Moreno-Noguer, Begona Calvo, and José María Martínez Montiel. Sequential non-rigid structure from motion using physical priors. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):979–994, 2016.

[81] Stefanie Wuhrer, Jochen Lang, and Chang Shu. Tracking complete deformable objects with finite elements. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 1–8. IEEE, 2012.

[82] Antonio Agudo and Francesc Moreno-Noguer. Force-based representation for non-rigid shape and elastic model estimation. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2137–2150, 2018.

[83] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007.

[84] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. In *ACM transactions on Graphics (TOG)*, volume 24, pages 1134–1141. ACM, 2005.

[85] HG Kenngott, JJ Wünscher, M Wagner, et al. OpenHELP (heidelberg laparoscopy phantom): development of an open-source surgical evaluation and training tool. *Surgical endoscopy*, 29(11):3338–3347, 2015.

[86] Christoph Schmalz, Frank Forster, Anton Schick, and Elli Angelopoulou. An endoscopic 3D scanner based on structured light. *Medical image analysis*, 16(5):1063–1072, 2012.

[87] Sven Haase, Jakob Wasza, Thomas Kilgus, and Joachim Hornegger. Laparoscopic instrument localization using a 3-D Time-of-Flight/RGB endoscope. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 449–454. IEEE, 2013.

[88] Mingsong Dou, Henry Fuchs, and Jan-Michael Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, pages 99–106. IEEE, 2013.

[89] Seth Billings, Nishikant Deshmukh, et al. System for robot-assisted real-time laparoscopic ultrasound elastography. In *SPIE Medical Imaging*, pages 83161W–83161W. International Society for Optics and Photonics, 2012.

[90] Jasper RR Uijlings, Arnold WM Smeulders, and Remko JH Scha. Real-time bag of words, approximately. In *Proceedings of the ACM international Conference on Image and Video Retrieval*, page 6. ACM, 2009.

[91] Stamatia Giannarou, Marco Visentini-Scarzanella, and Guang-Zhong Yang. Probabilistic tracking of affine-invariant anisotropic regions. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):130–143, 2013.

[92] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32 (11):1231–1237, 2013.

[93] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single RGB-D camera. *ACM Transactions on Graphics (TOG)*, 36(3):32, 2017.

[94] Elizabeth Yip and Richard Sincovec. Solvability, controllability, and observability of continuous descriptor systems. *IEEE Transactions on Automatic Control*, 26(3): 702–707, 1981.

[95] Antonio Agudo and Francesc Moreno-Noguer. Force-based representation for non-rigid shape and elastic model estimation. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2137–2150, 2017.

[96] Antonio Agudo and Francesc Moreno-Noguer. Robust spatio-temporal clustering and reconstruction of multiple deformable bodies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[97] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1272–1279, 2013.

[98] P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

[99] Teng Zhang, Kanzhi Wu, Jingwei Song, Shoudong Huang, and Gamini Dissanayake. Convergence and consistency analysis for a 3D invariant-EKF SLAM. *IEEE Robotics and Automation Letters*, 2(2):733–740, 2017.

[100] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.