Augmented Reality-assisted Surgical Guidance for Transnasal Endoscopy

by

Tong, Hon Sing

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Philosophy
at The University of Hong Kong

October 2022

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

Abstract of thesis entitled

Augmented Reality-assisted Surgical Guidance for Transnasal Endoscopy

Submitted by

Tong, Hon Sing

for the degree of Master of Philosophy

at the University of Hong Kong

October 2022

Augmented reality (AR), a technology that enables direct overlay of virtual images onto camera views, sparks a new opportunity to shape the future of the healthcare industry. Surgical ergonomics, efficiency and safety are expected to be further enhanced when compared with conventional surgical navigation. However, AR-assisted surgical guidance has not yet been adopted into mainstream clinical practice. One of the major concerns is the accuracy and stability of augmentation, which will not only cause visual disturbance but may also lead to unnecessary complications. Major factors affecting the accuracy and stability of AR-assisted guidance include but not limited to tracking modalities and the quality of patient's 3D anatomical models. Possible causes for sub-optimal tracking are electromagnetic (EM) tracking interference, optical tracking line-of-sight issues, and poor ergonomics due to bulky tracking tools. Regarding patient's 3D anatomical models, their accuracy varies based on scanning quality, reconstruction software and human operation. Augmenting poorly segmented models onto the endoscopic view gives rise to an observed depth that is not representative of the real surface during surgery. As a result, visual inconsistency may cause surgeon's fatigue. More severely, complications may arise if critical structures such as nerves and vessels are accidentally damaged. In light of these limitations, this thesis aims to explore innovative and alternative sensing solutions related to tracking and mapping in endoscopic procedures. The proposed approaches aim to provide more accurate, stable, and ergonomic AR-assisted guidance.

香港大學圖書館 HKU Libraries

First, a visual-strain fusion-based camera tracking method is proposed, such that reliable pose estimation is maintained even under adverse visual conditions such as presence of obstacles, complete darkness, and exaggerated lighting. Sparse strain measurement of a single-core fiber Bragg grating (FBG) fiber is utilized in an online learning process to estimate the tip pose of a soft manipulator. Simultaneously, an eye-in-hand mono-camera mounted at the soft manipulator tip is also utilized to estimate poses by simultaneous localization and mapping (SLAM). Sensing fusion is then performed between the FBG-derived pose and SLAM-derived pose to give robust pose feedback. Pose estimation experiments were performed in a LEGO® scenario. The mean estimation error was reduced from 3.116 mm to 1.324 mm when fusion was used in comparison to pure estimation by SLAM.

Second, a monocular depth estimation method is proposed, with the aim to obtain depth information *in-situ* without relying on pre-operatively segmented 3D anatomical models. A virtual endoscopic environment is utilized to train a supervised depth estimation network. During application, a generative adversarial network (GAN) first transfers image style from the real endoscopic view to a synthetic-like view. Next, the trained depth estimation network predicts framewise depth from the synthetic-like images in real-time. Regarding accuracy evaluation, framewise depth was predicted from images captured from within a nasal airway phantom and compared with ground truth, achieving a structural similarity (SSIM) value of 0.8310 ± 0.0655. In addition, 3D annotation on the endoscopic view was performed with the nasal airway phantom. The annotations created can anchor stably onto target anatomical surfaces even with camera movement.

(493 words)

香港大學圖書館 HKU Libraries

Declaration

I hereby declare that this whole dissertation titled *Augmented Reality-assisted Surgical Guidance for Transnasal Endoscopy* is my own work, except the parts with due acknowledgement, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed: _____

Tong, Hon Sing

香港大學圖書館 HKU Libraries

Acknowledgements

Three years ago when I was a final year undergraduate, while all my peers started sending CVs and attending interviews, I sent an email to Dr. Ka-Wai Kwok and asked for an opportunity for lab shadowing in IRIS. At that moment, I did not have a single clue about what "research" felt like, nor did I have any spectacular skills in robotics. Out of instinct, I just "felt like" I wanted to know what research and development is like in the field of robotics. And three years later, I am about to graduate again, but as a mechanical engineering MPhil this time.

I am grateful that Dr. Kwok granted me such a precious chance to be a part of IRIS. Not only did I strengthen my technical and problem-solving skills, but more importantly, I learnt to become a "better me" in terms of team collaboration, mental strength and being passionate. At the beginning of my study, Dr. Kwok once mentioned, that "the path of research is not easy". After overcoming all difficulties, I am proud to say I do not regret starting this path. Other than the patient guidance and advice given by Dr. Kwok, substantial support from lab members and collaborators was also an indispensable factor that guided me through this path. These individuals include but not limited to Mr. Justin Ho, Dr. Jason Chan, Dr. Xiaomei Wang, Mr. Alan Tang, Mr. Owen Ng, Mr. Jason Mak, Mr. Mengjie Wu, Mr. Liam He, Dr. Ge Fang, Dr. Kui Wang, Mr. Jing Dai, Miss Yingqi Li, Miss Becky Chung, Mr. Ian Sun and Miss Yun Lin Lew.

I also thank my parents and other family members who provide me with unconditional love. I would not have been able to pursue this MPhil degree without your care and support. Special thanks to my girlfriend Stella Ting, who accompanied me in this journey. Finally, this thesis is a gift to my grandfather who passed away at the beginning of my MPhil study. It would have been a more beautiful autumn if you could witness my graduation.

Chris Tong

香港大學圖書館 HKU Libraries

Table of Contents

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

List of Figures

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

香港大學圖書館 HKU Libraries

List of Table

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
6D	Six Dimensional
AC	Alternating Current
AOS	Algebraic One Step
AR	Augmented Reality
CG	Computer Graphics
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
cycleGAN	Cycle Generative Adversarial Network
DC	Direct Current
DiL	Dictionary Learning
DNN	Deep Neural Network
DoF	Degree of Freedom
DRF	Dynamic Reference Frame
ELM	Extreme Learning Machine
EM	Electromagnetic
ENT	Ear, Nose and Throat
ESS	Endoscopic Sinus Surgery
FBG	Fiber Bragg Grating
FDA	Food and Drug Administration
FLE	Fiducial Localization Error
fMRI	Functional Magnetic Resonance Imaging
FOV	Field-of-view
FRE	Fiducial Registration Error
GAN	Generative Adversarial Network
GPS	Global Positioning System
GPU	Graphics Processing Unit
HMD	Head-mounted Display

香港大學圖書館 HKU Libraries

IGS	Image-guided Surgery
IMU	Inertial Measurement Unit
IR	Infrared
LED	Light-emitting Diode
LSTM	Long Short-term Memory
MAE	Mean Absolute Error
MIS	Minimally Invasive Surgery
MP	Moore-Penrose
MRI	Magnetic Resonance Imaging
NRMSE	Normalized Root-mean-square Error
OFDR	Optical Frequency Domain Reflectometry
OST	Optical See-through
PET	Positron Emission Tomography
RAM	Random Access Memory
RBF	Radial Basis Function
RGB	Red-green-blue
RMS	Root-mean-square
RMSE	Root-mean-square Error
RPE	Reprojection Error
SfM	Structure-from-motion
SLAM	Simultaneous Localization and Mapping
SSIM	Structural Similarity
SVD	Singular Value Decomposition
TME	Total Mesorectal Excision
TRE	Target Registration Error
UDP	User Datagram Protocol
US	Ultrasound
VINS	Visual Inertial System
VO	Visual Odometry
vSLAM	Visual Simultaneous Localization and Mapping
WDM	Wavelength Division Multiplexing

香港大學圖書館 HKU Libraries

CHAPTER 1
INTRODUCTION

1.1 MOTIVATION AND OBJECTIVES

Advancements in tracking technology and navigation systems have changed medical care in the last decades. Surgical navigation systems can not only enhance anatomical knowledge and treatment experience of surgeons, but also bring about a higher level of precision and accuracy, potentially improving working efficiency, safety and lowering the cost of a surgical procedure. Augmented reality (AR), a technology that enables direct overlay of virtual images onto camera views, sparks a new opportunity to shape the future of the healthcare industry. By incorporating AR into conventional surgical navigation, extra information such as visualization of subsurface critical structures, pre-operatively planned surgical paths, and surgical annotation can be fused with the endoscopic view. AR-assisted surgical guidance inherits benefits brought by conventional surgical navigation, and simultaneously introduces new possibilities.

However, AR-assisted surgical guidance has not been adopted into mainstream clinical practice. Issues like poor depth perception and visual cluttering might have prevented the adoption of this technology. Spatial and temporal misalignment between virtual objects and physical anatomy may cause fatigue due to visual inconsistency. Most importantly,

香港大學圖書館 HKU Libraries

complications may arise if critical structures such as nerves and vessels are accidentally damaged, or if targeted tissue is not adequately resected. In fact, a system's accuracy highly depends on tracking modalities, quality of patient's 3D anatomical models, and registration techniques. Electromagnetic (EM) tracking and optical tracking are two commonly employed tracking modalities. Although both have promising accuracy and reliability in ideal situations, EM tracking accuracy may drastically deteriorate in the presence of magnetic interference while optical tracking suffers from the line-of-sight issues. Next, the accuracy of 3D anatomical models varies based on imaging quality, reconstruction software and human operation. Augmenting poorly segmented models onto the endoscopic view gives rise to an observed depth that is not representative of the real surface during a surgery. In case of poor segmentation, even with ideal tracking and accurate registration, misalignment between virtual objects and physical anatomy would still be observed.

Therefore, the objective of this work is to explore innovative sensing alternatives that might benefit tracking and mapping during an endoscopic procedure, eventually leading to a more accurate and stable AR-assisted guidance system. In particular, machine learning plays a major role in the proposed methods. Main contributions of this work are as follows:

Visual-strain fusion for camera tracking	- Online learning-based pose estimation using sparse strain measurement of single-core fiber Bragg grating (FBG) fiber. - Sensing fusion between mono-camera simultaneous localization and mapping (SLAM) and FBG-derived localization information. - Experimental validation of the proposed sensing fusion method under normal and poor visual conditions.
Real-to-virtual domain transfer-based depth estimation	- Monocular depth estimation for achieving real-time AR in surgical guidance. - Supervised depth estimation network trained entirely in a virtual environment and used to predict depth from endoscopic images in real-time. Cycle Generative Adversarial Network (cycleGAN)-based real-to-virtual style transfer is implemented on endoscopic images.

香港大學圖書館 HKU Libraries

- Predicted depth evaluation against ground truth depth in a nasal airway phantom.

- Overall system stability assessment in terms of temporal alignment and depth consistency.

1.2 THESIS ORGANISATION

Chapter 2 presents an overview of the development and the state-of-the-art of both conventional surgical navigation and AR-assisted guidance. Fundamental technical components that constitute the technologies are introduced, which include general tracking modalities, tool calibration, registration methods, and error quantification methods.

Chapter 3 focuses on the development of an online learning-based pose estimation method for an eye-in-hand camera. This method involves sensing fusion between SLAM-based pose estimation and FBG-derived localization information. Experimental validation shows that pose estimation from this visual-strain fusion strategy can give promising results under both normal and poor visual conditions. This work is based on the co-authored paper "*Learning-based Visual-Strain Fusion for Eye-in-hand Soft Robot Pose Estimation and Control"*.

Chapter 4 introduces a monocular depth estimation method for achieving 3D annotations in transnasal surgery. During the training phase, a virtual endoscopic environment is utilized to train a supervised depth estimation network. During the testing phase, real endoscopic views are style-transferred to synthetic-like views before an image is input into the depth estimation network, wherein framewise depth can be obtained in real-time. This work is based on the first-authored paper "*Real-to-Virtual Domain Transfer-based Depth Estimation for Real-time 3D Annotation in Transnasal Surgery: A Study of Annotation Accuracy and Stability"*.

Chapter 5 provides a conclusion of this thesis, including future research directions.

香港大學圖書館 HKU Libraries

1.3 RESEARCH ACCOMPLISHMENTS IN STUDY PERIOD

First-authored paper

- H.S. Tong, Y.L. Ng, Z. Liu, J.D.L. Ho, P.L. Chan, J.Y.K. Chan, and K.W. Kwok, "*Real-to-Virtual Domain Transfer-based Depth Estimation for Real-time 3D Annotation in Transnasal Surgery: A Study of Annotation Accuracy and Stability*", **International Journal for Computer Assisted Radiology and Surgery (IJCARS)**. 2021;16(5):731-739.

Co-authored paper

- X. Wang, J. Dai, H.S. Tong, K. Wang, G. Fang, X. Xie, Y.H. Liu, S. K.W. Au, and K.W. Kwok, "*Learning-based Visual-Strain Fusion for Eye-in-hand Soft Robot Pose Estimation and Control*", **IEEE Transactions on Robotics (T-RO)**
  (Submitted for 3rd revision)

- Z.L. He, J. Dai, H.S. Tong, G. Fang, J.D.L. Ho, L.Y. Liang, H.C. Chang and K.W. Kwok, "*Design of MRI-guided Robot with Soft Fluidic-driven Actuator for Bilateral Stereotactic Neurosurgery*".
  (In preparation)

Filed patent

- "*Patient-specific maxillary template for surgical navigation*"
  US Provisional Pat.: US 63/123,506 (Filed on 10 Dec 2020)
  Inventors: K.W. Kwok, J.Y.K. Chan^, J.D.L. Ho, H.S. Tong

Poster presentation

- C.L. Lam, C.I. Lam, H.L. Leung, H.S. Tong, J.D.L. Ho, K.W. Kwok, and J.Y.K. Chan, "*A Novel Augmented Reality Navigation System for Flexible Endoscopic Sinus Surgery*", **Hong Kong College of Otorhinolaryngologists Annual Scientific Meeting 2021**
  (President Prize for Best Poster Presentation)

香港大學圖書館 HKU Libraries

CHAPTER 2
STATE-OF-THE-ART OF AUGMENTED REALITY IN SURGICAL NAVIGATION

2.1 INTRODUCTION

Augmented reality (AR) is a technology that enables the fusion of virtual images with the real world [1]. Unlike virtual reality (VR), computer graphics (CG) in AR are intended for enhancing the natural vision of a user instead of entirely replacing it. While AR has been applied to alter the user experience of smartphones and games, there are also possibilities in changing the medical industry, potentially improving surgical safety and efficacy in the near future [2]. In particular, AR can be incorporated into conventional surgical navigation to become AR-assisted surgical guidance. Surgeons are required to pay attention to information displayed on different monitors during conventional surgical navigation. When an extra AR module is incorporated, surgeons may focus more on the surgical site because patient-specific information can be augmented onto the endoscopic view or on a head-mounted display (HMD). In addition, patient scans obtained pre-operatively are presented in a 2D manner in conventional surgical navigation. In AR-assisted surgical guidance, 2D scans are processed and presented in the form of 3D anatomical models in the endoscopic view, providing intuitive spatial orientation in the surgeon's perspective. This chapter first gives an overview of conventional surgical navigation, which is the prerequisite for AR-assisted guidance. Next, state-of-the-art in AR-assisted surgical guidance is introduced, as well as basic technical components constituting these technologies.

香港大學圖書館 HKU Libraries

2.1.1 Surgical Navigation

The development of medical imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), or even functional imaging such as functional MRI (fMRI), and positron emission tomography (PET) are examples that mark the modern era of healthcare. These imaging technologies have enriched our understanding of target anatomies and pathologies and have become indispensable parts of any surgical navigation system by providing patient-specific "maps" of the target anatomy. Furthermore, they form the basis for positioning any surgical tool or anatomical target during a surgery [3].

Sukegawa *et al.* [4] has made an appropriate analogy between surgical navigation and global positioning system (GPS), as illustrated in **Fig. 2.1a** and **Fig. 2.1b**. GPS localizes your phone or vehicle and displays it on a geographical map in real-time, analogous to showing the location of a surgical instrument with respect to CT or MRI scans during a surgery. Extra information such as critical areas marked as "no-fly zone" and a path leading to the target anatomy/destination can also be marked prior to surgery and displayed intra-operatively. Therefore, patient scans and a real-time tracking system are the essential building blocks to achieving surgical navigation. Next, a registration process is required to spatially correlate the scans and the patient before beginning navigation. Without registration, the navigation system would not be able to localize the instruments with respect to the patient's frame of reference.

Most surgical navigation systems have an assumption that the body being tracked is a rigid body [5]. To elaborate, it is assumed the body does not change in shape or position during a pre-operative scan or during a surgery. Subsequently, a navigation system is commonly applied to anatomical sites that are relatively rigid, which include but not limited to skull-base surgeries, spinal surgeries, and orthopaedics [3]. For example, neurosurgery is one of the pioneering areas that incorporates navigation. It has been used to assist brain tumor resection more than two decades ago [6-8]. It is reasonable for neurosurgery to be one of the first medical specialties to adopt navigation because the brain is encased by a rigid skull, making deformation minimal during imaging and surgery. Also, there is a demand for high precision to avoid damage of healthy tissue. As depicted in **Fig. 2.1b**, a conventional surgical navigation display visualizes the position of a biopsy tool tip on the coronal, sagittal and axial views, which are the tri-planar views that surgeons would usually refer to during diagnosis and treatment.

香港大學圖書館 HKU Libraries

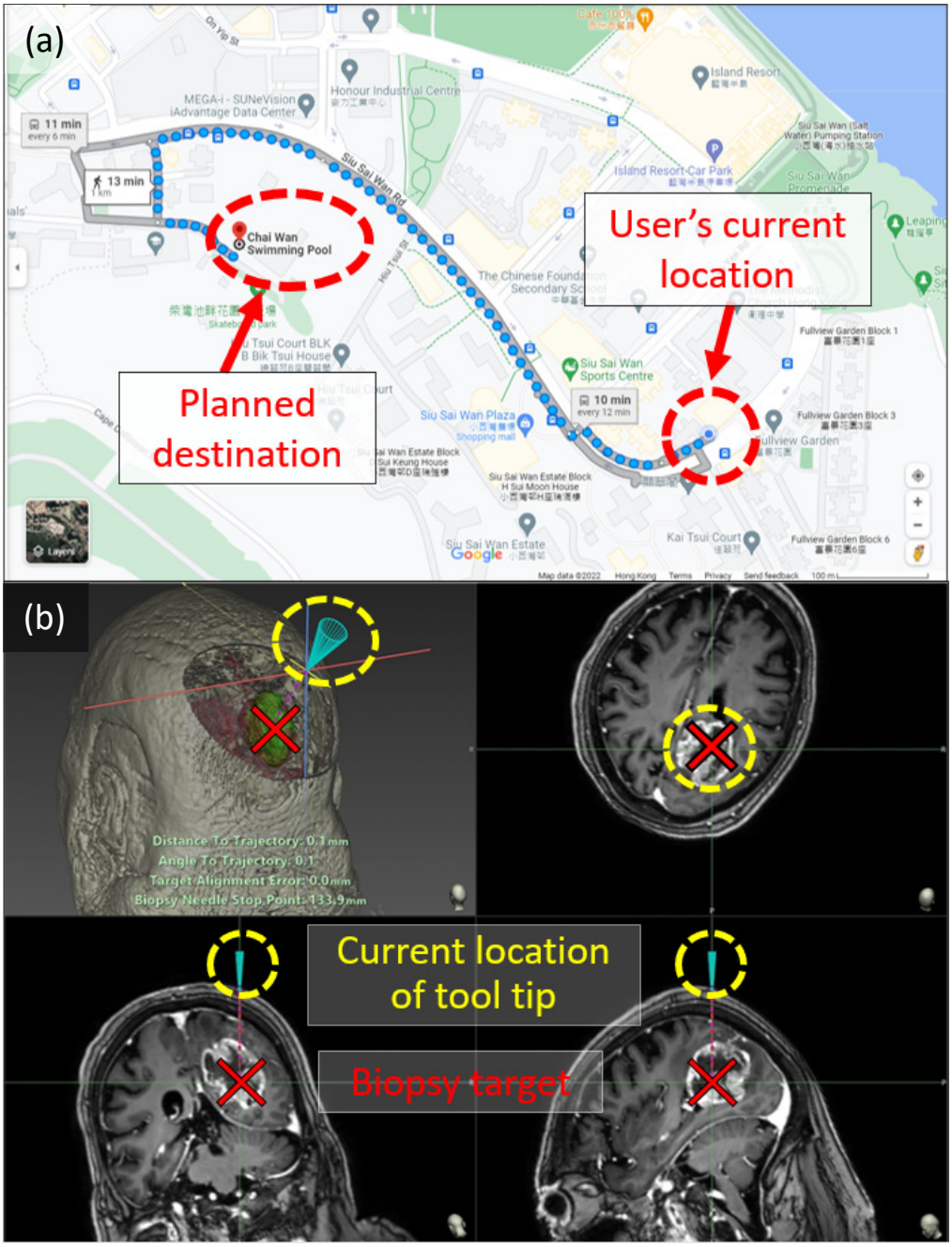

Fig. 2.1 Analogy between global positioning system (GPS) and conventional surgical navigation. **(a)** GPS locates the user's current geographical location on a map and guides the user to a planned destination. **Image source: Google Map. (b)** Navigation-assisted brain biopsy. Biopsy tool tip and target are localized within MRI scans of the patient in real-time. **Image source:[9].**

Ear, nose and throat (ENT) surgery, like neurosurgery, also has a demand for surgical instrument localization. Since the early 1990s, instrument navigation has been a critical tool for ENT surgery [10]. ENT surgery is often characterized by the presence of delicate bony soft tissue structures, narrow spaces and close proximity to critical structures such as the carotid arteries and optic nerves [11]. By knowing the instrument locations in real-time, surgeons can work with an expanded comfort zone. For instance, in a study of the influence

香港大學圖書館 HKU Libraries

posed by surgical navigation on sinus surgery performance, Reardon *et al.* [12] revealed that surgeons tend to have more sinuses reached when navigation was used. Confidence level of surgeons becomes higher as they are provided with extra spatial information that reduces the chance of damaging critical structures. The Medtronic StealthStation™ ENT, as depicted in **Fig. 2.2a**, is an example of a navigation system for anatomy and instrument localization. The system implements EM tracking technology to measure the pose of both the instruments and the patient, as shown in **Fig. 2.2b** and **Fig. 2.2c**. Compared with first-generation systems, recent products such as the StealthStation™ ENT have improved user-friendliness in terms of registration, and are robust in terms of instrument localization accuracy. However, core functionalities are essentially the same [10].

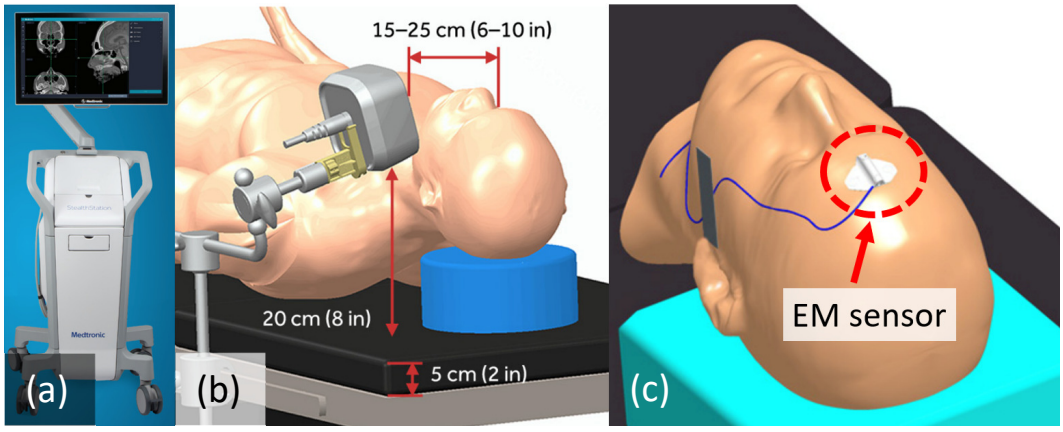

Fig. 2.2 **(a)** StealthStation™ ear, nose and throat (ENT) navigation system by Medtronic, which implements electromagnetic (EM) tracking. **(b)** Bedside-mounted EM field generator placed 15-25 cm from the patient's head, **(c)** EM sensor attached to the forehead of the patient, which tracks movement of the head. **Image source: Medtronic.**

Compared with neurosurgery and otolaryngology, spinal and orthopaedic surgeries are even more suitable candidates for navigation because of the high rigidity of bones. Due to precise registration, optimal navigation precision can be achieved [5]. A product that exemplifies this is the Mazor X StealthEdition™ by Medtronic, as depicted in **Fig. 2.3**. It is a robot-guided platform that performs tracking of the instrument by calculating the robot forward kinematics. Khan [13] has provided evidence of the efficacy of pedicle screw placement using the Mazor X, showing a screw placement accuracy of 99.5%, where 189 out of 190 screws were placed with Ravi Grade I accuracy (completely within the pedicle) and 1 screw with Ravi Grade II accuracy (<2mm pedicle wall breach) [14]. After acquisition of Mazor X by Medtronic, an extra optical tracking system was added to create a "double safety net" for navigation accuracy in the cases that localization by robot kinematics fails.

香港大學圖書館 HKU Libraries

Fig. 2.3 Mazor X StealthEdition™ by Medtronic, a navigation system for spinal and orthopaedic surgery. Surgical instrument localization is achieved by i) robot kinematics and ii) optical tracking of reflective markers. **Image source: Medtronic.**

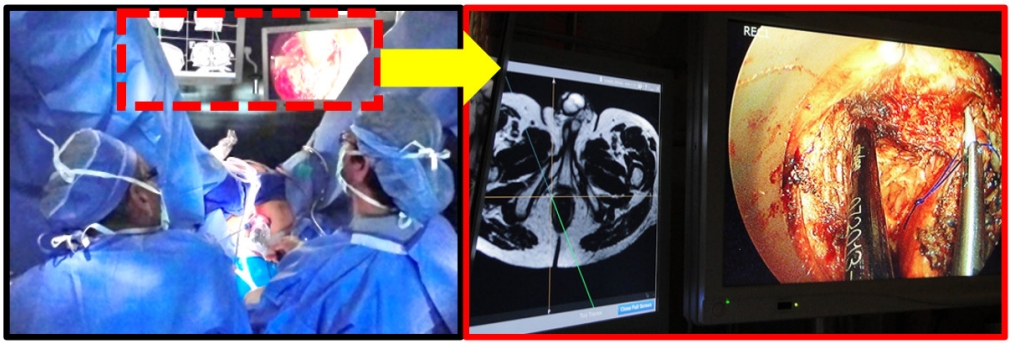

Fig. 2.4 Transanal total mesorectal excision (TME) assisted by surgical navigation. Surgeons are required to pay attention to two monitors at the same time. Monitor on the left shows the position of the laparoscopic dissector tip with respect to the patient MRI scan. Monitor on the right shows the laparoscopic view. **Image source: [15].**

The application of surgical navigation has progressed even further in recent years. In 2015, Atallah *et al.* [15] reported the first frameless stereotactic navigation for transanal total mesorectal excision (TME) as illustrated in **Fig. 2.4**. Tracking modality employed in this study was optical tracking by a stereoscopic infrared camera. Compared with upper abdominal organs, the pelvis is less affected by respiratory and pneumoperitoneum movement [3], making it a suitable site for performing surgical navigation. In this study, 3 patients underwent TME, measuring a navigation accuracy of ± 3.69 mm. Although the study recorded a 47-minute increase in case time mainly due to navigation setup, the author concluded that the navigation technique is beneficial for i) maintaining an accurate dissection plane, ii) preventing the damage of critical anatomical structures, and iii) monitoring the dissection progress.

香港大學圖書館 HKU Libraries

Regardless of the advancement of surgical navigation mentioned above, potential risk of navigation inaccuracy still exists. In 2017, Food and Drug Administration (FDA) issued an analysis of surgical navigation, reporting events in which navigation systems experienced errors that have led to complications, prolonged procedures, or even death [16]. Therefore, a surgical navigation system should only be considered an assistive tool that aids with surgeon's decision making. It is believed that as long as healthcare providers are aware of the possibility of navigation error occurrence, risks should be outweighed by the benefits that surgical navigation brings [3]. Surgical navigation is an advancement that emerged decades ago, with both industry and the medical community aiming to create further breakthroughs in this technology. A possible enhancement in the near future is AR, enabling the direct overlay of extra visual information onto the surgical site.

2.1.2 Augmented Reality (AR)-assisted Surgical Guidance

In the previous section, we discussed that surgical navigation requires i) the patient's pre-operative scan which acts as a map; 2) tracking modalities which act as GPS to track both the instruments and the patient's anatomy; and 3) a registration process that registers patient scans with the surgical site. With the combination of these well-developed tools and methods, AR-assisted surgical guidance can be realized.

To prepare for AR application in surgeries, pre-operative imaging such as CT and MRI are usually performed, such that digitalization of the patient's anatomy can be obtained. Next, segmentation is performed to derive 3D patient anatomy from 2D pre-operative scans. Anatomical information in the form of CG can then be superimposed on surgical video that is captured with an endoscope and displayed on a monitor. Medical AR performed in this manner is known as video-based AR. An example is illustrated in **Fig. 2.5a**, which shows the overlay of sub-surface blood vessels in a transoral endoscopic surgery [17], and in **Fig. 2.5b**, which shows the overlay of the kidney in a partial nephrectomy [18]. Another category of medical AR is optical see-through (OST) AR, as illustrated in **Fig. 2.5c**. Instead of involving images streamed from an endoscope, OST AR requires the user to wear a head-mounted display (HMD) during a surgery, which is an "eyeglasses-like" device that has special projectors for displaying CG on the see-through glasses. Both types of medical AR differ from conventional surgical navigation by requiring extra calibration processes. These calibration processes are essential for i) characterising intrinsic parameters of the

香港大學圖書館 HKU Libraries

endoscope or camera involved, and ii) finding the spatial offset between the camera optical center and the tracking sensor attached on it.

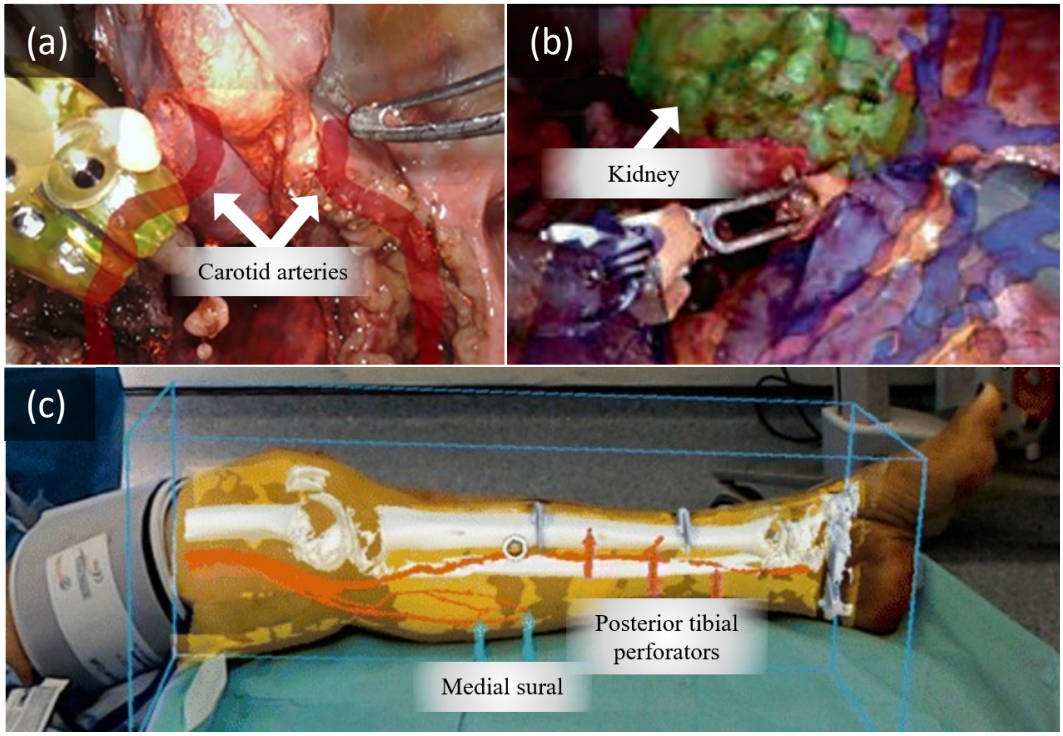

Fig. 2.5 Examples of AR-assisted surgical guidance in different medical specialties. **(a)** Underlying carotid arteries visualized in a transoral endoscopic surgery. **Image source: [17]**; **(b)** Kidney visualized in a partial nephrectomy. **Image source: [18]**; **(c)** Bones and blood vessels visualized using an orthopaedic surgical guidance system. **Image source: [18].**

Similar to conventional surgical navigation, AR also allows the fusion of information from different modalities like CT, MRI and US, enabling pre-operative analysis and planning. However, its major difference from conventional navigation is that, as revealed in **Fig. 2.6**, the surgeon does not need to redirect his/her eyes constantly between the surgical site and the intra-operative guidance display during the surgery [5, 10], resulting in improved ergonomics. In addition, AR may visualize subsurface critical structures like nerves, blood vessels, major organs [2, 19] and pre-operatively planned surgical trajectories, potentially leading to higher surgical efficiency, improved surgeon confidence and lower risk of complications [11, 20-22]. AR also brings more convenience as it makes additional assistive functions possible, such as labelling, 3D annotations [23], surface measurement [24] and extended virtual field-of-view (FOV) [25]. All this extra information presented to the surgeon aids with decision making in terms of visual guides that are directly overlaid onto the surgical site. It opens a new perspective beyond conventional surgical navigation that only involves tri-planar patient scans presented in 2D.

香港大學圖書館 HKU Libraries

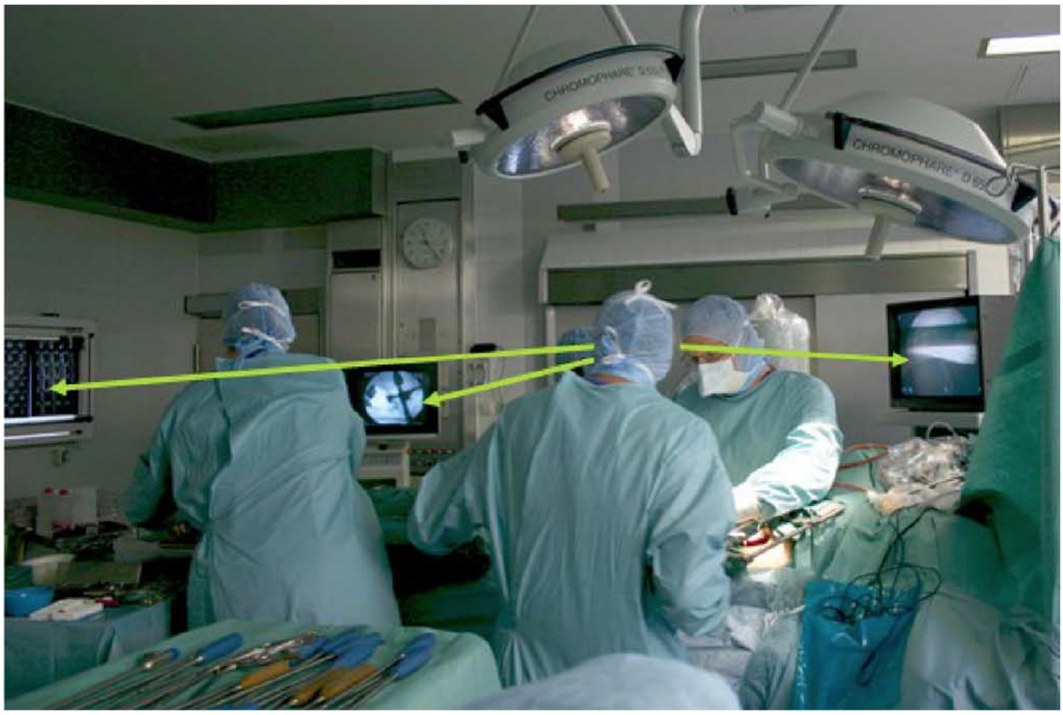

Fig. 2.6 Example of an operating theater setup where surgeons constantly redirect their eyesight between the surgical site and several intra-operative guidance displays during a conventional surgical navigation. **Image source: [26].**

Applying AR to surgeries may give an impression of being surreal and futuristic. In fact, the idea of medical image augmentation for surgical guidance was proposed as early as 1982 by Kelly *et al.* [27] for neurosurgery. They augmented CT-imaged tumor outlines onto a microscope that was on a stereotactic frame. In 1986, Roberts *et al.* [28] further incorporated an ultrasonic tracking system to track the movement of the operating microscope. Interestingly, the abovementioned pioneers developed the idea of applying AR to surgery even earlier than the "birth" of the term "augmented reality", which was coined by Boeing researchers in 1990 [29]. In 1992, Bajura *et al.* [30] presented another early work that applied AR to medicine. It involved the use of an HMD that visualizes 3D ultrasound images on a pregnant human subject, as illustrated in **Fig. 2.8a** and **Fig. 2.8b**. At the same time, as video endoscopy became more common for minimally invasive surgery (MIS) in the 1990's, endoscopic augmentation also emerged in 1998 for brain surgery [31]. In 2000, Lapeer *et al.* [32] proposed an ENT AR framework that displays augmented objects on a stereo microscopic view. In 2012, Navab *et al.* [33] marked a milestone by presenting the first deployment of AR in operating theaters.

香港大學圖書館 HKU Libraries

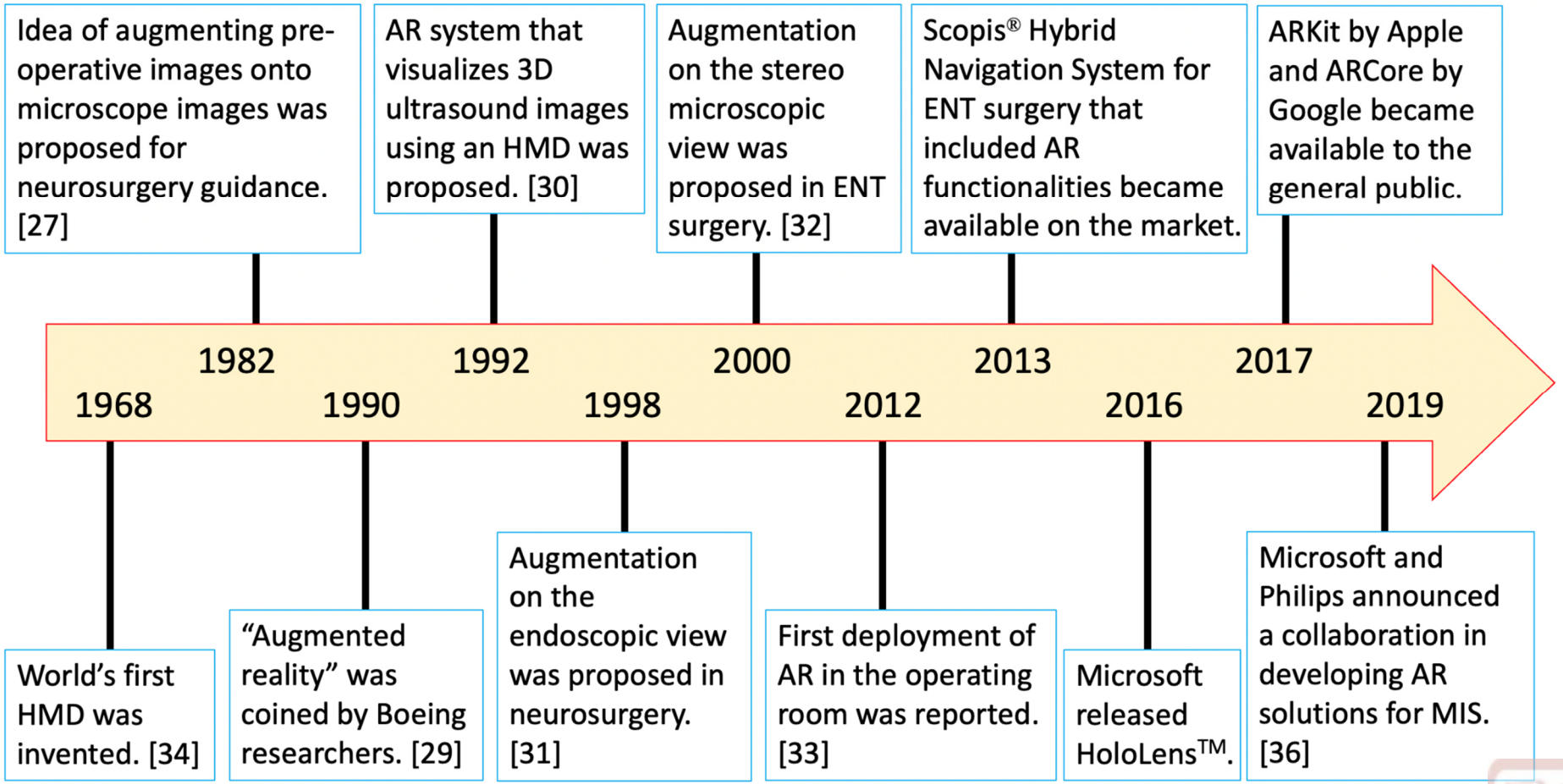

Fig. 2.7 Development timeline of AR research and application in the medical field.

香港大學 圖書館 HKU Libraries

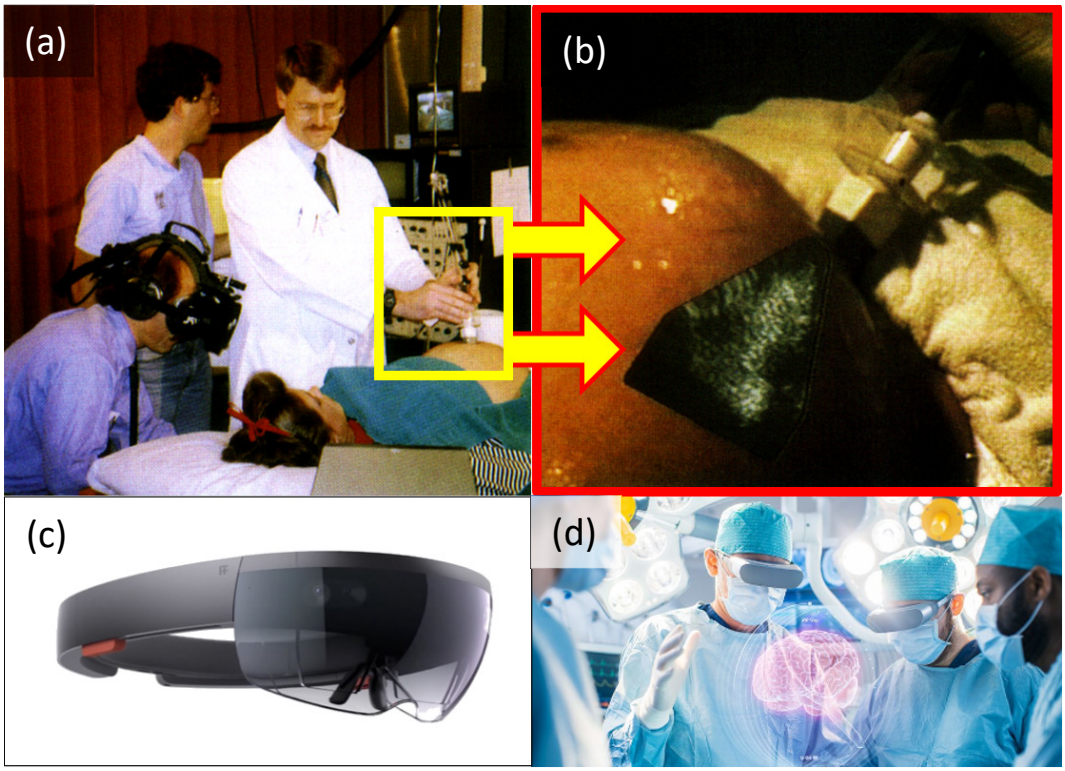

Fig. 2.8 Medical AR guidance using head-mounted displays (HMDs). **(a)** Early example of AR guidance with an observer viewing the patient through a video see-through HMD while ultrasound imaging is performed. **(b)** Superimposed 2D ultrasound image on the patient's abdomen as observed through the HMD. **Image source: [30]**; **(c)** HoloLens™ by Microsoft, an example of a modern optical see-through (OST) HMD. **Image source: Microsoft**; **(d)** Graphical mock-up of pre-operative planning between surgeons with HMDs. **Image source: [35].**

Medical AR guidance has also experienced rapid development in industry over the past decades due to advancement in technologies such as tracking, display, rendering, computation and video endoscopic systems. Scopis® GmbH released the Scopis® Hybrid Navigation System for endoscopic sinus surgery (ESS) in 2013, which included AR functionalities to aid with surgical guidance. In the meantime, popularity of AR continues to increase as software development kits for mobile phones like the ARKit by Apple and ARCore by Google became available to the general public in 2017. Microsoft and Philips also announced a plan in developing AR solutions for MIS in 2019 [36], which is based on the Microsoft HoloLens™ platform (**Fig. 2.8c**) for development.

However, as the medical industry currently lacks a standardized method to quantify overlay error in both the spatial and temporal contexts, surgical AR is still not sufficiently convincing for surgeons to adopt confidently. As a result, the limited uptake of AR is a general phenomenon, even in specialties such as orthopedics and skull base surgeries mostly involving rigid tissue, which are considered as a suitable application of AR [2]. To

香港大學圖書館 HKU Libraries

elaborate, the major challenges and limitations are related to the concern on ergonomics and surgical safety [2, 20, 37]. For example, state-of-the-art AR HMD devices, as illustrated in **Fig. 2.8c** and **Fig. 2.8d**, still suffer from poor ergonomics, with existing devices not being sufficiently lightweight such that surgeons still feel comfortable to put it on for hours throughout the procedure. Current HMDs also have limited FOV for visualizing augmented objects, requiring surgeons to frequently turn their heads to keep the augmented objects within the FOV. Most importantly, error in augmentation might arise during surgery. Possible reasons are i) improper initial registration, ii) deterioration of registration accuracy due to loosening of sensors from the patient and iii) discrepancy between patient's 3D anatomical models and physical anatomy under dynamic surgical scenes. Severe complications may occur in situations where critical structures are accidentally damaged. In sum, despite the benefits of AR and its advancing facilities (e.g. improved ergonomics of using HMD), AR is still not commonly adopted in clinical practice. AR does not add sufficient confidence to surgeons due to their worries about deficiency of pre-clinical validation and well-defined performance indices.

2.1.3 AR-assisted Ear, Nose and Throat (ENT) Surgery

ENT surgery is one of the appropriate candidates for applying AR-assisted surgical guidance. First, it is a category related to skull base surgery, which is often technically demanding due to close proximity to critical structures [38], such as optic nerves and carotid arteries. Visual aids by AR may provide alerts for avoiding these critical structures, which is especially useful for junior surgeons who might not have enough experience, potentially shortening the learning curve. Second, surgical navigation has been a critical tool for paranasal and adjacent skull base surgeries since the 1990s [5]. There exists a solid technological foundation in terms of tracking, display, endoscopic systems, and registration methods, making the development of AR for ENT surgery legitimate and practicable. Third, benefits provided by current conventional navigation are limited [39]. Ergonomics is sub-optimal as surgeons need to constantly redirect his/her eyesight between the surgical field and the intra-operative guidance display [5, 10]. In addition, surgical workflow may be hindered as a probe is repeatedly used to point at known surgical sites for verifying navigation accuracy [40]. Applying AR to ENT surgical navigation can potentially address these limitations.

香港大學圖書館 HKU Libraries

In view of the changes that AR can bring to ENT surgical navigation, interest from both academia and industry has been increasing over the last decade [20, 40]. Studies that apply AR to ENT included, but are not limited to transnasal endoscopic and skull base surgery [25, 41, 42], cochlear implantation [43], transoral robotic surgery [44] and parathyroidectomy [45]. In 2011, Winne *et al.* [46] published a brief clinical report that confirmed the feasibility of augmenting target anatomical models on video images for ESS in a cadaveric setting. However, clinical utility of the augmentation is not detailed in this study. Citardi *et al.* [42] carried out another study of AR in ESS. In this study, the Scopis® Hybrid Navigation System (Stryker, USA) was employed for both pre-operative planning and AR-assisted intra-operative guidance on cadaveric subjects. Before the start of a surgery, surgeons can highlight desired surgical pathways and target anatomy to dissect, which are then aligned and fused with live endoscopic video images during surgery.

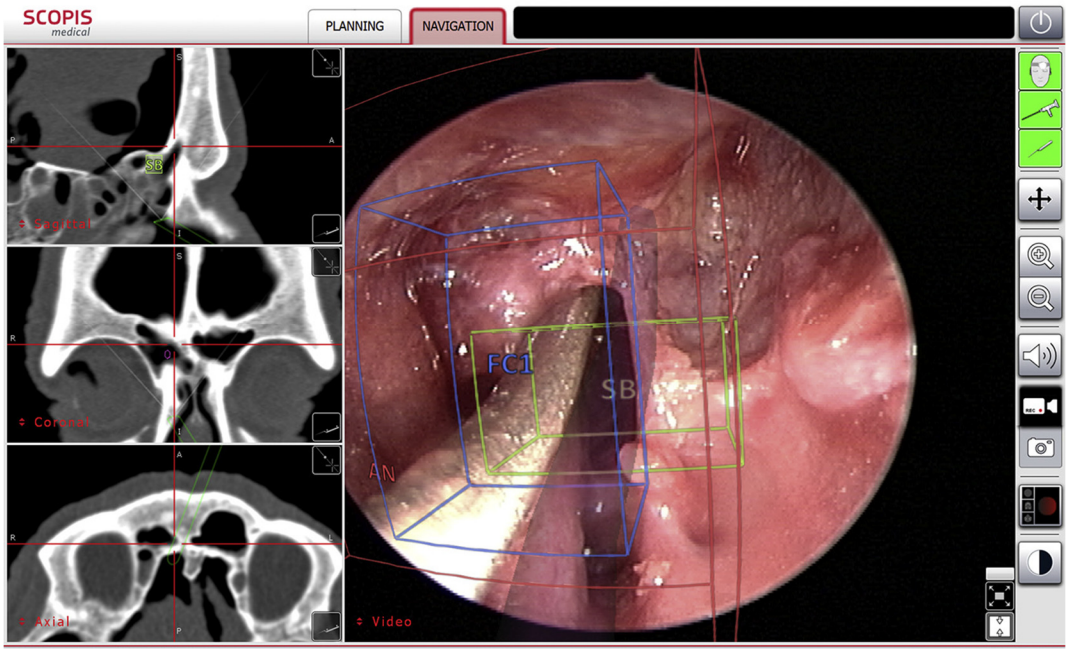

Fig. 2.9 User interface of the Scopis® Hybrid Navigation System (Stryker, USA). The endoscopic view (right) shows intra-operative overlay of bounding boxes indicating dissected frontal recess cells. Tri-planar views (left) show the endoscope tip location with respect to patient CT scans in real-time. **Image source: Stryker.**

As shown in **Fig. 2.9** and **Fig. 2.10**, bounding boxes on dissected frontal recess cells and a navigation pathway leading to the frontal sinus are overlaid on the endoscopic view intra-operatively. These visual aids guide surgeons to efficiently and safely arrive at the target dissection site based on pre-operatively planned information. Tri-planar views are standard representations for conventional surgical navigation and are displayed beside the augmented endoscopic view to help visualize the location of the instrument tip with respect

香港大學圖書館 HKU Libraries

to the patient's CT or MRI scans. Surgeons may also adapt to using AR-assisted guidance more effectively with the presence of traditional tri-planar views. Target registration error (TRE) in this study was estimated at 1.5 mm, implying that this system was clinically feasible because TRE was below the "golden standard" of 2 mm [47], which is the largest acceptable value for surgical navigation. Benefits of AR-assisted ENT surgery in terms of clinical utility were also explained. To elaborate, tracking of the instrument tip and augmentation of frontal sinus outflow tract allowed the cannulation of frontal and sphenoid ostia without the need for a formal ethmoid dissection. In addition, the highlighting of anatomical structures with bounding boxes helped facilitate efficient dissection and enhanced surgical safety through the avoidance of critical structures.

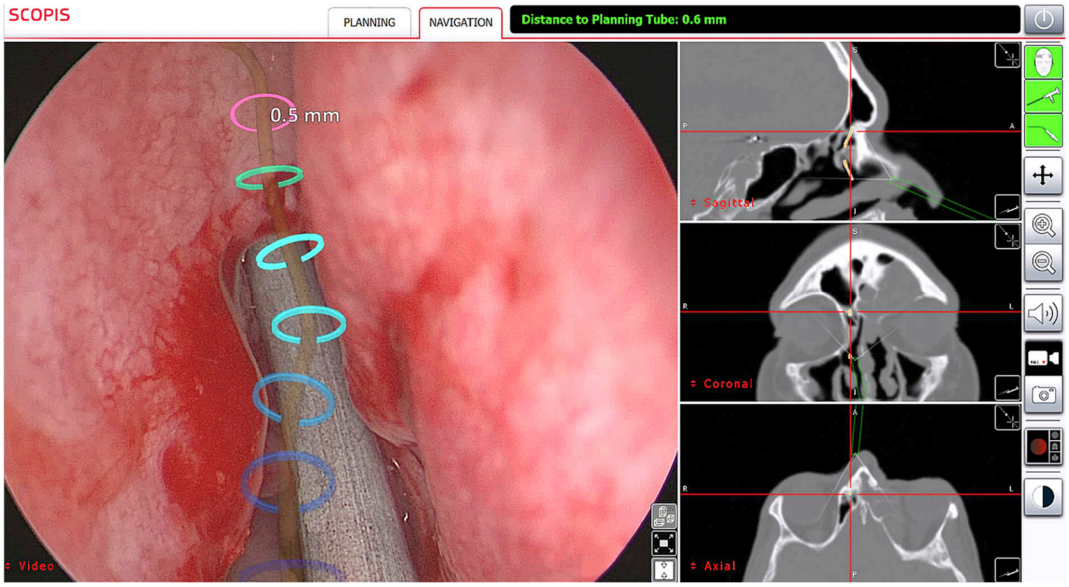

Fig. 2.10 Intra-operative overlay of a navigation pathway leading to the frontal sinus (Scopis® Hybrid Navigation System, Stryker, USA). The path is indicated with rings that show the forward direction along the path (left). Corresponding endoscope tip location with respect to patient CT scans displayed in tri-planar views (right). **Image source: Stryker.**

A study by Li *et al.* [25] presented AR-assistance in an innovative manner by overlaying real-world endoscopic images onto a virtual environment, which is inverse to the method typically used in AR. This study involved 15 otorhinolaryngologists performing maxillary and frontal sinus expansion, sphenoidotomy, ethmoidectomy and intracavernous internal carotid artery dissection on cadaveric heads. These procedures were either performed with a self-developed AR guidance platform or a conventional navigation system without AR. Comparing results with and without AR, average TRE was 1.28 ± 0.45 mm and 1.32 ± 0.41 mm, respectively, while procedure completion time was 88.27 ± 20.45 min and 104.93 ± 24.61 min, respectively. Despite having a small difference in TRE values, procedure

香港大學圖書館 HKU Libraries

completion time that involved AR was statistically lowered ($P < 0.05$). It was also observed that junior surgeons had a larger extent of time reduction than more experienced ones. The author concluded that AR-assisted guidance effectively reduces the mental workload and operation time of surgeons, especially for junior surgeons who have less experience. In this study, FOV of the nasal endoscope is expanded to give an extended view beyond what the endoscope observes in reality, as shown in **Fig. 2.11a**. More importantly, sub-surface anatomical structures such as blood vessels, eyeballs, optic nerves, target lesions and brain can be visualized (**Fig. 2.11b** and **Fig. 2.11c**), enabling critical structure avoidance and relieving surgeon's mental burden.

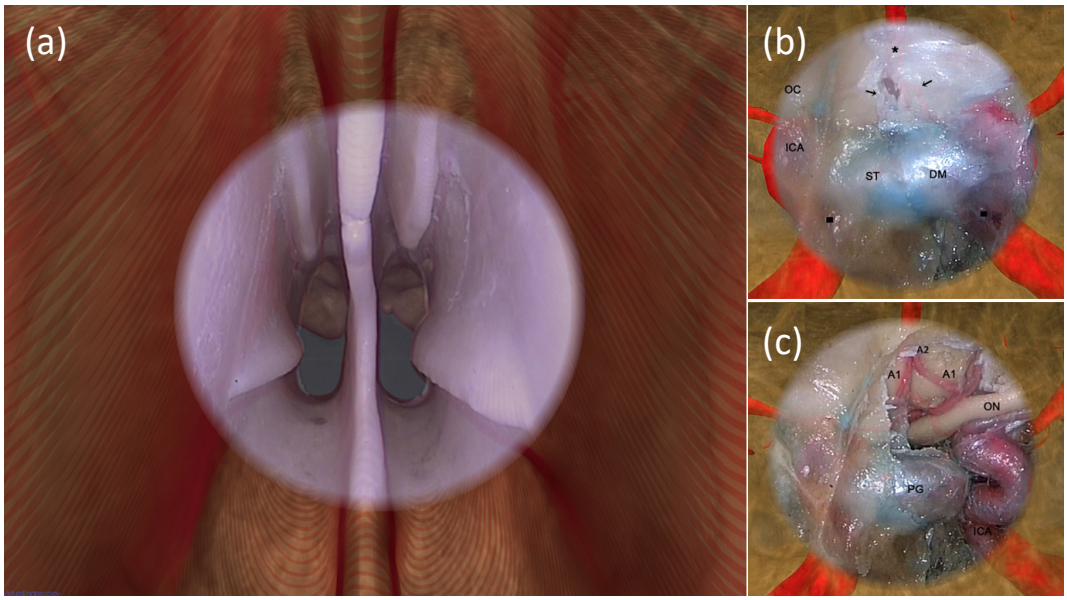

Fig. 2.11 AR achieved by overlaying real-world endoscopic images onto a virtual environment. **(a)** Augmentation in the nasal airway of a phantom. **(b)** Augmentation of a cadaver sphenoid sinus that shows an exposed dura mater after bone removal in the lateral, posterior and superior lateral walls. After the dura mater is opened, **(c)** shows that the actual locations of the blood vessels are consistent with their projections in the extended virtual view. **Image source: [25].**

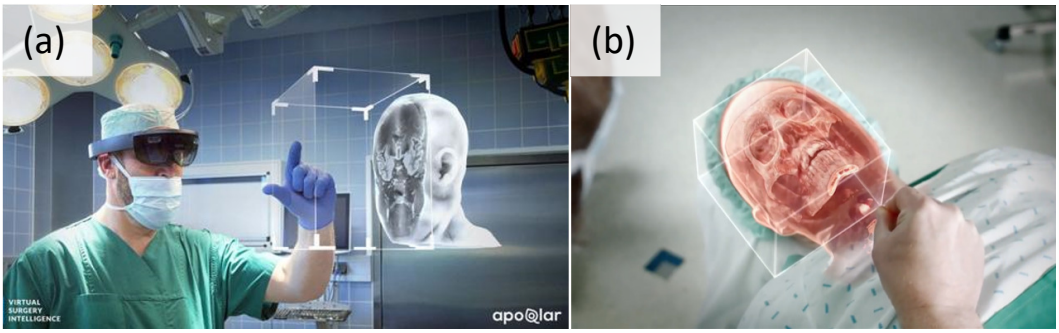

Fig. 2.12 Illustration of the VSI HoloMedicine® system being used to perform **(a)** pre-operative planning and **(b)** overlay of a 3D anatomical model onto the patient. **Image source: Apoqlar.**

香港大學圖書館 HKU Libraries

While the Scopis® Hybrid Navigation (Stryker, USA) utilized by Citardi *et al.* [42] is a well-known product that exemplifies video-based AR, there also exists OST AR in the ENT surgery market. The VSI HoloMedicine® system, a medical mixed reality software platform developed by ApoQlar, takes advantage of the Microsoft HoloLens™ hardware to achieve AR-assisted healthcare and education, as illustrated in **Fig. 2.12a** and **Fig. 2.12b**. This solution is capable of facial surface recognition and alignment of pre-operative patient models for AR guidance of ENT surgery. It is foreseeable that interest from both academia and industry on AR-assisted ENT surgery will continue to grow.

2.2 GENERAL INSTRUMENT TRACKING TECHNOLOGIES IN SURGERY

Currently, there are two main types of tracking modalities used in surgical navigation solutions, namely optical and EM tracking. Both tracking modalities are reported to achieve submillimeter accuracy under ideal conditions [1]. Although they have fundamentally different working principles, they share the same objectives, which is to track the position and orientation of i) patient anatomy, ii) surgical instruments and iii) endoscopes in real-time. This section provides a brief introduction on the working principles, advantages, and disadvantages of these tracking solutions.

2.2.1 Optical-based Navigation

Optical-based navigation usually refers to tracking of infrared (IR) markers by a two-camera sensor as shown in **Fig. 2.13a**. These markers are either active or passive as depicted in **Fig. 2.13c**. "Active" means the markers are mounted with light-emitting parts, while "passive" means the markers rely on reflecting IR light emitted from illuminators. Having a minimum of two calibrated cameras, the 3D position of a marker can be calculated by stereoscopic triangulation. Position tracking of several markers mounted in a pre-defined configuration (**Fig. 2.13b**) may further determine the object's orientation.

香港大學圖書館 HKU Libraries

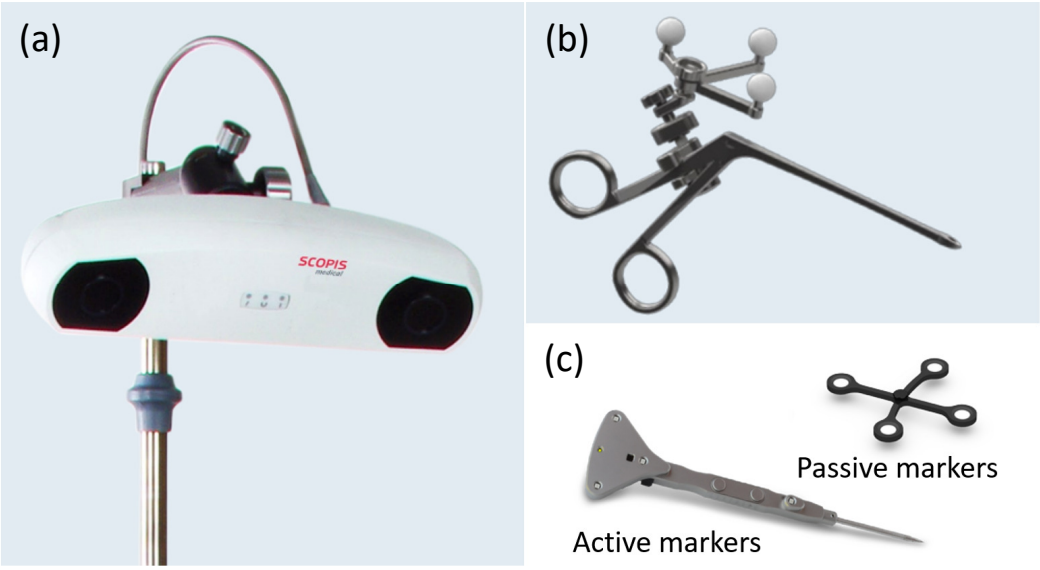

Fig. 2.13 (a) Stereo-camera of an optical tracking system (Stryker, USA). **(b)** Instrument mounted with reflective optical markers. **Image source: Stryker**; **(c)** Active markers with light-emitting components, and passive reflective markers (Atracsys, Swiss). **Image source: Atracsys.**

A benefit of using optical tracking is that markers attached on the patient or an instrument can be a standalone structure without requiring wiring for transmitting sensor data. However, these marker frames are usually bulky in the form of a triangular shape [48], causing inconvenience during an operation. Most importantly, tracking becomes invalid when markers are not within the line-of-sight of the cameras. This situation may frequently occur because the arm or the body of a surgeon is usually close to the marker frames, subsequently obstructing the line-of-sight.

2.2.1.1 Camera Calibration

To achieve stereoscopic triangulation, the two cameras should be calibrated. Camera calibration specifically refers to the process that attains camera parameters such as distortion coefficients, intrinsic and extrinsic parameters. A starting point of introducing this concept is the pinhole camera model as depicted in **Fig. 2.14**. In this model, light rays enter the camera from a focal point to reach the sensor. Light rays are then projected onto the sensor to form an image.

香港大學圖書館 HKU Libraries

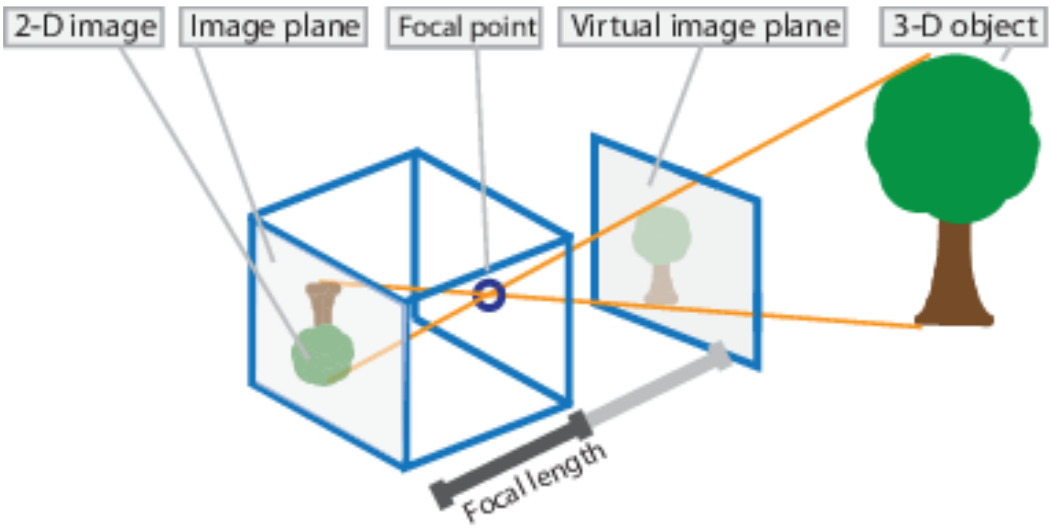

Fig. 2.14 Schematic diagram of a pinhole camera that consists of a single small aperture. As light rays pass through the aperture, an inverted image is formed on the image plane. **Image Source: [49].**

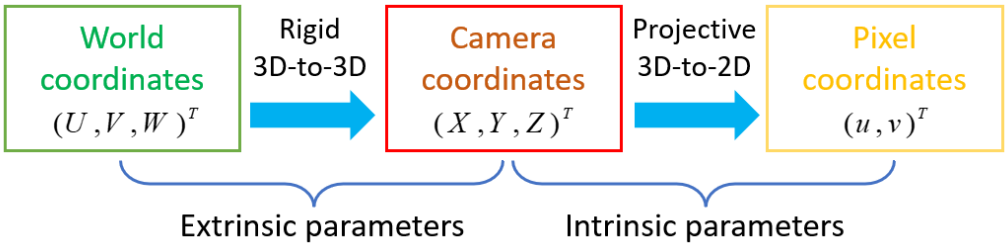

Fig. 2.15 Forward projection from a point $(U,V,W)^T$ in the 3D world to a pixel $(u,v)^T$ on a 2D image plane. Extrinsic parameters describe a rigid 3D-to-3D transformation from the world coordinate frame to the camera local coordinate frame, while intrinsic parameters describe a projective 3D-to-2D transformation from the camera local coordinate frame to pixel coordinates.

Mathematically, a point $(U,V,W)^T$ in the 3D world coordinate system undergoes a rigid transformation ${}^c\mathbf{T}_w$ to give $(X,Y,Z)^T$ with respect to the camera local frame of reference:

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = {}^c\mathbf{T}_w \begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix}, \tag{2.1}$$

where ${}^c\mathbf{T}_w$ is also called extrinsic parameters that describe the world's reference frame with respect to the camera local frame. Next, a 3D-to-2D perspective projection is performed with a 3-by-3 camera intrinsic matrix $\mathbf{K}$ to give 2D pixel coordinates $(u,v)^T$. Specifically, assuming square pixels:

香港大學圖書館 HKU Libraries

$$w\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{2.2}$$

where w is a scale factor, f_x and f_y are focal lengths, $(c_x, c_y)^T$ is the optical center where the optical axis and the image plane intersect. f_x, f_y and $(c_x, c_y)^T$ are expressed in the unit of pixels.

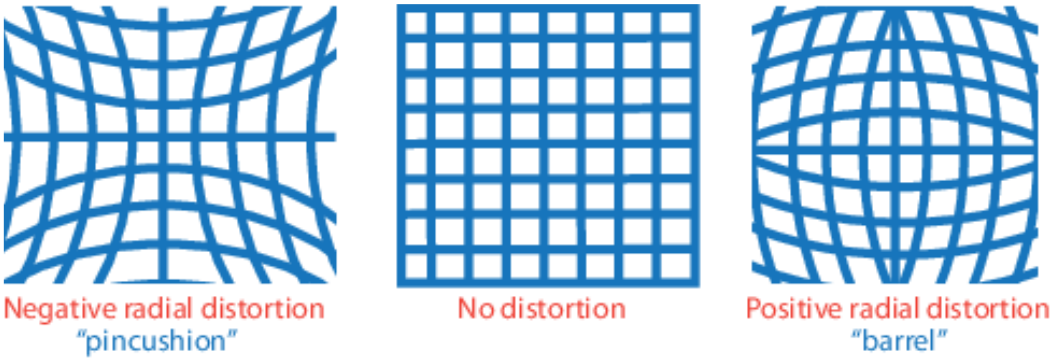

Fig. 2.16 Illustration of radial distortion caused by light rays bending more at the edges of a lens, giving rise to either a "pincushion" or a "barrel" effect. **Image Source: [49].**

Apart from attaining $\mathbf{K}$, camera calibration also gives distortion parameters of a camera. In practice, every lens is imperfect and would normally exhibit distortion effects. As illustrated in **Fig. 2.16,** radial distortion occurs when light rays bend more at the edges of a lens than at the optical center. It can be characterized by constants k_1 and k_2. For highly distorted endoscopes, a third constant term k_3 can be added:

$$x' = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \tag{2.3}$$

$$y' = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6), \tag{2.4}$$

where x and y are pixel coordinates before distortion while x' and y' are pixel coordinates after distortion. Coordinates x, y, x' and y' are in normalized pixel units, and $r^2 = x^2 + y^2$.

香港大學圖書館 HKU Libraries

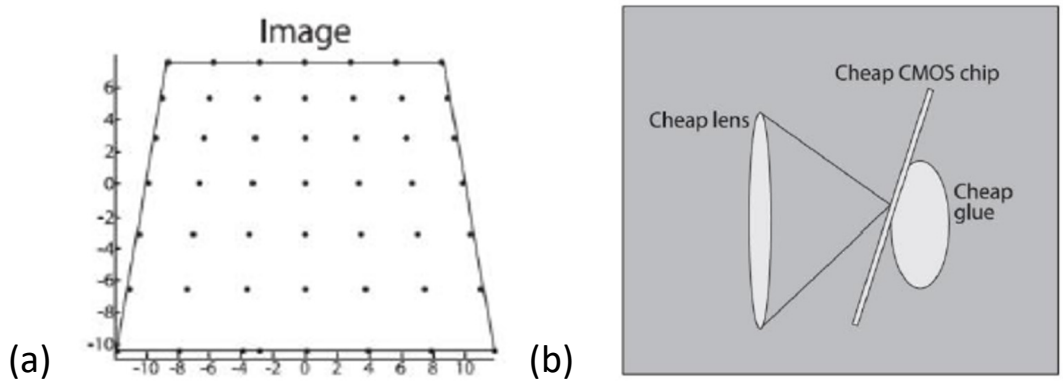

Fig. 2.17 (a) Tangential distortion illustrated on a grid. **(b)** Lens and image sensor not parallel to each other, leading to tangential distortion. **Image source: [50].**

As shown in **Fig. 2.17a**, another type of distortion is tangential distortion. Due to manufacturing flaws, inferior product quality or deteriorated product condition, the lens might not be parallel to the image sensor (**Fig. 2.17b**), leading to tangential distortion. It can be characterized by two constants p_1 and p_2, given by:

$$x' = x + [2p_1xy + p_2(r^2 + 2x^2)] \tag{2.5}$$

$$y' = y + [p_1(r^2 + 2y^2) + 2p_2xy], \tag{2.6}$$

where x and y are pixel coordinates before distortion while x' and y' are pixel coordinates after distortion. Coordinates x, y, x' and y' are in normalized pixel units, and $r^2 = x^2 + y^2$.

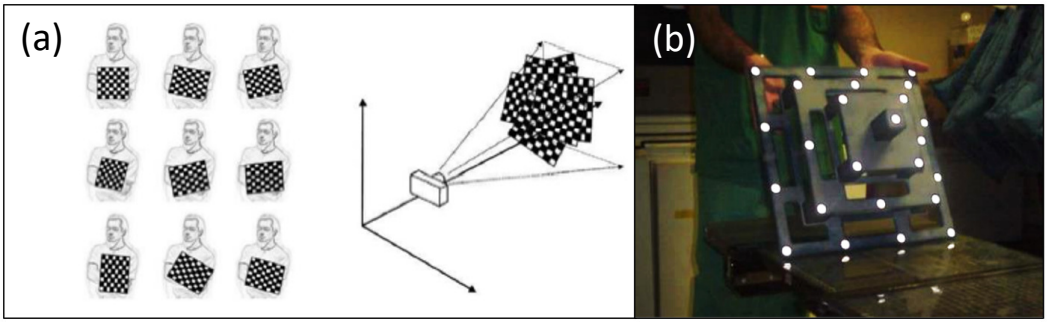

Fig. 2.18 Camera calibration using **(a)** chessboard and **(b)** 3D calibration pyramid mounted with optical markers. When calibrating with a chessboard, images of the chessboard are taken at different viewing angles. Images are taken by either fixing the chessboard while moving the camera or fixing the camera while moving the chessboard. **Image source: [50].**

香港大學圖書館 HKU Libraries

In summary, the intrinsic parameters of a camera are $(f_x, f_y, c_x, c_y, k_1, k_2, k_3, p_1, p_2)$. During the process of camera calibration, images are taken as the camera is pointed towards objects (calibrators). These calibrators can be a chessboard grid with known square numbers and sizes as shown in **Fig. 2.18a**, or a frame with known geometry mounted with optical markers as shown in **Fig. 2.18b**. Commonly used open-source toolboxes include the MATLAB built-in toolbox "Single Camera Calibrator App" [49] and the OpenCV camera calibration module [51]. These toolboxes perform calibration with a chessboard grid. Pattern keypoints, which are the corners of squares in a grid, are first detected on a calibration image. Next, corresponding world points of these keypoints are projected onto the same image. Intrinsic parameters are then estimated by minimizing the distance between detected keypoints and re-projected points.

2.2.1.2 Computing the Position of an Optical Marker in 3D

With all the intrinsic parameters of a camera, a reverse process of forward projection becomes possible. A detected optical marker $(u, v)^T$ on an image can be "back-projected" to give the 3D location $(X, Y, Z)^T$ of the marker with respect to the camera local frame of reference, given by the following:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{K}^{-1} \cdot d \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{2.7}$$

where $\mathbf{K}$ is the camera intrinsic matrix, and d is the depth of the marker from the camera's optical center.

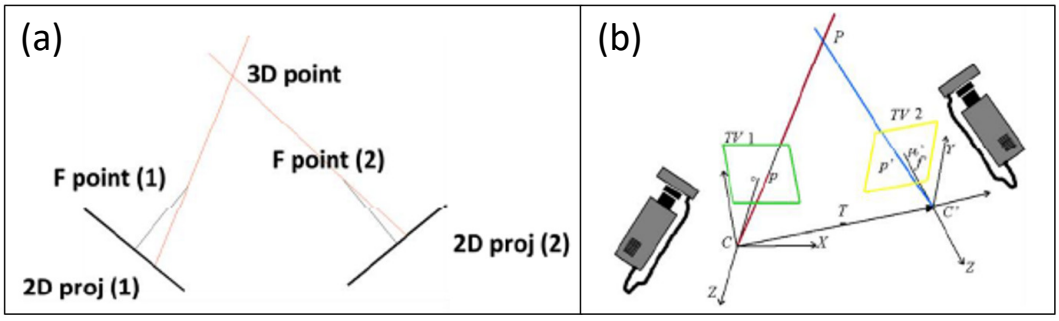

Fig. 2.19 3D reconstruction of a point with a stereo-camera optical tracking system by stereoscopic triangulation, as illustrated **(a)** from a top view and **(b)** in 3D. **Image source: [50].**

However, depth d is not known because the exact location of the marker can be anywhere along a line that intersects the optical center and $(u, v)^T$ that is on the image plane. As such, the point of interest needs to be reconstructed by stereoscopic triangulation. When at least two calibrated cameras with known relative position and orientation are employed, the location of the marker can be calculated by finding the intersection of the lines mentioned above, as illustrated in **Fig. 2.19a** and **Fig. 2.19b**. In practice, these lines may not intersect because of numerical errors. In this circumstance, the point can be reconstructed by finding a position that is closest to these lines in a least-squares sense.

2.2.2 Electromagnetic (EM) Tracking

In an EM tracking system, a magnetic field generator is positioned near the target anatomy as depicted in **Fig. 2.20d**. The generator produces a magnetic field, which in turn induces small amplitudes of voltage in inductor coils embedded inside EM sensors. Orientation and position of EM sensors with respect to the field generator are then decoded from the induced voltage. As shown in **Fig. 2.20a**, an EM sensor is more compact in size and shape when compared with optical markers, and can be conveniently mounted onto the patient (**Fig. 2.20c**) or an instrument (**Fig. 2.20b**). Most importantly, unlike optical tracking, EM tracking does not suffer from line-of-sight issues. Object obstruction between the field generator and the sensor does not affect sensor localization.

However, there are also disadvantages of using EM tracking. Generally, EM tracking systems have a higher latency than optical tracking systems due to the process of internal filtering [52]. Wu *et al.* [53] compared the latency of the Aurora system (NDI, Canada) with an optical tracking system and found its latency to be 80 ms more. When applied to surgical AR, a temporal misalignment between physical anatomy and the overlaid virtual object may become obvious, creating a feeling of overlaid objects “floating” unstably on the endoscopic view. Additionally, EM tracking systems suffer from field distortion errors in the presence of ferromagnetic or conductive materials [52]. In particular, Schicho *et al.* [54] tested the disturbance on an Aurora system with metallic instruments and recorded errors of up to 5 mm. Therefore, it is of paramount importance to avoid the presence of ferromagnetic or conductive materials near or inside the tracking volume to avoid interference. Furthermore, EM sensors usually require wiring for transmitting signals to a tracking console, as illustrated in **Fig. 2.21**.

香港大學圖書館 HKU Libraries

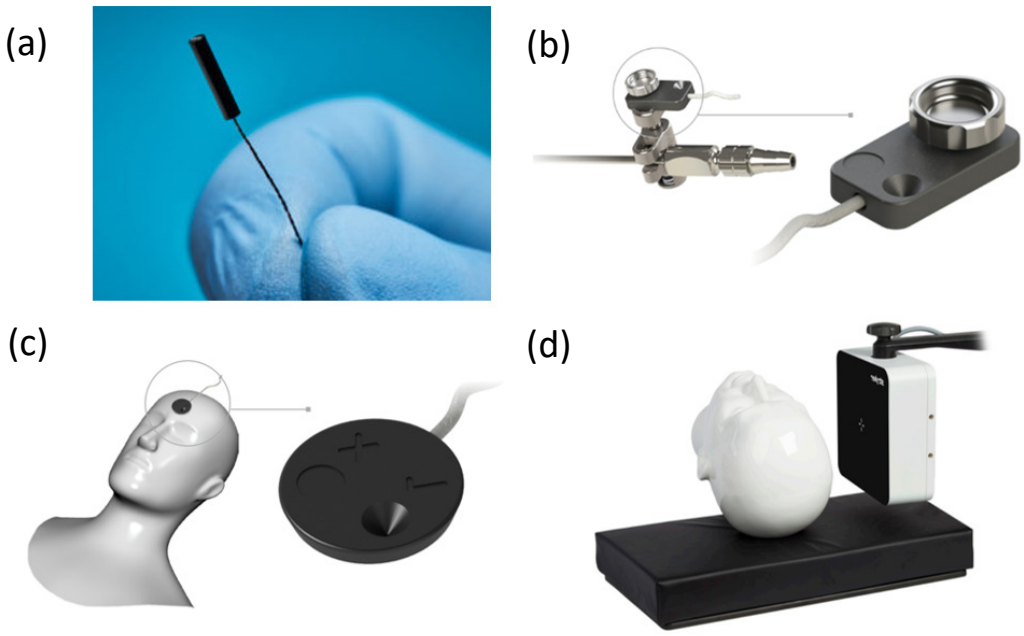

Fig. 2.20 (a) 6-DoF EM sensor (Aurora, NDI, Canada) with a small size of ø 1.8 x 9.0 mm. **Image source: NDI**; **(b)** EM sensor (Stryker, USA) coupled to an endoscope using an adaptor. **(c)** EM sensor (Stryker, USA) attached onto the patient's forehead for tracking movement of the head. **(d)** Example placement of an EM field generator (Stryker, USA) near a patient to localize sensors. **Image source: Stryker.**

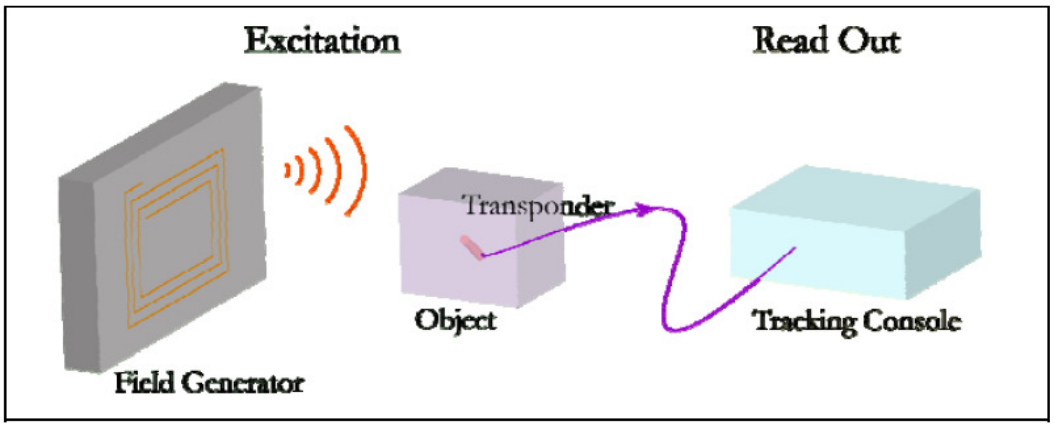

Fig. 2.21 Illustrative example setup of an EM tracking system. The field generator is composed of transmitting coils that generate changing magnetic fields. A voltage is induced in the transponder, or receiving coil, which transmits a signal to the tracking console where the transponder's location and orientation is computed. **Image source: [55].**

The field generator encompasses differently aligned inductors (transmitting coils) for producing changing magnetic fields along multiple axes [52]. Next, a wired sensor, which is also an inductor (receiving coil), experiences changes in magnetic flux, resulting in an induced voltage. The signal is then transmitted to a tracking console. Computation at the

香港大學圖書館 HKU Libraries

tracking console then decodes the position and orientation of sensor. EM tracking can be categorized into alternating current (AC)-based and direct current (DC)-based [56]. For AC-based tracking, receiving coils are search coils that measure induced voltage produced by an alternating magnetic field. For DC-based tracking, fluxgate sensors are used to vectorially measure static or low frequency magnetic fields [52]. The basic working principle of EM tracking systems is illustrated in **Fig. 2.22**.

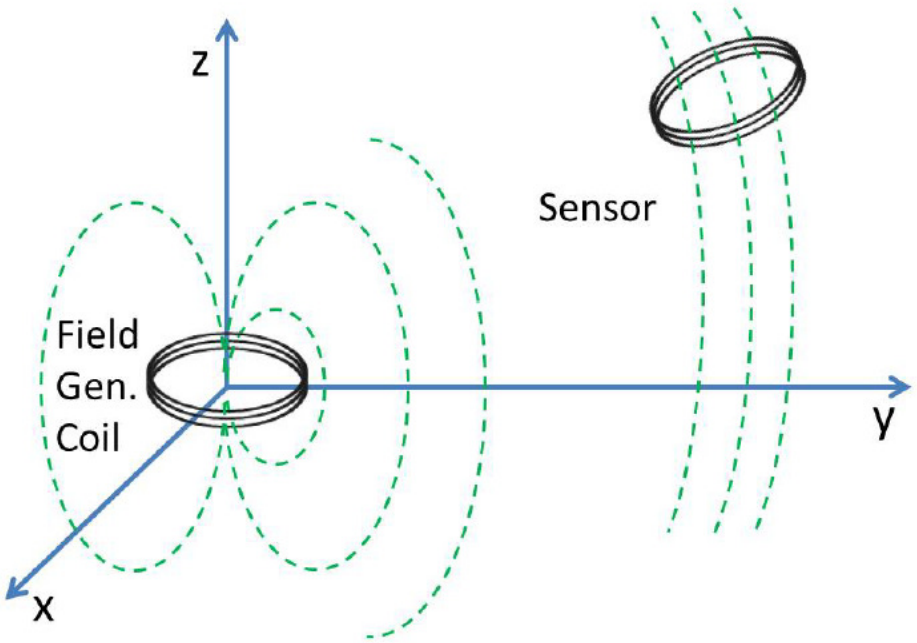

Fig. 2.22 Basic working principle of EM tracking systems. Field generating coils produce a magnetic field with a varying field strength at different locations. Voltage is induced at the receiving coil according to Faraday's Law of Induction. Location of the receiving coil is estimated by minimizing a cost function that describes difference between measured and theoretical voltages. **Image source: [56].**

In general, a transmitting coil generates a magnetic field, which can be approximated by the equivalent dipole field [57-59]:

$$\mathbf{B}(x,y,z,t)=\frac{\mu_0}{4\pi}\left[\frac{3(\mathbf{M}\cdot\mathbf{r})\mathbf{r}}{|\mathbf{r}|^5}-\frac{\mathbf{M}}{|\mathbf{r}|^3}\right]e^{-j\omega t}, \tag{2.8}$$

where μ_0 is the permeability of free space, $\mathbf{M}$ is the coil's magnetic moment, $\mathbf{r}$ is the vector from the coil to the observation point, t is time and ω is the operating frequency. Specifically, $\mathbf{M}=NAI\hat{\mathbf{n}}$, where N is the coil's number of turns, A is the area encircled by the coil, I is the current, and $\hat{\mathbf{n}}$ is a unit normal vector of area A. When a receiving coil enters a time-varying magnetic field generated by the transmitting coil, a voltage is

香港大學圖書館 HKU Libraries

induced at the receiving coil according to Faraday's Law of Induction:

$$V(x, y, z, t) = -j\omega NA\mathbf{B}(x, y, z, t)\hat{\mathbf{n}}, \tag{2.9}$$

where N is the receiving coil's number of turns, A is the area encircled by the coil, ω is the operating frequency at the transmitting coil, and $\hat{\mathbf{n}}$ is the coil area's unit normal vector. By arranging k transmitting coils in a known configuration, the position of a receiving coil can be estimated. Estimation is achieved by minimizing a cost function that describes the difference between i) the measured voltage and ii) the theoretical voltage from equation (2.9). An example cost function is as follows:

$$\Phi = \sum_{i=1}^{k} (V_i^t - V_i^m)^2, \tag{2.10}$$

where V^t is the theoretical voltage and V^m is the measured voltage. For every transmitter-receiver coil couple, a residual error $V_i^t - V_i^m$ exists. The sum of squared residual errors is then minimized, giving the corresponding estimated location. By composing the cost function with magnetic field amplitudes instead of voltage values, this method can be adjusted and applied on DC-based EM tracking too [56].

The simplified model described above only computes the position of a sensor. To derive orientation, magnetic dipole can be considered. Calculating mutual inductance between coils [60, 61] is an example method. However, as a magnetic dipole is axially symmetric, available rotation information is limited to two degrees of freedom (DoF) instead of three [52]. Therefore, it is common to see five-DoF sensors available on the market which include three translation DoFs and two rotation DoFs. To achieve six-DoF tracking, two five-DoF sensors can be combined, or multiple inductor coils can be assembled to give a sensor with multiple dipoles.

香港大學圖書館 HKU Libraries

2.3 TOOL CALIBRATION

As introduced in **Section 2.2**, optical and EM tracking are the two main technologies employed in both conventional surgical navigation and AR-assisted guidance. Before localizing an instrument or an endoscope intra-operatively, calibration processes are

required. This is because the tracking point defined by an EM sensor or optical markers does not coincide with the instrument tip or the optical center of an endoscope. Therefore, a calibration process that specifies the instrument tip as a target tracking point is necessary. This process is commonly referred to as "pivot calibration" for instruments. To characterize the transformation between an endoscope's optical center and the sensor, a calibration process known as "hand-eye calibration" is required.

2.3.1 Pivot Calibration for Instrument Tip Localization

As an optical marker frame or an EM sensor is attached onto an instrument, they form a dynamic reference frame (DRF) for the optical or EM system to track. When a DRF is attached onto an instrument, the tracking system does not recognize the geometrical relationship between the DRF and the target tracking point, or in other words the instrument tip. Therefore, pivot calibration is needed to inform the system about this spatial transformation.

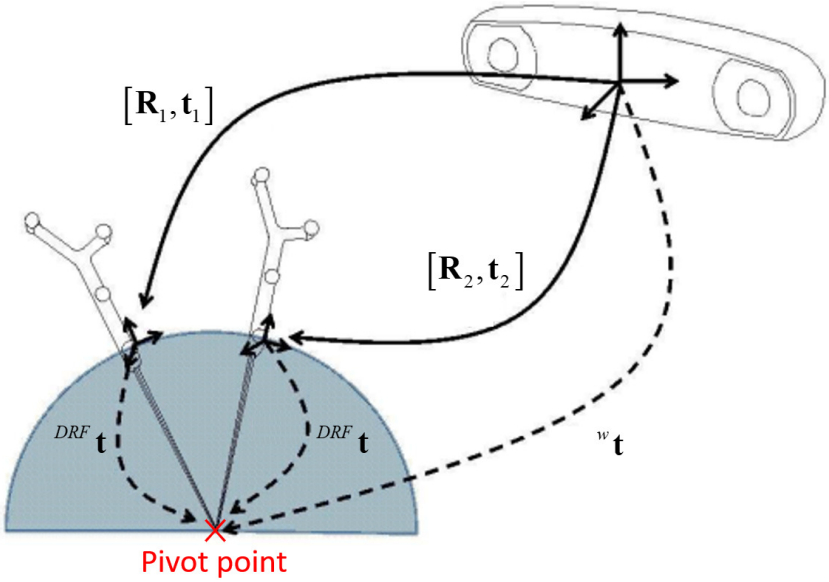

Fig. 2.23 Pivot calibration of a pointer tool. The dynamic reference frame (DRF) is defined by the marker pattern on the tool. During pivot calibration, the tool tip is fixed at the pivot point and moved along a hemispherical surface in 3D. Two poses of the DRF $[\mathbf{R}_i, \mathbf{t}_i], i = 1, 2$ detected by the stereo-camera are denoted by solid lines, while unknown translations are denoted by dashed lines. After pivot calibration, translations $^{DRF}\mathbf{t}$ and $^{w}\mathbf{t}$ are calculated. Translation $^{DRF}\mathbf{t}$ describes instrument tip's location relative to the DRF origin, while translation $^{w}\mathbf{t}$ describes pivot's position with respect to the tracking system origin. **Image source: [62].**

香港大學圖書館 HKU Libraries

During pivot calibration, the tip of an instrument is anchored at a pivot point. The instrument is then manually rotated around this pivot as illustrated in **Fig. 2.23**. Mathematically, pivot calibration is defined as follows: Estimate the position ${}^{DRF}\mathbf{t}$ of the instrument tip with respect to the DRF's origin, given that m transformations $\left[\mathbf{R}_i, \mathbf{t}_i\right], i = 1, 2, ..., m$, which describe the poses of the DRF with respect to the optical or EM tracking system's origin, are obtained during the calibration. Note that $\mathbf{R}$ is a 3-by-3 rotation matrix and $\mathbf{t}$ is a translation vector. There are different methods for solving this problem, one of them is the Algebraic One Step (AOS) as documented by Ziv [62]. As the instrument is rotated around the pivot, ${}^{DRF}\mathbf{t}$ and the pivot position ${}^{w}\mathbf{t}$ with respect to the tracking system origin (world origin) are constants because the pivot point is fixed. For each $\left[\mathbf{R}_i, \mathbf{t}_i\right]$ measured, we have the following:

$$\left(\mathbf{R}_i\right)\left({}^{DRF}\mathbf{t}\right) + \mathbf{t}_i = {}^{w}\mathbf{t} \tag{2.11}$$

After m poses $\left[\mathbf{R}_i, \mathbf{t}_i\right], i = 1, 2, ..., m$ are sampled, the following overdetermined system is established:

$$\begin{bmatrix} \mathbf{R}_1 & -\mathbf{I} \\ \vdots & \vdots \\ \mathbf{R}_m & -\mathbf{I} \end{bmatrix} \begin{bmatrix} {}^{DRF}\mathbf{t} \\ {}^{w}\mathbf{t} \end{bmatrix} = \begin{bmatrix} -\mathbf{t}_1 \\ \vdots \\ -\mathbf{t}_m \end{bmatrix} \tag{2.12}$$

By solving this equation, ${}^{DRF}\mathbf{t}$ and ${}^{w}\mathbf{t}$ can be obtained. In theory, as there are only two unknown translation vectors, two equations established from two instrument poses are sufficient to compute ${}^{DRF}\mathbf{t}$, that is:

$$\begin{cases} \left(\mathbf{R}_1\right)\left({}^{DRF}\mathbf{t}\right) + \mathbf{t}_1 = {}^{w}\mathbf{t} \\ \left(\mathbf{R}_2\right)\left({}^{DRF}\mathbf{t}\right) + \mathbf{t}_2 = {}^{w}\mathbf{t} \end{cases} \tag{2.13}$$

Solving this system of two equations and two unknowns may compute ${}^{DRF}\mathbf{t}$, where:

$${}^{DRF}\mathbf{t} = -\left(\mathbf{R}_1 - \mathbf{R}_2\right)^{-1}\left(\mathbf{t}_1 - \mathbf{t}_2\right) \tag{2.14}$$

Although two measured poses are sufficient for this computation, solving from more DRF poses is statistically beneficial as the effect from random errors can be minimized.

香港大學圖書館 HKU Libraries

2.3.2 Hand-eye Calibration for Endoscope Pose Estimation

Similar to instrument localization, when an optical marker frame or an EM sensor is attached onto the endoscope, a DRF is established for the system to track. Therefore, a calibration process called "hand-eye calibration" is necessary to characterize the transformation between the DRF and the endoscope's optical center. Accurate hand-eye calibration is critical for AR-assisted surgical guidance because overlay accuracy of all virtual elements on the endoscopic view heavily depends on this calibration process. To elaborate, an improper transformation characterisation by hand-eye calibration, especially rotational relationship, leads to a prominent overlay error that would even deteriorate when target objects are further away from the endoscope, resembling a "lever arm effect". For a flexible endoscope, the DRF must be mounted near the endoscope tip to ensure the transformation between the DRF and the optical center is fixed, as shown in **Fig. 2.24a** and **Fig. 2.24b**. For rigid endoscopes, the DRF can be positioned near the handle similar to how sensors and marker frames are mounted on instruments.

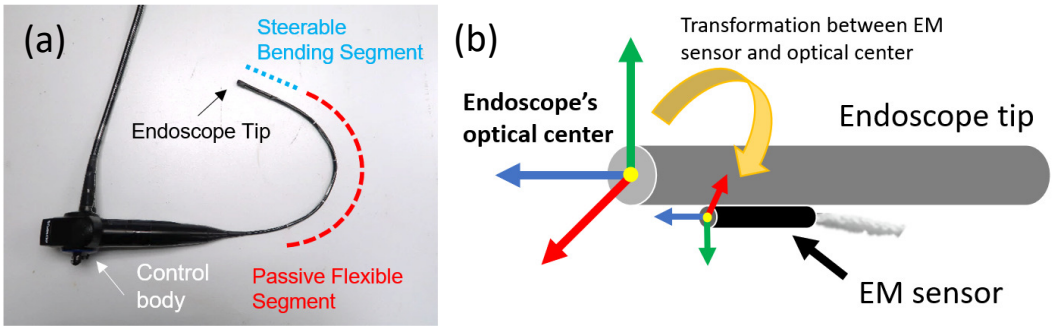

Fig. 2.24 (a) Flexible rhinolaryngoscope (ENF-VH, Olympus) with a steerable bending segment at the tip. **(b)** Illustration of the 6-DoF EM sensor (Aurora, NDI, Canada) anchored at the tip of the flexible endoscope. Transformation between the EM sensor and the endoscope's optical center can be obtained from hand-eye calibration.

In the formulation of hand-eye calibration, "hand" refers to the DRF while "eye" refers to the endoscope's optical center. As the DRF is attached on the endoscope, this hand-eye problem is specifically an eye-in-hand problem. During the calibration process, an endoscope, which is attached with optical markers or an EM sensor, is moved around a calibration pattern, often a chessboard pattern, and takes images of it, as illustrated in **Fig. 2.25**. This pattern defines a fixed frame of reference, so for every image taken by the endoscope, the pose of this "pattern frame" with respect to the "eye" ${}^{e}\mathbf{T}_{p}$ is stored. Pose

香港大學圖書館 HKU Libraries

${}^{e}\mathbf{T}_{p}$ is equivalent to the extrinsic parameters introduced in **Section 2.2.1.1**. Simultaneously, pose of the DRF ${}^{w}\mathbf{T}_{DRF}$ with respect to the tracking system (world) is also stored when every image is captured. Considering two images captured at two endoscope poses, by naming ${}^{w}\mathbf{T}_{DRF}$ and ${}^{e}\mathbf{T}_{p}$ at position 1 as $\mathbf{A}_i$ and $\mathbf{B}_i$, while that for position 2 as $\mathbf{A}_j$ and $\mathbf{B}_j$ respectively, expression simplification is made as follows:

$$\begin{cases} \mathbf{A} = \mathbf{A}_j \mathbf{A}_i^{-1} \\ \mathbf{B} = \mathbf{B}_j^{-1} \mathbf{B}_i \end{cases} \tag{2.15}$$

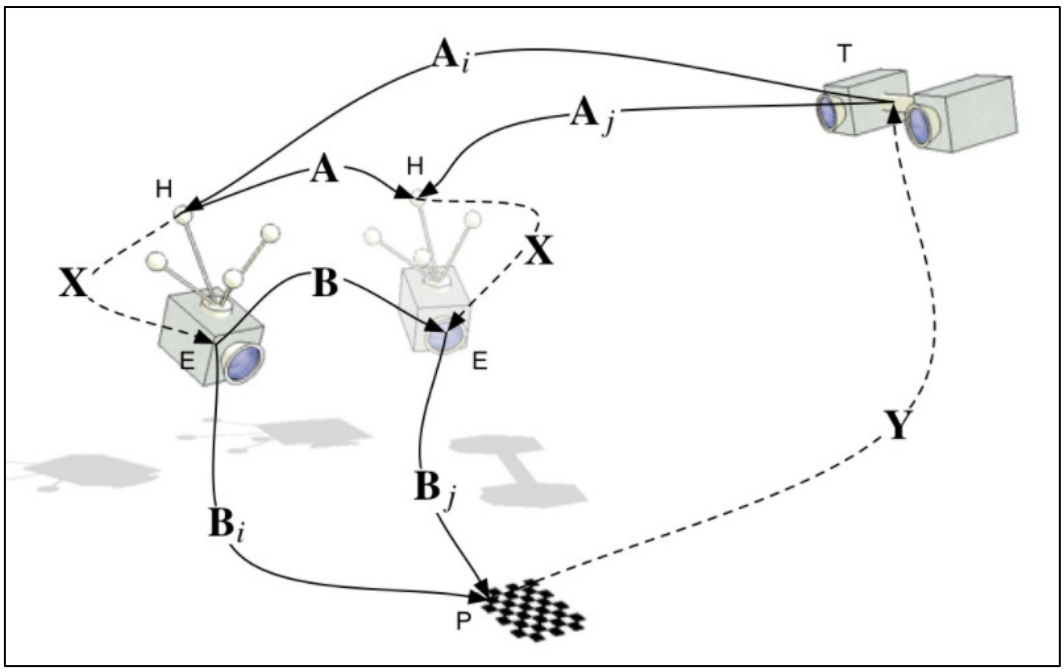

Fig. 2.25 Hand-eye calibration in an eye-in-hand configuration. At two different camera poses, the camera has extrinsic parameters $\mathbf{B}_i, \mathbf{B}_j$ and corresponding tracked poses $\mathbf{A}_i, \mathbf{A}_j$ provided by an external tracking system. By solving $\mathbf{AX} = \mathbf{XB}$ (equation 2.16), transformation $\mathbf{X}$ from the camera's optical center to the DRF is calculated. **Image source: [63].**

$\mathbf{X}$ is then defined as the transformation from the endoscope's optical center to the DRF, which is the target transformation to be computed. Therefore, for every pair of endoscope poses, there will be a geometrical relationship formed:

$$\mathbf{AX} = \mathbf{XB} \tag{2.16}$$

香港大學圖書館 HKU Libraries

Solving the $\mathbf{AX = XB}$ relationship is a well-studied problem with different approaches [64-67]. In general, $\mathbf{AX = XB}$ can be decomposed into a matrix equation and a vector equation:

$$\mathbf{R}_A\mathbf{R}_X = \mathbf{R}_X\mathbf{R}_B \tag{2.17}$$

$$(\mathbf{R}_A - \mathbf{I})\mathbf{t}_X = \mathbf{R}_X\mathbf{t}_B - \mathbf{t}_A \tag{2.18}$$

As a rotation matrix is orthogonal, equation (2.17) can be written as follows:

$$\mathbf{R}_A = \mathbf{R}_X\mathbf{R}_B\mathbf{R}_X{}^T \tag{2.19}$$

This similarity transformation reveals that $\mathbf{R}_A$ and $\mathbf{R}_B$ share the same eigenvalues. As every rotation matrix has a unit eigenvalue with a value of 1, by letting $\mathbf{n}_B$ be the eigenvector associated with this unit eigenvalue for $\mathbf{R}_B$, multiplying it with equation (2.17) may yield the following:

$$\mathbf{R}_A\mathbf{R}_X\mathbf{n}_B = \mathbf{R}_X\mathbf{R}_B\mathbf{n}_B \tag{2.20}$$

$$\mathbf{R}_A\mathbf{R}_X\mathbf{n}_B = \mathbf{R}_X\mathbf{n}_B$$

Here, it is apparent that $\mathbf{n}_A = \mathbf{R}_X\mathbf{n}_B$. In summary, solving $\mathbf{AX = XB}$ is equivalent to solving $\mathbf{n}_A = \mathbf{R}_X\mathbf{n}_B$ and equation (2.18). There are approaches that decouple the problem by solving the rotation before the translation [64-66], while methods that solve both at the same time also exist [67, 68]. In any circumstance, three different endoscope poses are required to compute a unique $\mathbf{X}$.

2.4 3D-TO-3D RIGID REGISTRATION IN ENDOSCOPY

In both conventional surgical navigation and AR-assisted guidance, the patient's pre-operatively obtained images or models must be aligned with the target anatomy before the start of a surgery. To achieve this alignment, a 3D-to-3D rigid registration process is required. The simplest form of registration is manual registration, where the patient's images or 3D models are manually shifted, rotated, or scaled until an ideal alignment is achieved. However, alignment quality is then highly dependent on the user's technique. There is no systematic mechanism to ensure that the clinical standard for safety is met,

香港大學圖書館 HKU Libraries

which is a maximum TRE of 2 mm as mentioned in **Section 2.1.3**. In addition, the process of manual registration is time-consuming, which can affect the smoothness of a surgical workflow due to a lengthened setup. As a result, alternative registration methods such as i) point-based registration and ii) surface-based registration are adopted in most commercially available navigation systems and relevant academic studies.

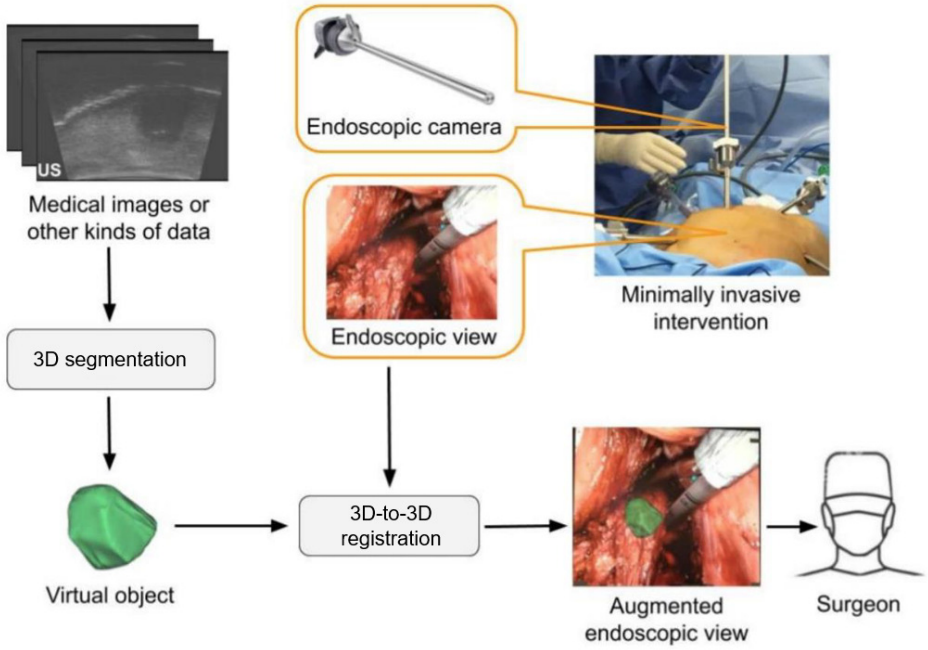

Fig. 2.26 AR implementation in endoscopic surgery. Through segmentation, 3D virtual anatomical models are acquired from 2D patient scans. A rigid registration process is then required to align 3D models with the patient. Target anatomy, endoscopes, and instruments are localized in real-time during surgery by EM or optical tracking systems. With the addition of camera calibration and hand-eye calibration, endoscopic overlay of virtual anatomical models is achieved. **Image source: [69].**

2.4.1 Point-based Method

Being the "golden standard" for rigid registration [70], point-based registration can achieve sub-millimeter accuracy in dry-lab conditions [71, 72]. The purpose of point-based rigid registration is to find a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ that can best align i) a set of points $\{x_i\}$ in the pre-operative image reference frame and ii) a corresponding set of points $\{y_i\}$ on the patient in the reference frame of the tracking device. These "points" are usually called "fiducials", which can be natural anatomical landmarks or artificially made landmarks. In a study on registration accuracy of an ENT AR navigation system conducted by Winne *et al.* [46], four anatomical fiducials were chosen for point-based registration. Nonetheless, it was concluded that registration based on anatomical fiducials introduced

香港大學圖書館 HKU Libraries

more errors compared to artificial fiducials. A possible explanation is that rigid bony structures assigned as fiducials are usually covered by other soft tissues. They also suggested inserting bone screws on the patient to achieve registration. However, this approach is highly invasive and will greatly hinder clinical workflows. Finally, point-based registration assumes the patient remains static once the registration is finished. Any movement of the patient during surgery (without a DRF tracking the patient) would make the registration invalid, leading to the need for a repeated registration process.

2.4.2 Surface-based Methods

To perform surface-based registration, point clouds or 3D meshes of the patient's anatomical surfaces must first be generated e.g. through monocular simultaneous localization and mapping (SLAM) [24] or by stereo-based methods [73]. Next, surface matching is implemented between these point clouds or 3D meshes and pre-operatively obtained 3D anatomical models. Unlike point-based registration, surface-based registration also takes into account the deformation of anatomy. Though, it is challenging to perform registration with partial surface data to give a robust registration [1], especially when the endoscope navigates inside the patient's body such as the narrow nasal passages in an endoscopic sinus surgery.

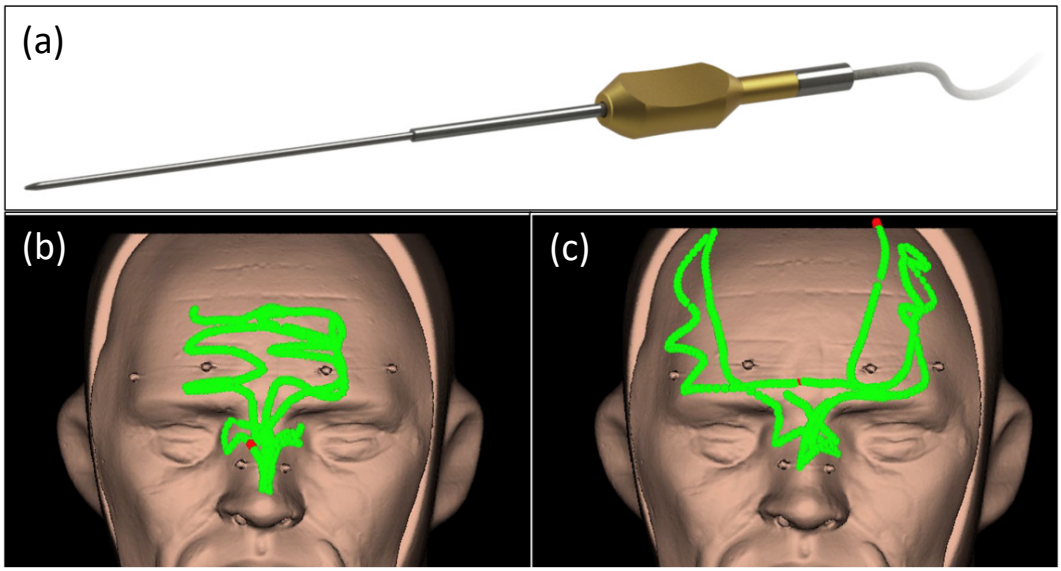

Fig. 2.27 (a) EM pointer tracker (Stryker, USA). **Image source: Stryker**; Surface-based rigid registration by drawing **(b)** narrow field registration contours and **(c)** wide field registration contours using the pointer tracker**.** Wide field contours result in a higher registration accuracy than narrow field contours. **Image source: [70].**

香港大學圖書館 HKU Libraries

A more commonly adopted surface-based method in skull base or ENT navigation systems is contour recording, as depicted in **Fig. 2.27b** and **Fig. 2.27c**. Using a tracked pointer as shown in **Fig. 2.27a**, the surgeon traces along the face of the patient. The contour is then matched with the pre-operatively segmented 3D model of the skull to compute a rigid transformation matrix. This matrix registers pre-operatively segmented anatomies onto the skull. Talmadge *et al.* [70] carried out a study on contour map point distribution, revealing that the choice of contour pattern affects registration accuracy. To be specific, a contour that spreads across a larger area results in a higher registration accuracy than a narrower contour. Similar to point-based registration, surface-based registration by contour recording also becomes invalid once the head is moved. Therefore, most skull base or ENT surgical navigation systems include a sensor that is mounted on the forehead of the patient, which acts as a DRF that tracks movements of the head, as depict in **Fig. 2.20c**. Therefore, the contour recording process registers pre-operative information to the DRF instead of directly registering to the patient.

Despite its widespread usage on the commercial market, the accuracy of contour-based registration has not been well studied [70]. In a recent study on neuronavigation by Mongen *et al.* [74], registration by surface matching was shown to produce twice the error of point-based registration when using commercially available systems. Specifically, the TRE reported after point-based registration was 2.49 ± 0.86 mm, while that of surface matching was 5.35 ± 1.64 mm. Registration error by surface matching may be due to i) skin movement during contact with the tracked pointer, ii) nasogastric tube placement, iii) presence of hair and iv) facial distortion in circumstance of decreased muscle tone.

2.5 QUANTIFICATION OF REGISTRATION AND OVERLAY ERROR

Navigation accuracy is always the first and foremost criterion of a safe surgical navigation system. No matter how many visual aids or guiding functionalities a navigation system has, subpar accuracy would compromise the benefits of surgical navigation and may result in harm to the patient due to accidental damage to critical anatomical structures. From the perspective of the users, error quantification is necessary to indicate if a system is safe and reliable. From the perspective of system developers, quantified errors are essential

香港大學圖書館 HKU Libraries

feedback information for error rectification. Error quantification is also a major factor that determines if a navigation system can be granted regulatory approval such as through the Food and Drug Administration (FDA). In this section, several commonly utilized error quantification metrics regarding rigid registration accuracy are first introduced.

For AR-assisted surgical guidance, registration accuracy alone is not sufficient for describing spatial or temporal misalignment between augmented virtual objects and physical anatomy. Temporal misalignment may occur when there is a latency discrepancy between the streaming of endoscopic images and the streaming of tracking sensors' data, leading to a feeling of augmented objects "floating" unstably on the endoscopic view. Detailed discussion about temporal misalignment is given in **Chapter 4**. In this section, error quantification metrics that specifically describe spatial misalignment in AR-assisted surgical guidance are introduced.

2.5.1 Rigid Registration Error

There are three common quantities for assessing registration accuracy, namely i) fiducial localization error (FLE), ii) fiducial registration error (FRE), and iii) target registration error (TRE). FLE means the distance between the measured fiducial position in the pre-operative image space and the true position of the fiducial in the physical space. FRE means the distance between the measured fiducial position in the physical space and the measured fiducial position in the pre-operative image after it is registered to the physical space. TRE means the distance between the position of an anatomical target (after being registered to the physical space) and its true position in the surgical field.

Currently, TRE is the standard for assessing registration accuracy because this metric is of direct surgical interest that reveals how accurate an instrument can navigate in the surgical field [75]. In particular, the "golden standard" for a navigation system to provide a safe guidance is to achieve a TRE of 2 mm or less [47]. In practice, TRE can only be directly determined when the surgeon compares the actual anatomical structure with the navigation localization [75]. Unlike TRE, FRE can be examined directly after registration, which is given by:

$$\mathbf{FRE}^2 = \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{R}x_i + \mathbf{t} - y_i \|^2 , \tag{2.21}$$

香港大學圖書館 HKU Libraries

where rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ are computed by optimally aligning fiducial set $\{x_i\}$ in the pre-operative image frame to its correspondent fiducial set $\{y_i\}$ on the patient. N is the number of fiducials, usually more than six [75]. FRE can easily be computed because vector $\mathbf{FRE}_i = \mathbf{R}x_i + \mathbf{t} - y_i$ is essentially the residual displacement error between y_i and the corresponding x_i that is registered from the image space to the physical space [1]. Although being the only directly measurable quantity by a navigation system [75], FRE has been statistically proven to be uncorrelated with TRE [76]. Thus, surgeons may still take reference from any FRE values provided by a system, but they should always be cautious and have their own judgment rather than heavily relying on the system because a low FRE value does not guarantee a low TRE value.

As it is not practical to directly measure TRE during a clinical procedure, various models have been proposed to predict this metric [76-80]. Fitzpatrick and his team, one of the pioneers who contributed to defining TRE in 1992 [81], derived the relationship between FLE and TRE in 1998 [80]. By their definition, for a zero-mean and normally distributed FLE, expected value of the squared TRE magnitude is:

$$\left\langle \mathbf{TRE}^2 \right\rangle = \frac{\left\langle \mathbf{FLE}^2 \right\rangle}{N}\left(1+\frac{1}{3}\sum_{k=1}^{3}\frac{d_k^2}{f_k^2}\right), \tag{2.22}$$

where $\langle\ \rangle$ denotes expected value, N is the number of fiducials, d_k is the distance from some target point to a principal axis k of the fiducial configuration, f_k is the root-mean-square (RMS) distance from fiducials to a principal axis k of the fiducial configuration. From this equation, it is easily observable that having more fiducials for registration may in theory reduce TRE. Also, at the centroid of the fiducials, TRE is minimized. Fitzpatrick *et al.* [80] also concluded that a higher spread of fiducials can result in a lower TRE. However, when this relationship is utilized to estimate TRE from FLE, one should be aware that this relationship assumes FLE values of different fiducials are independent of one another. In reality, this assumption might not always hold, because they may all be affected by systematic tracking and calibration errors, as well as errors attributed by tissue motion. This can potentially lead to an underestimated TRE [82].

Nonetheless, TRE can be measured in a laboratory setting with the means to directly measure the location of targets. When performing rigid registration, "fiducials" are natural anatomical landmarks or artificial markers inserted on the target tracking entity. To compute TRE, extra fiducial markers are attached that act as "targets" for TRE computation

香港大學圖書館 HKU Libraries

instead of "fiducials" that serve for registration. Therefore, similar to FRE, TRE is given by:

$$\mathbf{TRE}^2 = \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{R}x_i + \mathbf{t} - y_i \|^2 \tag{2.23}$$

Same as equation (2.21), rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ are computed from rigid registration. $\{x_i\}$ is a set of "targets" in the image frame while $\{y_i\}$ is the corresponding set in the physical space. In this case, "targets" serve for accuracy validation instead of rigid registration. N is the number of targets at which position measurements are taken for accuracy validation.

2.5.2 Re-projection Error (RPE)

TRE can sufficiently reflect overall registration accuracy, however, it cannot directly describe overlay accuracy of an AR-assisted guidance system. In simple terms, overlay accuracy depicts how accurate a virtual object is overlaid onto a real target object on a display. A low TRE value does not guarantee a high overlay accuracy because overlay accuracy is also affected by error induced by an erroneous camera calibration or hand-eye calibration.

To quantify overlay accuracy, a prerequisite is to have features that are visible on both the virtual overlaid objects and the surgical field [83]. Under a laboratory setting, this can be simply prepared by inserting artificial targets on anatomy. The virtual correspondent of an inserted artificial target can then be derived from CT or MRI scans of that anatomy inserted with targets. As depicted in **Fig. 2.28**, when both a real target and an overlaid target are visible on the endoscopic view during accuracy validation, the overlay error can be expressed as the pixel distance between the real target and the overlaid target. However, this error quantification alone does not consider the distance of a target to the endoscope. Assuming the geometric distance between a real target and an overlaid target is fixed, the observed error in terms of pixels would naturally increase as the endoscope gets nearer to the target. As such, pixel error does not render surgeons with a meaningful piece of information about the physical geometric error. To fill this gap, Thompson *et al.* [84] proposed a metric to quantity overlay accuracy, namely re-projection error (RPE).

香港大學圖書館 HKU Libraries

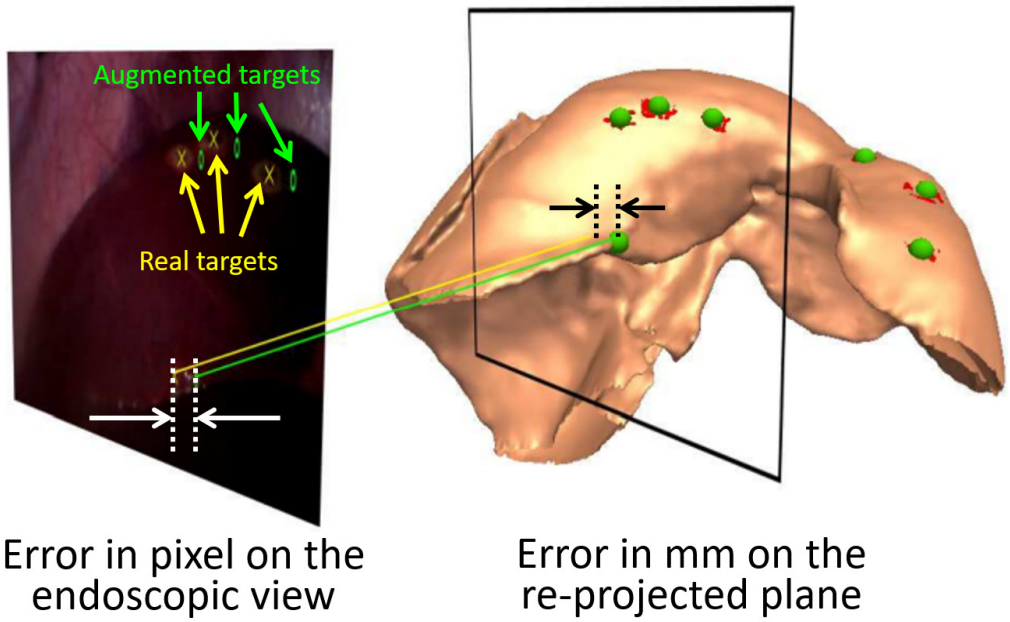

Fig. 2.28 Overlay error in terms of distance between the augmented target and real target. Error is in units of pixel on the endoscopic view and in units of mm on the re-projected plane. The experimental setting in this example is a porcine liver. **Image source: [83].**

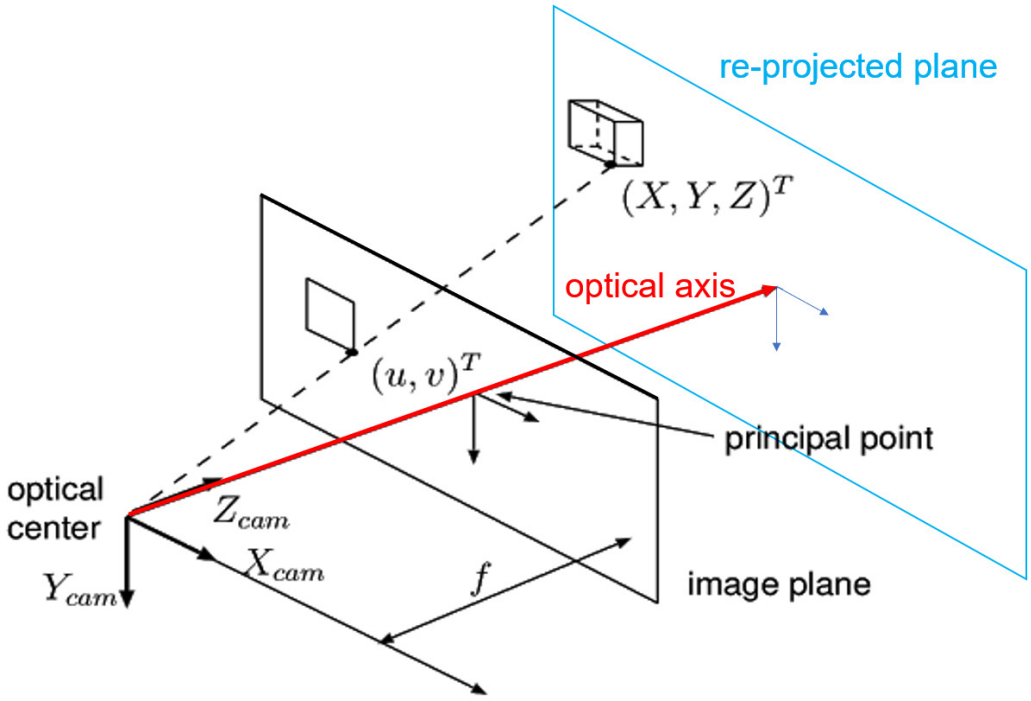

Fig. 2.29 Re-projected plane as illustrated in a pinhole camera setting. The re-projected plane is parallel to the image plane and coincides with the target point on the overlaid object. **Image source: [85].**

香港大學圖書館 HKU Libraries

In essence, RPE is computed by "re-projecting" the pixel error from the image plane onto a plane that i) coincides with the overlaid target and ii) is parallel to the image plane, illustrated as the "re-projected plane" in **Fig. 2.29**. RPE is thus an error in terms of a

physical length (mm) instead of pixel. Under this definition, RPE for a point target is given by:

$$\text{RPE} = err_p \left(\frac{d}{f_p} \right), \tag{2.24}$$

where err_p is the pixel error measured on the image plane, f_p is the focal length of the endoscope expressed in units of pixel, d is the distance (in mm) between the endoscope optical center and the re-projected plane, or in other words, the depth of the point target with respect to the endoscope's frame of reference. It should be noted that, although RPE reflects both registration and calibration errors, this metric does not reveal any geometric error in the direction along the endoscope optical axis (z-direction in **Fig. 2.29**).

2.6 CONCLUSION

This chapter first discusses the development and the state-of-the-art of conventional surgical navigation in different medical specialties. Next, by enabling direct overlay of patient-specific 3D anatomical models onto the surgical field, conventional surgical navigation is incorporated with AR capability and becomes AR-assisted surgical guidance. Greater focus has been placed on AR-assisted ENT surgery, which is closely related to the study that is introduced in **Chapter 4**. Basic technical components of both conventional surgical navigation and AR-assisted surgical guidance systems are also reviewed, which included commonly used tracking modalities, methods of rigid registration, system calibration and spatial error quantification. Although existing standard sensing techniques have shown promising robustness in ideal conditions, limitations such as i) EM tracking interference in the presence of ferromagnetic or conductive materials and ii) optical tracking line-of-sight issue may hinder system accuracy and stability. In the coming chapters, innovative sensing alternatives that can be incorporated in an endoscopic procedure are explored, namely "*Visual-strain Fusion for Camera Tracking"* (**Chapter 3**) and "*Real-to-virtual Domain Transfer-based Depth Estimation"* (**Chapter 4**). With these sensing alternatives, we aim to reduce the reliance on external tracking modalities while achieving clinically safe tracking accuracy and stability.

香港大學圖書館 HKU Libraries

CHAPTER 3
VISUAL-STRAIN FUSION FOR CAMERA TRACKING

3.1 INTRODUCTION AND RELATED WORK

Image processing algorithms have significantly extended the practical value of the eye-in-hand camera, enabling and promoting its applications for quantitative measurements. Pose estimation, one of such applications, has potential in surgical navigation or in scene reconstruction. However, fully vision-based estimation methods sometimes encounter difficulties handling cases with deficient features. Single-core fiber inscribed with fiber Bragg gratings (FBGs) is capable of configuration-related strain measurement when integrated on a soft manipulator or a flexible endoscope. In this study, visual information is fused with sparse strain data collected from FBG fiber to facilitate pose estimation of a continuum robot. An improved extreme learning machine (ELM) algorithm with selective training data updates is implemented to establish and refine the FBG-based pose estimator online. The integration of FBG-based pose estimation can improve sensing robustness by reducing the number of times that visual tracking is lost given moving visual obstacles and varying lighting. In particular, this integration solves pose estimation failures under full occlusion of the tracked features or complete darkness.

香港大學圖書館 HKU Libraries

Recent advances in computer vision enable the detection of the robot configuration in unstructured environments [86, 87], similar to human visual perception that allows us to interpret body movement relative to our surroundings. In computer vision, camera pose estimation is a fundamental problem that has been widely studied in the areas of structure from motion (SfM), visual odometry (VO) [88], simultaneous localization and mapping (SLAM), and even commonly applied in augmented reality (AR) [89] as well as autonomous navigation [90]. Pose estimation by means of temporally coherent features in a sequence of 2D/3D images [91] can avoid the complicated integration of additional positional sensors.

However, feature-based estimation using cameras are inherently subject to the image quality, which is inevitably affected by unstable light exposure, vision occlusion, and rapid viewpoint changes [88]. This weakness is made more apparent with cameras used in the eye-in-hand configuration, where the camera (i.e., the *eye*) is fixed on the robot end-effector (the *hand*), to see its surroundings. Although the eye-in-hand approach is intuitive and provides active visual perception, it requires effective end-effector movement for pose detection [92, 93], and greatly demands for consistent motion patterns. In addition, as the camera usually points closer towards objects of interest [94], the effect of local lighting variations and specular reflection will be dominant in the camera view. To compensate for the pose error induced by the lack of high-quality image features, fusing computer vision data with other sensing feedback has become a promising option.

The most prevalent type of fusion approach is to integrate cameras with inertial measurement units (IMUs) [89, 95], that is, the visual-inertial system (VINS), which has been generally developed for rigid-link robots. Detected acceleration and angular velocity could be utilized by employing statistical filtering techniques such as extended Kalman filters [89], or fusion methods that are learning-based, e.g., convolutional neural networks (CNN) [96] and long short-term memory (LSTM) [97] . Although the extrinsic calibration and accumulated drift in VINS were widely discussed [98-100], mechanical integration of IMUs would require tailor-made or compact packing with the camera at the tip of a soft robot or a flexible endoscope, whereas rigid robots have the freedom to fix the IMUs anywhere along their links. To this end, there remains a demand for alternative sensors that can directly measure the pose of soft robots and flexible endoscopes.

香港大學圖書館 HKU Libraries

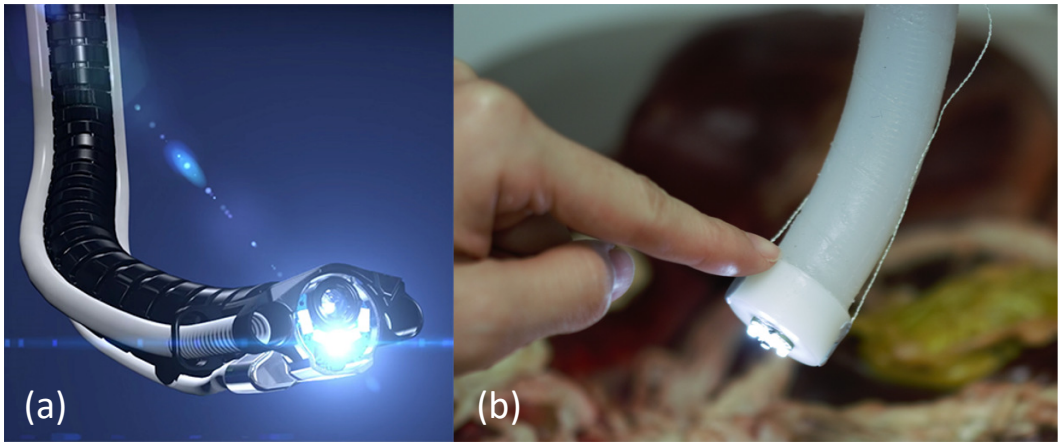

Fig. 3.1 Application of flexible endoscope in robot-assisted surgery. **(a)** Medrobotics Flex® Robotic System for otolaryngology and colorectal surgeries. **Image source: [101]**; **(b)** Camera and LED modules mounted at the tip of a soft manipulator for obtaining an endoscopic view.

As illustrated in **Fig. 3.1**, soft robots and flexible endoscopes involve deformation, of which the strain changes on the body surface would give strong cues to estimate its configuration. Real-time strain sensing achieved with fiber Bragg grating (FBG) optical fiber is a potential candidate that can utilize these strain changes for feedback [102-104]. FBG sensors provide several advantages over electronic strain sensors, including the capability for dense strain measurements with a single connection, and insusceptibility to water submersion and electromagnetic fields. As a result, FBGs have been investigated in thin surgical tools such as biopsy needles [105], or even in magnetic resonance imaging (MRI) environments [106, 107]. Optical frequency domain reflectometry (OFDR) interrogation applied on continuous-grating multi-core fiber is one form of FBG sensing that is capable of stand-alone 3D curvature sensing with a single fiber and is typically integrated in manipulators or instruments with strict diameter requirements [108, 109]. For pose estimation of fluid-driven soft robots, single-core optic FBGs using the common wavelength division multiplexing (WDM) method would be more appropriate considering its advantages of higher sensing sampling rate (100-3,000 Hz) and significantly lower cost. When helically wound onto robot surface [110, 111], the fiber can sensitively detect small deformations at high frequencies enabling reliable closed-loop robot control. Task space control of the soft robot using absolute FBG-detected strain would be more reliable than using IMU feedback which needs to calculate the integral of relative acceleration/velocity. The working principle of FBGs is illustrated in **Fig. 3.2**.

The mapping from FBGs measurement to continuum robot configuration can be established using either analytical modeling [107] or machine learning [112] approaches. Sefati *et al.* [111] had compared their tip positional sensing accuracy of a planar bending continuum

香港大學圖書館 HKU Libraries

manipulator equipped with *three* parallel FBG fibers. The results demonstrated improved sensing performance using the data-driven method without prior information of the FBG allocation. In learning-based methods, positional markers need to be employed as the ground truth to complete the mapping. In our previous work [113], we also proposed a flexible surface sensing system, in which only *one* single-core fiber inscribed with FBGs was embedded in a soft substrate. Offline learning was needed to "train" the mapping between FBG strains and the surface morphology detected by motion capture cameras.

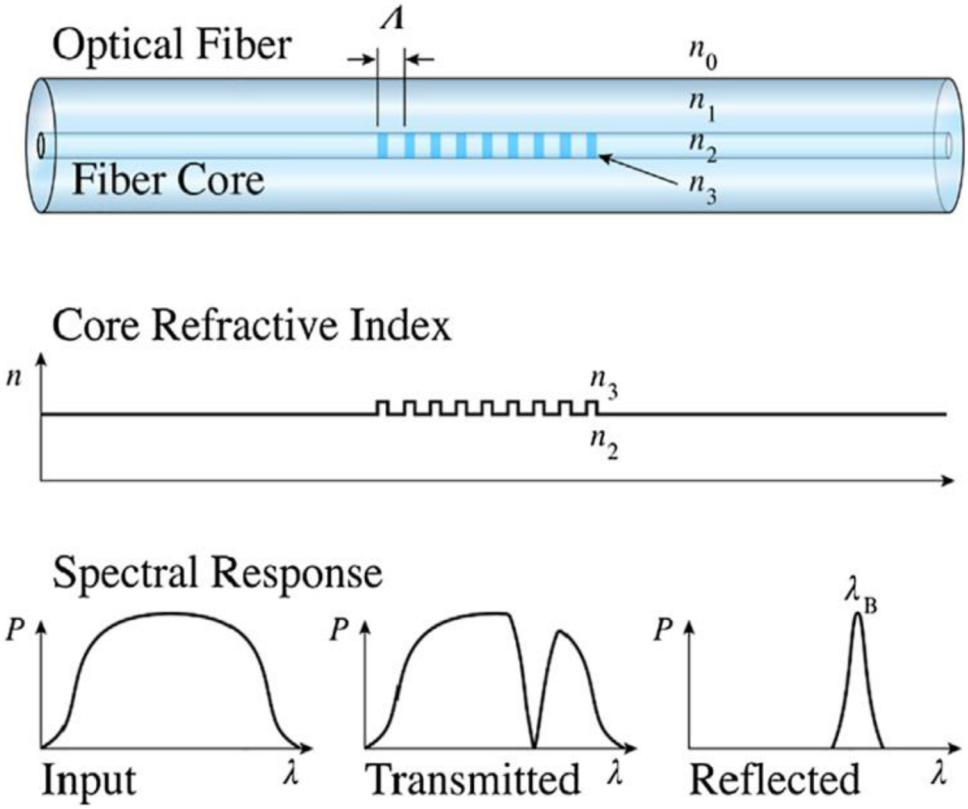

Fig. 3.2 Working principle of FBGs. Gratings implies periodic changes of refractive index in the core of a fiber. Different spacing between gratings result in reflected light with different wavelengths. It subsequently infers strain measurement from reflected wavelength variations. **Image source: [114].**

Considering the small form factor of FBG fibers and their ease of integration with devices/instruments, researchers have also aimed to leverage them with various camera configurations. In other previous work, we employed a single-core FBG fiber on a continuum robot to enhance 2D motion estimation and path tracking in the endoscopic camera view [115]. However, these types of 3D shape and 2D motion estimators need to be trained by additional positional sensors in advance, heavily relying on prior data exploration and accurate ground truth data. Alambeigi *et al.* [116] also proposed a sensor

香港大學圖書館 HKU Libraries

fusion technique to address the shape/position estimation of continuum robots. As an illustration, the intermittent external information provided by an eye-to-hand camera calibrated the continuous imperfect FBG feedback to achieve accurate 2D positional sensing in obstructed environments. This work is one example of the few visual-strain fusion combinations for positional sensing, with even fewer examples using cameras integrated into the robot tip for eye-in-hand feedback.

Therefore, our aim in this study is to utilize a self-contained camera to serve as the pose ground truth in ordinary cases, while the online initialized and updated FBG sensor can be fused to settle estimation error caused by poor-quality images. No external sensors would be applied in the algorithm since we would like to simplify the employed devices. A widely adopted sensor may be used in the test but just to prove the accuracy of camera-based pose estimation as the training ground truth. The sensing dimension is also extended from 3D position/shape to 6D pose, offering more flexibility in robot applications such as spatial image stitching.

3.2 POSE ESTIMATION OF SOFT MANIPULATOR

To improve the eye-in-hand pose estimation stability by integrating a single-core FBG fiber, the FBGs can be evenly distributed on the robot body. In our case, the fiber is helically wrapped on the cylindrical surface, thus reflecting the robot's overall deformation via wavelength shifts. Helically wrapping the fiber allows us to use one single fiber to cover the cylindrical surface continuously, instead of using multiple fibers. More importantly, helical wrapping enables deformation detection of elongation, compression, and torsion, which cannot be achieved by routing the fibers just along the longitudinal axis. We hypothesize that the camera-based estimation in feature-abundant scenarios is the primary choice of sensing information, which may suffer from inadvertent poor image quality. For such circumstances, the FBG-based pose estimation model, trained by accurate camera-based estimations, could act as a stable backup and guarantee the operation of the entire framework. **Fig. 3.3** shows the configuration of the soft robot and wavelength shifts in forms of plots.

香港大學圖書館 HKU Libraries

Accredited to Dr. Xiaomei Wang, the first author of the paper related to this chapter, the idea of visual-strain fusion-based robot pose estimation was proposed. She was also the

major contributor of the training and testing phases of the pose estimation network, assisted by Mr. Jing Dai and me mainly on experimental setup, data collection and processing.

3.2.1 Definition of Task Space

The eye-in-hand camera shares the same pose with the robot tip because it is fixed on the robot end-effector. The origin of the global coordinate system (green coordinate frame in **Fig. 3.3a**) is taken at the initial position of the end-effector when it is not actuated, and the robot central axis is set as the z-axis with downward as the positive direction. In the SLAM algorithm, the initial pose is taken by default as the origin of its measuring frame. The camera pose estimated by SLAM at time step k is defined as $\boldsymbol{z}_c = \left[\boldsymbol{p}_c(k) \quad \boldsymbol{q}_c(k)\right]$, where $\boldsymbol{p}_c(k) \in \mathbb{R}^3$ represents a position and $\boldsymbol{q}_c(k) \in \mathbb{R}^4$ is an orientation expressed in quaternion. The actual pose becomes $\boldsymbol{z} = \left[\boldsymbol{p}(k) \quad \boldsymbol{q}(k)\right]$. Under stable and smooth movement, the SLAM estimation $\boldsymbol{z}_c$ in feature-abundant camera views can be deemed as the approximated pose of robot end-effector, i.e., $\boldsymbol{z}_c \approx \boldsymbol{z}$. It is worth noting that SLAM would not be the sole option to provide $\boldsymbol{z}_c$; the feedback from other pose measuring approaches will also be valid to train the FBG estimation model, such as EM tracking.

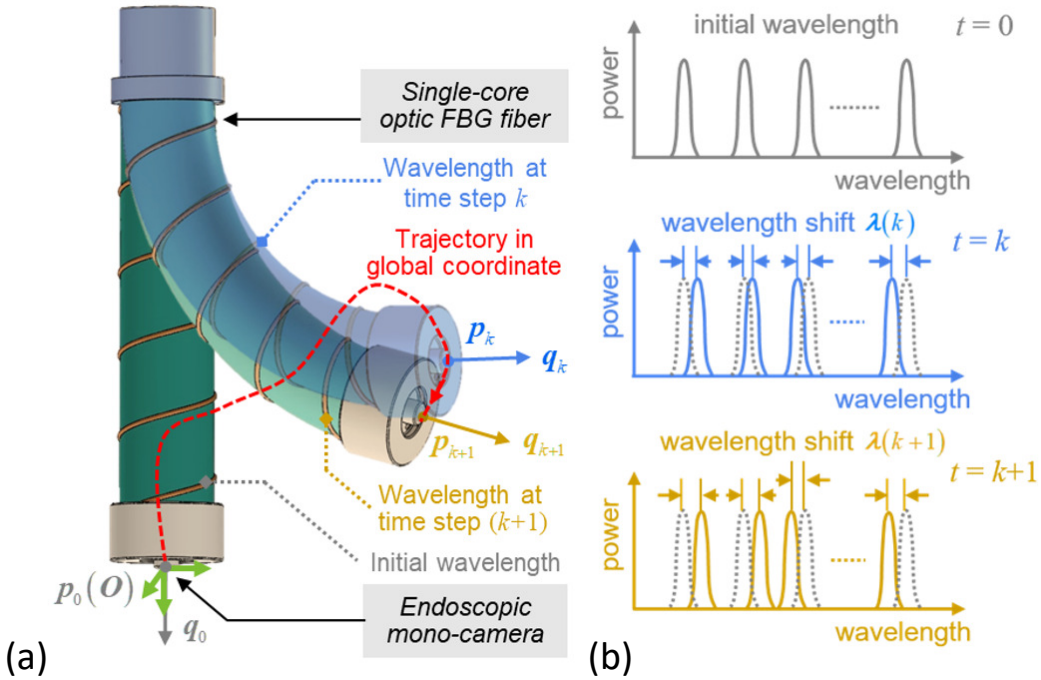

Fig. 3.3 Single-core FBG fiber wrapped on a soft continuum robot. **(a)** Camera poses obtained at each time step k based on SLAM algorithm. **(b)** FBG wavelengths shifted correspondingly, i.e., from $\boldsymbol{\lambda}(k)$ to $\boldsymbol{\lambda}(k+1)$.

香港大學圖書館 HKU Libraries

Adopting SLAM-based pose is advantageous in the sense that no external sensing devices are required other than the integrated camera itself. At equilibrium state, the actuator input is represented as $\boldsymbol{u}(k) \in \mathbb{R}^m$, where the dimension of actuation space is denoted by m. In essence, an actuation command $\Delta\boldsymbol{u}(k)$ is computed to achieve a desired movement denoted by $\Delta\boldsymbol{p}^*(k)$ and $\Delta\boldsymbol{q}^*(k)$. Next, the single-core FBG fiber wrapped helically on the continuum robot gives strain measurements. l wavelength/strain measurement points are provided by the multiplexing l units of FBGs inscribed, which are independent of each other. Difference between the original wavelength vector $\boldsymbol{\lambda}_0$ at the initial robot configuration and wavelength vector at time step k is depicted as a wavelength shift vector $\boldsymbol{\lambda}(k) \in \mathbb{R}^l$ as illustrated in **Fig. 3.3b**.

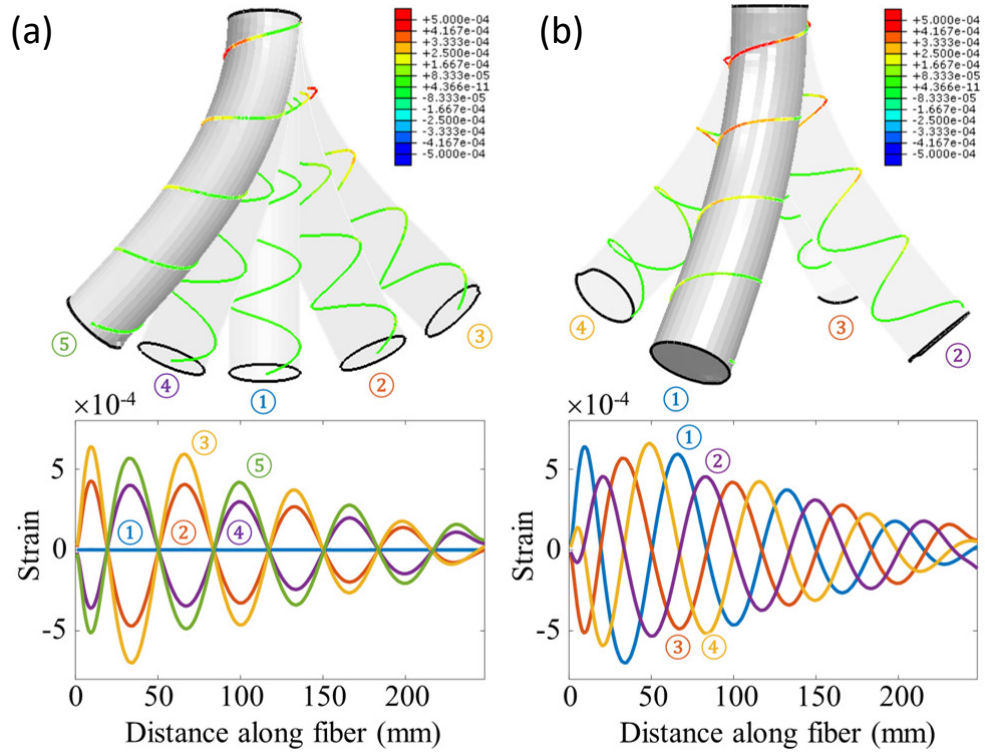

Fig. 3.4 Finite element modeling (FEM) of the strains helically distributed along an elastic continuum manipulator. **(a)** Strains varying in amplitude when the manipulator bends on the same plane/direction. **(b)** Strains under four different-bending directions distinguished by their phase differences.

3.2.2 Learning-based Pose Estimation by Fiber Bragg Grating (FBG)

The consideration of optic fiber integration is that the wavelength shift/strain sequence of all FBGs should be mapped uniquely to reflect the end-effector pose, but not altering the

香港大學圖書館 HKU Libraries

original soft robot mechanical properties. Details about the fiber placement can be found in our previous work [115]. Simply, the fiber was wound helically on the robot body. No rigorous requirements on the wrapping structure were set, as long as the FBGs can be dispersedly distributed. Distances between adjacent turns could also vary without strict consistency. Details of the fiber used in our experiments can be found in **Section 3.3.1**. The pose estimation $\boldsymbol{z}_c(k)$ and the wavelength shift $\boldsymbol{\lambda}(k)$ at time step k are obtained at 20~50 Hz, where the exact sampling rate depends on computer and camera performance. Then, FBG feedback is used to train a pose estimation model, in which a mapping relationship between camera pose and FBG feedback is established. In the following paragraphs, the i) training, ii) prediction, and iii) updating phases of the pose estimation method are explained in detail. This workflow was developed by the first author, Dr. Xiaomei Wang, of the paper related to this chapter.

Training: With a command sequence $\boldsymbol{U}=\left[\boldsymbol{u}(1) \quad \boldsymbol{u}(2) \quad \cdots \quad \boldsymbol{u}(N_0)\right]$ that has N_0 steps in total, the robot is first actuated. The pose estimation model is initialized by collecting a few sample pairs. The corresponding FBG wavelength shift sequence $\boldsymbol{\Lambda}$ and camera pose sequence $\boldsymbol{Z}_c$ are:

$$\boldsymbol{\Lambda}=\left[\boldsymbol{\lambda}(1) \quad \boldsymbol{\lambda}(2) \quad \cdots \quad \boldsymbol{\lambda}(N_0)\right] \in \mathbb{R}^{l\times N_0}$$
$$\boldsymbol{Z}_c=\left[\boldsymbol{z}_c(1) \quad \boldsymbol{z}_c(2) \quad \cdots \quad \boldsymbol{z}_c(N_0)\right] \in \mathbb{R}^{7\times N_0}$$

The following mapping is to be learned:

$$\boldsymbol{z}_c(k)=f\left(\boldsymbol{\lambda}(k)\right) \tag{3.1}$$

Having N_0 distinct training samples composing of input $\boldsymbol{\Lambda}$ and output $\boldsymbol{Z}_c$, we can express the output of a SLFN [117, 118] with N_0 hidden nodes by:

$$\boldsymbol{o}_j=\sum_{i=1}^{N}\boldsymbol{\beta}_i\phi_i\left(\boldsymbol{\lambda}(j)\right)=\sum_{i=1}^{N}\boldsymbol{\beta}_i\phi\left(\boldsymbol{\lambda}(j),\boldsymbol{a}_i,b_i\right),\ j=1,2,\ldots N_0 \tag{3.2}$$

where $\boldsymbol{a}_i=\left[a_{i1} \quad a_{i2} \quad \cdots \quad a_{il}\right]^{\mathrm{T}}$ and $\boldsymbol{\beta}_i=\left[\beta_{i1} \quad \beta_{i2} \quad \cdots \quad \beta_{i,7}\right]^{\mathrm{T}}$ are weighting vectors of the input and output nodes. $\phi\left(\boldsymbol{\lambda}(j),\boldsymbol{a}_i,b_i\right)$ is an activation function, with b_i

香港大學圖書館 HKU Libraries

being the threshold of the i^{th} node. Specifically, the activation function is a radial basis function (RBF):

$$\phi_i\left(\boldsymbol{\lambda}(j)\right)=\phi\left(\boldsymbol{\lambda}(j),\boldsymbol{a}_i,b_i\right)=\exp\left(\frac{\left\|\boldsymbol{\lambda}(j)-\boldsymbol{a}_i\right\|^2}{b_i}\right) \tag{3.3}$$

By rewriting $\boldsymbol{a}_i$ and $\boldsymbol{\lambda}(j)$ as an inner product $\boldsymbol{a}_i\boldsymbol{\lambda}(j)$, equation (3.2) becomes:

$$\boldsymbol{\Phi}\boldsymbol{\beta}=\boldsymbol{O}, \tag{3.4}$$

where

$$\boldsymbol{\Phi}=\begin{bmatrix}\phi(\boldsymbol{a}_1\boldsymbol{\lambda}(1)+b_1) & \cdots & \phi(\boldsymbol{a}_N\boldsymbol{\lambda}(1)+b_N)\\ \vdots & \cdots & \vdots\\ \phi(\boldsymbol{a}_1\boldsymbol{\lambda}(N_0)+b_1) & \cdots & \phi(\boldsymbol{a}_N\boldsymbol{\lambda}(N_0)+b_N)\end{bmatrix}\in\mathbb{R}^{N_0\times N}$$

$$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}} & \boldsymbol{\beta}_2^{\mathrm{T}} & \cdots & \boldsymbol{\beta}_N^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\in\mathbb{R}^{N\times 7},\ \boldsymbol{O}=\begin{bmatrix}\boldsymbol{o}_1^{\mathrm{T}} & \boldsymbol{o}_2^{\mathrm{T}} & \cdots & \boldsymbol{o}_{N_0}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\in\mathbb{R}^{N_0\times 7}$$

By setting appropriate parameters in the hidden layer output matrix $\boldsymbol{\Phi}$, the following is obtained:

$$\boldsymbol{\Phi}\boldsymbol{\beta}=\boldsymbol{Z}_c \tag{3.5}$$

Next, the "appropriate parameters" are to be trained by minimizing the cost function $\left\|\boldsymbol{O}-\boldsymbol{Z}_c\right\|$. To elaborate, a vector containing $\boldsymbol{a}_i^{\mathrm{T}}$, $\boldsymbol{\beta}_i^{\mathrm{T}}$, and b_i^{T}, where i = 1, 2, …, N, is to be computed. In this study, a least-square approach is adopted to minimize the cost function. By utilizing an ELM algorithm with a given input set $\boldsymbol{\Lambda}$, equation (3.5) can be solved in a least-square sense to give a resultant $\hat{\boldsymbol{\beta}}$:

$$\left\|\phi(\boldsymbol{A},\boldsymbol{b})\hat{\boldsymbol{\beta}}-\boldsymbol{Z}_c\right\|=\min_{\boldsymbol{\beta}}\left\|\phi(\boldsymbol{A},\boldsymbol{b})\boldsymbol{\beta}-\boldsymbol{Z}_c\right\|, \tag{3.6}$$

where $\boldsymbol{A}=\{\boldsymbol{a}_1\ \ \boldsymbol{a}_2\ \ \cdots\ \ \boldsymbol{a}_N\}$, and $\boldsymbol{b}=\{b_1\ \ b_2\ \ \cdots\ \ b_N\}$. Analytically, the output weights $\boldsymbol{\beta}$ is computed as follows:

$$\hat{\boldsymbol{\beta}}=\boldsymbol{\Phi}^{\dagger}\boldsymbol{Z}_c, \tag{3.7}$$

where $\boldsymbol{\Phi}^{\dagger}$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Phi}$. Now, prediction may begin based on the global nonlinear mapping model in equation (3.1) that is computed by the

香港大學圖書館 HKU Libraries

ELM. The initialization step hereto is the whole procedure of standard ELM, which is an offline training method. Its robustness is determined by the Moore-Penrose (MP) inverse, possibly resulting in low overall estimation accuracy. However, the MP inverse is only employed in the initialization phase for network weights calculation, and the subsequent online update can weaken the adverse effects of MP inverse.

Prediction: At the k^{th} time step, wavelength shift $\boldsymbol{\lambda}(k)$ is mapped to the robot end-effector pose $\boldsymbol{z}_w(k)$:

$$\boldsymbol{z}_w(k) = f\left(\boldsymbol{\lambda}(k)\right),\ k = 1, 2, \cdots \tag{3.8}$$

It is worth noting that during prediction phase, pose prediction by the model is independent of pose estimation $\boldsymbol{z}_c(k)$ from the camera. Therefore, the prediction model can be deemed as another pose estimator, of which the prediction results can undergo fusion with $\boldsymbol{z}_c(k)$.

Updating: The training phase gives a prediction vector $\boldsymbol{\beta}^{(0)}$ from training dataset $\mathbf{D}^{(0)}$ that contains N_0 distinct sample pairs of $\boldsymbol{\Lambda}$ and $\boldsymbol{Z}_c$. Equation (3.7) now becomes:

$$\boldsymbol{\beta}^{(0)} = \left(\left(\boldsymbol{\Phi}^{(0)}\right)^{\mathrm{T}}\boldsymbol{\Phi}^{(0)}\right)^{-1}\left(\boldsymbol{\Phi}^{(0)}\right)^{\mathrm{T}}\boldsymbol{Z}_c = \left(\mathbf{K}^{(0)}\right)^{-1}\left(\boldsymbol{\Phi}^{(0)}\right)^{\mathrm{T}}\boldsymbol{Z}_c \tag{3.9}$$

When the ELM receives a new training dataset $\mathbf{D}^{(1)}$ that contains N_1 distinct sample pairs, a new weighting vector $\boldsymbol{\beta}^{(1)}$ is computed as follows:

$$\boldsymbol{\beta}^{(1)} = \begin{bmatrix}\boldsymbol{\Phi}^{(0)}\\ \boldsymbol{\Phi}^{(1)}\end{bmatrix}^{+}\begin{bmatrix}\boldsymbol{Z}_c^{(0)}\\ \boldsymbol{Z}_c^{(1)}\end{bmatrix} = \left(\mathbf{K}^{(1)}\right)^{-1}\left(\boldsymbol{\Phi}^{(1)}\right)^{\mathrm{T}}\left(\boldsymbol{Z}_c^{(1)} - \boldsymbol{\Phi}^{(1)}\boldsymbol{\beta}^{(0)}\right) + \boldsymbol{\beta}^{(0)}, \tag{3.10}$$

where

$$\mathbf{K}^{(1)} = \begin{bmatrix}\boldsymbol{\Phi}^{(0)}\\ \boldsymbol{\Phi}^{(1)}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\Phi}^{(0)}\\ \boldsymbol{\Phi}^{(1)}\end{bmatrix} = \left(\boldsymbol{\Phi}^{(1)}\right)^{\mathrm{T}}\boldsymbol{\Phi}^{(1)} + \mathbf{K}^{(0)}$$

After the input of the k^{th} training dataset $\mathbf{D}^{(k)}$, the ELM model is updated as follows [119]:

$$\boldsymbol{\beta}^{(k)} = \left(\mathbf{K}^{(k)}\right)^{-1}\left(\boldsymbol{\Phi}^{(k)}\right)^{\mathrm{T}}\left(\boldsymbol{Z}_c^{(k)} - \boldsymbol{\Phi}^{(k)}\boldsymbol{\beta}^{(k-1)}\right) + \boldsymbol{\beta}^{(k-1)}, \tag{3.11}$$

where

香港大學圖書館 HKU Libraries

$$\mathbf{K}^{(k)} = \left(\mathbf{\Phi}^{(k)}\right)^{\mathrm{T}} \mathbf{\Phi}^{(k)} + \mathbf{K}^{(k-1)}$$

In consideration of the possible deteriorated camera-based estimations due to the poor image quality, it is necessary to set an activation threshold of the model updating mechanism. The re-projection error e_S in the ORB-SLAM2 algorithm can be utilized as such an indication to determine whether the newly obtained sample should be incorporated for online learning. When e_S is larger than the threshold (> 1.4), the matrix $\boldsymbol{\beta}^{(k)}$ will keep the value as in the last iteration step.

The reduction of effects from previous data when updating the ELM model is accomplished by the adjustment of some weight parameters of previous measurements. Equation (3.11) can be expressed by:

$$\boldsymbol{\beta}^{(k)} = \left(\begin{bmatrix} \mathbf{\Phi}^{(k-1)} \\ \mathbf{\Phi}^{(k)} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{\Phi}^{(k-1)} \\ \mathbf{\Phi}^{(k)} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{\Phi}^{(k-1)} \\ \mathbf{\Phi}^{(k)} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{Z}_c^{(k-1)} \\ \boldsymbol{Z}_c^{(k)} \end{bmatrix} = \mathbf{H}^{(k)} \mathbf{M}^{(k)}, \tag{3.12}$$

where

$$\mathbf{H}^{(k)} = \left[\left(\mathbf{\Phi}^{(k)}\right)^{\mathrm{T}} \mathbf{\Phi}^{(k)} + \left(\mathbf{\Phi}^{(k-1)}\right)^{\mathrm{T}} \mathbf{\Phi}^{(k-1)} \right]^{-1} \tag{3.13}$$

$$\mathbf{M}^{(k)} = \left(\mathbf{\Phi}^{(k)}\right)^{\mathrm{T}} \boldsymbol{Z}_c^{(k)} + \left(\mathbf{\Phi}^{(k-1)}\right)^{\mathrm{T}} \boldsymbol{Z}_c^{(k-1)} \tag{3.14}$$

By adding a weighting factor *w* to the variables associated with previous training datasets, expression (3.13) and (3.14) become:

$$\widehat{\mathbf{H}}^{(k)} = \left[\left(\mathbf{\Phi}^{(k)}\right)^{\mathrm{T}} \mathbf{\Phi}^{(k)} + w\left(\mathbf{\Phi}^{(k-1)}\right)^{\mathrm{T}} \mathbf{\Phi}^{(k-1)} \right]^{-1} \tag{3.15}$$

$$\widehat{\mathbf{M}}^{(k)} = \left(\mathbf{\Phi}^{(k)}\right)^{\mathrm{T}} \boldsymbol{Z}_c^{(k)} + w\left(\mathbf{\Phi}^{(k-1)}\right)^{\mathrm{T}} \boldsymbol{Z}_c^{(k-1)} \tag{3.16}$$

Then, the Sherman-Morrison formula [120] gives a recursive expression of (3.15) by:

$$\widehat{\mathbf{M}}^{(k)} = \frac{\widehat{\mathbf{M}}^{(k-1)}}{w} - \frac{\mathbf{N}^{(k)} \left(\mathbf{N}^{(k)}\right)^{\mathrm{T}}}{w\left[w + \mathbf{\Phi}^{(k)} \mathbf{N}^{(k)} \right]}, \tag{3.17}$$

where $\mathbf{N}^{(k)} = \mathbf{\Phi}^{(k)} \widehat{\mathbf{M}}^{(k-1)}$.

香港大學圖書館 HKU Libraries

3.2.3 Sensing Fusion of Camera and FBG

The re-projection error e_S in ORB-SLAM2 algorithm indicates the accuracy of pose estimation z_c by the camera. Therefore, camera-derived pose z_c and FBG-derived pose z_w can be fused with a variable weighting factor based on error e_S. The criterion for attaining the final fused pose estimation is given by:

$$z = \begin{cases} z_c, & e_S \le E_L \\ K_S\left(e_S - E_L\right) z_w - \left[1 - K_S\left(e_S - E_L\right)\right] z_c, & E_L < e_S < E_U \\ z_w, & e_S \ge E_U \end{cases} \tag{3.18}$$

where E_U and E_L are the boundaries determining whether or not to completely adopt or discard ORB-SLAM2 estimation. K_S is an adjusting factor. Regarding data collection from ORB-SLAM2, I was the main contributor on system setup and coding amendment.

3.3 EXPERIMENTS, RESULTS AND DISCUSSION

Using our self-fabricated continuum soft robot that was integrated with a mono-camera and FBG-fiber, the visual-strain fusion-based pose estimation method was employed to estimate pose in a LEGO® scenario. Pose estimation by ORB-SLAM2 was first proved to be reliable for fulfilling the purpose of model training. Next, several tests were carried out to prove the robustness of our method against poor quality of visual features.

香港大學圖書館 HKU Libraries

3.3.1 Soft Robot with Mono-camera and FBG

The continuum robot was molded by silicone rubber (Ecoflex30, Smooth-on Inc.), with a 3D printed tip cap and a fixation base as shown in **Fig. 3.5a**. Three pneumatic chambers are distributed in a distance of 5.1 mm to the robot's central axis and an angle of 120° between each other as illustrated in **Fig. 3.5b**, providing omni-directional bending [121, 122]. The chamber inflation was regulated by an actuation unit comprising of three pairs

of stepper motors and cylinders. Precise angular position control could be implemented on the motors, thus adjusting the volumes of sealed cylinders connected to chambers. **Fig. 3.5c** shows an endoscopic camera and an LED PCB module anchored on the robot tip cap. A single-core optical FBG fiber with 17 FBGs (6-mm long gratings, 20-mm spacing) was helically wrapped and adhered on the silicone continuum body. For the convenience of fabrication, the distances between adjacent turns of fiber were set at approximately 16.5 mm. The robot's outer diameter was 20 mm, and the bendable part was 90 mm in length. As the robot base was fixed in the experiment and the twisting is negligible, the roll orientation would not be controllable. Dr. Xiaomei Wang, the first author of the paper related to this chapter, was responsible for robot fabrication.

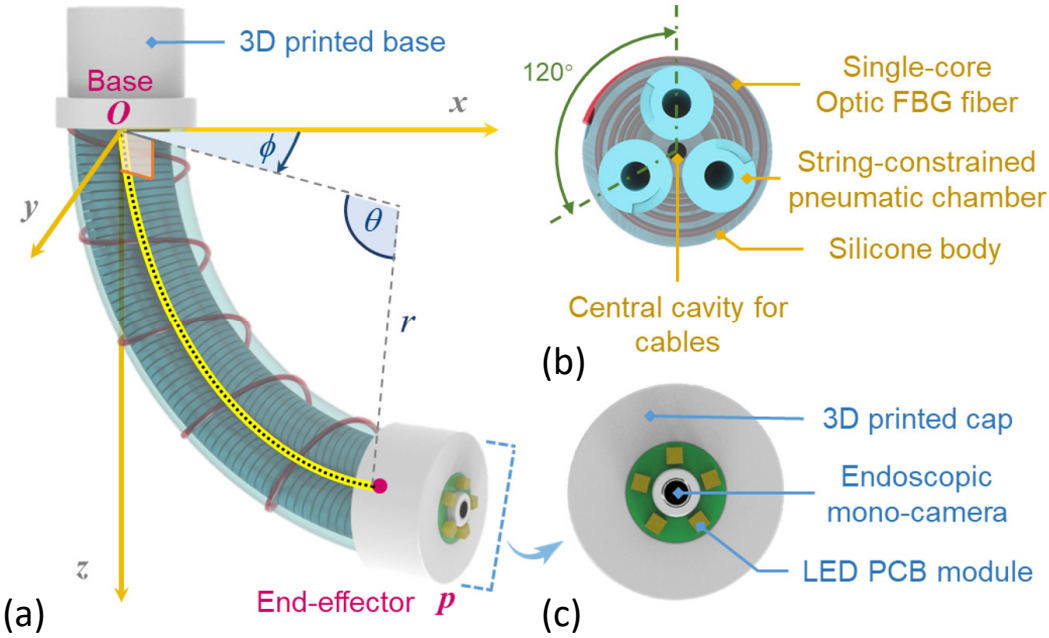

Fig. 3.5 Structural diagram of the continuum robot mounted with LEDs and a camera at its tip. **(a)** Configuration parameters r, θ and ϕ defined to describe a spatial arc of the constant curvature-based model. **(b)** Cross-section showing three air chambers for robot actuation. **(c)** Endoscopic camera providing real-time visual feedback to ORB-SLAM2 for camera pose estimation.

3.3.2 Pose Estimation by ORB-SLAM2

By constructing a LEGO® scenario, the accuracy of pose estimation by ORB-SLAM2 was first validated. The robot was actuated to follow a path that encircles and spirals out from the initial position. Actuation was achieved by assigning actuation steps $\boldsymbol{U}$ for three stepper motors. The actuation trajectory is shown in **Fig. 3.6b**.

Ground truth pose z was recorded by a pair of EM sensors attached on the robot tip. Camera calibration was performed to obtain intrinsic parameters at the beginning. Moreover, monocular metric scale was also determined in an initialization procedure. To perform initialization, the robot was moved slowly in a random direction until visual features were sufficient. When there was a *N*-step actuation $\boldsymbol{U}$, a set of SLAM-derived poses $\hat{\boldsymbol{Z}}_c = \begin{bmatrix} z_c(1) & z_c(2) & \cdots & z_c(N) \end{bmatrix}$ and a set of EM sensor measurements $\hat{\boldsymbol{Z}} = \begin{bmatrix} z(1) & z(2) & \cdots & z(N) \end{bmatrix}$ were recorded. By computing the affine transformation from $\boldsymbol{P}$ to $\hat{\boldsymbol{P}}_c$, SLAM-derived pose is calibrated to become $\boldsymbol{P}_c$:

$$\boldsymbol{p}_c = \mathbf{R}\hat{\boldsymbol{p}}_c \cdot \mathbf{k} + \hat{\boldsymbol{p}} \tag{3.19}$$

where $\mathbf{R}$ is a rotation matrix, $\mathbf{k}$ is a scaling factor, and $\hat{\boldsymbol{p}}$ is a translation vector. By comparing $\boldsymbol{P}_c$ and $\boldsymbol{P}$, mean absolute errors (MAE) of ORB-SLAM2 estimated poses in *x*, *y* and *z* directions were 0.508, 0.596 and 0.385 mm respectively. In total, the root-mean-square error (RMSE) was 0.998 mm as revealed in **Fig. 3.6a**. With the presence of sufficient visual features in our LEGO® setting, pose estimation by ORB-SLAM2 was proved to be reliable for fulfilling the purpose of model training.

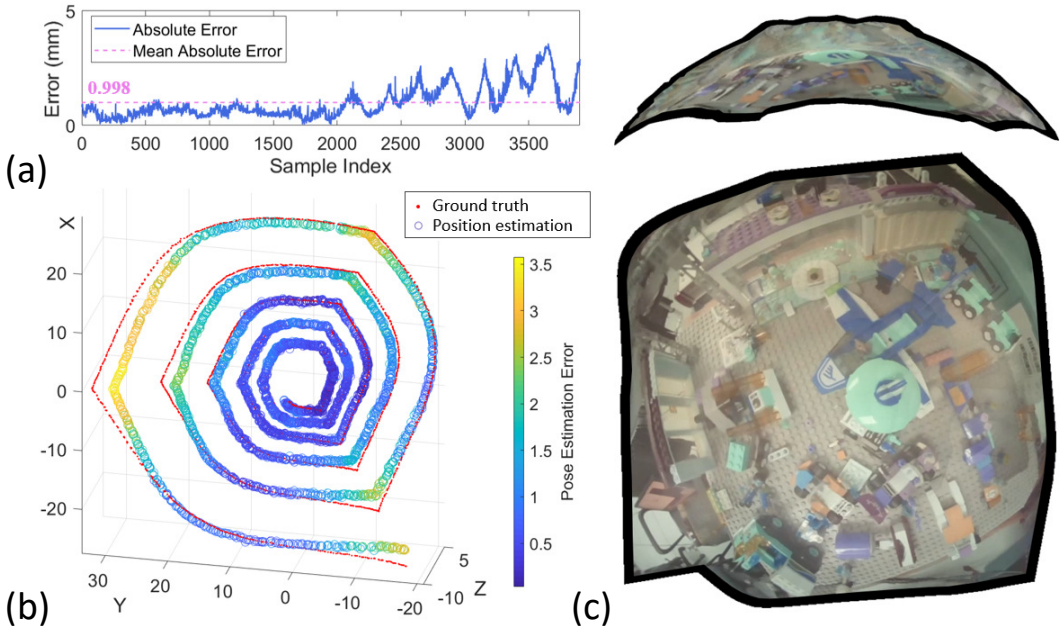

Fig. 3.6 Camera-based pose estimation results, where SLAM-based estimation was compared with ground truth measured by EM sensor. **(a)** Pose estimation errors. **(b)** Ground truth path and ORB-SLAM2 estimated path. **(c)** Front and side views of the stitched images in 3D, which are reconstructed using the SLAM pose estimation and image feedback.

香港大學圖書館 HKU Libraries

To showcase the reliability of ORB-SLAM2 tracking, an extra image mosaicking was performed as depicted in **Fig. 3.6c**. A 6D image stitching process was possible as both image frames and the corresponding estimated poses were available. Thanks to the robot actuation code developed by Dr. Xiaomei Wang and image-pose pair collection executed by Mr. Jing Dai, I was able to achieve 6D image stitching using Unity3D.

3.3.3 Sensor Fusion Pose Estimation

As mentioned in **Section 3.2.2**, ELM was adopted for the training and updating phases for our FBG pose estimation model. Wavelength shifts were utilized for end-effector pose estimation and model update simultaneously. In our model, the number of hidden nodes was 200 while number of initialization samples was 450. These numbers were determined by manual tuning based on the accuracy of ELM estimation. Time interval between steps was 50 ms.

Experimental results showed that with a larger sample number N_0, ELM prediction tends to improve. Though, improvement of accuracy became saturated when convergence of prediction was reached. At this point, significant improvement in accuracy would not be observed anymore. Prediction results by the ELM were then compared with camera-derived pose estimation, which was proven in **Section 3.3.2** to be reliable under the presence of sufficient visual features. Path estimated by the model was observed to approach closely to camera-derived poses. Numerically, MAE along the x, y and z directions were 1.82×10^{-4}, 3.95×10^{-4} and 4.39×10^{-4} mm respectively. In total, the mean spatial error was 8.28×10^{-4} mm. As a result, we have demonstrated the capability of an ELM to learn the mapping between FBG wavelength shift and pose information of the end-effector. After that, further experiments were carried out to test the effectiveness and robustness of our method when SLAM fails to estimate accurate poses due to adverse conditions. There were two tested conditions, including the presence of moving visual obstacles and a varying lighting condition.

香港大學圖書館 HKU Libraries

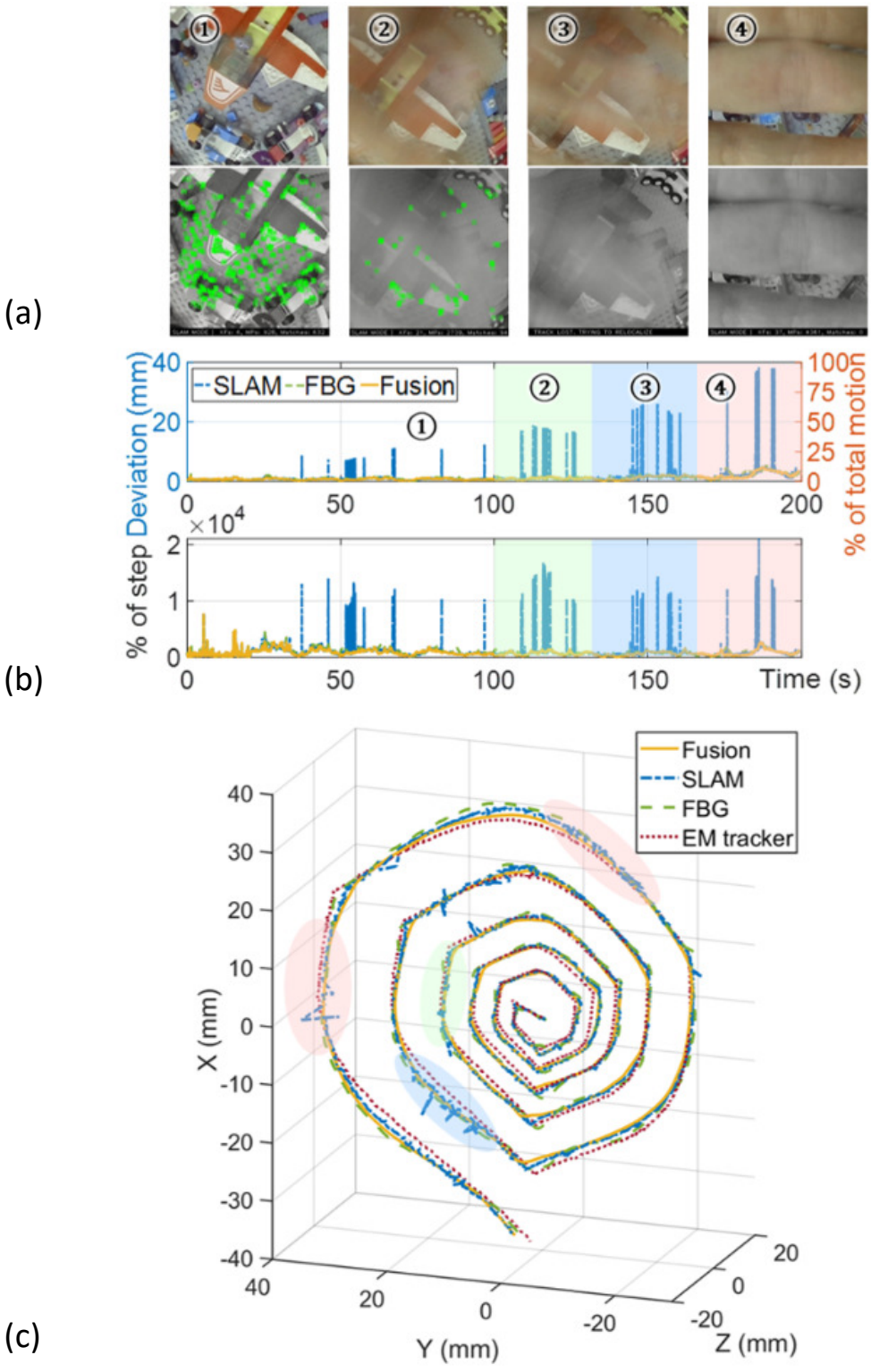

Fig. 3.7 Sensor fusion performance in the presence of visual obstructions. **(a)** Camera view and corresponding feature points in circumstances of: ① LEGO®-constructed scenario where feature points were in abundance; ② a hand that was moving in front of the camera where detected feature points drastically reduced; ③ a moving hand with no detected feature points; ④ a hand placed static in front of the camera that obscured all feature points for some seconds. **(b)** Deviations of fusion-based and SLAM-based pose estimation compared with EM sensors-measured ground truth poses. Percentages of error with respect to total motion range and each-step motion are provided. **(c)** Four paths depicting fusion-, SLAM- , FBG-based camera positions and EM-based ground truth path.

香港大學圖書館 HKU Libraries

3.3.3.1 In the Presence of Moving Obstacles

Actuation of the robot was again in a spiral path. As illustrated in **Fig. 3.7a**, moving and static obstacles were put in front of the camera view as a disturbance when the robot was actuated. It was also possible to reduce the number of visual features by artificially altering SLAM settings. However, this method may only adjust the density of features on the camera view. In contrast, using physical obstacles could directly create dynamic change in feature loss throughout the whole tracking process. Both the speed of obstacle motion and area of obstruction on the camera view can also be adjusted. In the first 100 s, the camera view was free of disturbance as shown in ①. To achieve ideal tracking as for a testing scenario, around 600 features in each frame were guaranteed. MAE of SLAM-derived poses and fusion-derived poses were 0.840 mm and 0.768 mm respectively. Next, a hand was put statically in front of the camera for around 30 s as shown in ②, followed by quick movement of the hand for another 30 s as shown in ③. In both ② and ③, visual features were not consistently covered. At the same time, number of features might even reduce to zero in a few image frames. MAE for SLAM was 1.694 mm, which was 4.2% of the largest distance away from the initial robot position. The maximum error was 26.272 mm. For fusion-based tracking, MAE was 1.132 mm (2.8% of the largest distance) and maximum error was 2.573 mm (6.4% of the largest distance). For the last 30 s of actuation, the hand statically covered the entire camera view for every other 2 s as shown in ④. In this situation, loss of visual features was consistent. SLAM estimation was severely affected, giving a maximum error of 38.483 mm, while fusion-based estimation gave a comparatively higher accuracy at a maximum error of 4.747 mm. It was notable that during ③, SLAM estimation might even be paused when all visual features were blurred. However, our proposed fusion-based tracking was still functioning as feedback from FBG could compensate for the loss in visual feedback. To cope with feedback loss from SLAM, the SLAM pose was set to be the latest valid value to avoid discontinuation of data streaming. As illustrated in **Fig. 3.7b** and **Fig. 3.7c**, pure SLAM-derived pose deteriorated in adverse conditions while fusion-based pose experienced a minimal amount of effect.

Judging from **Fig. 3.7c**, we can see that using FBG could provide a consistently stable pose estimation outcome. Apparently, fusion with SLAM seems not necessary. However, it is notable that FBG may not be reliable in every scenario. For instance, FBGs are thermally sensitive, its accuracy of strain measurement would be affected by local temperature changes caused by ablation or coagulation. Secondly, FBGs may only detect the change of robot configuration itself. Robot base motion would not affect robot configuration. In this

香港大學圖書館 HKU Libraries

case, FBG would not give any hint of pose feedback. The synergetic use of FBGs and SLAM is still necessary and practical. In addition, applying filters such as a median filter or a low-pass filter on the SLAM-derived poses seems to be an alternative to using FBGs. Apparently, most outliers due to SLAM tracking lost could be eliminated by these common filters. However, it inevitably introduces time delays; in contrast, the proposed visual-strain fusion method does not. Furthermore, such simple filters could be ineffective particularly when SLAM tracking lost lasts for several seconds or more.

3.3.3.2 In the Presence of Varying Lighting

Same as the actuation sequence in **Section 3.3.3.1**, robot moved in a spiral path. In this experiment, varying lighting conditions were produced as a disturbance during robot movement. The robot started moving under normal lighting provided by incandescent lamp as shown in **Fig. 3.8a**①.

Then, light was gradually dimmed to complete darkness as shown in ②, followed by the return to initial lighting. Complete darkness occurred in the period 75 ~ 150s. Next, in the period 150 ~ 200 s, a white-colored LED was directed towards the camera as shown in ③. During a low-light circumstance, visual features in the image frame reduced, subsequently inducing noise in poses estimated from SLAM as revealed in **Fig. 3.8b**. When light was removed, visual features were absent as the whole image frame was darkened, leading to interruption of SLAM estimation. On the other hand, a moving LED also brought noise to SLAM. Visual features would also be lost as brightness became saturated in some regions on an image frame, resulting in an increased SLAM tracking error. Numerically, SLAM exhibited an accuracy of RMSE at 3.116 mm, while our proposed fusion-based estimation presented an accuracy of RMSE at 1.324 mm. Results showed that fusion-based estimation achieved a stable estimation under varying lighting conditions and brought about a significantly improved accuracy compared with pure SLAM estimation.

香港大學圖書館 HKU Libraries

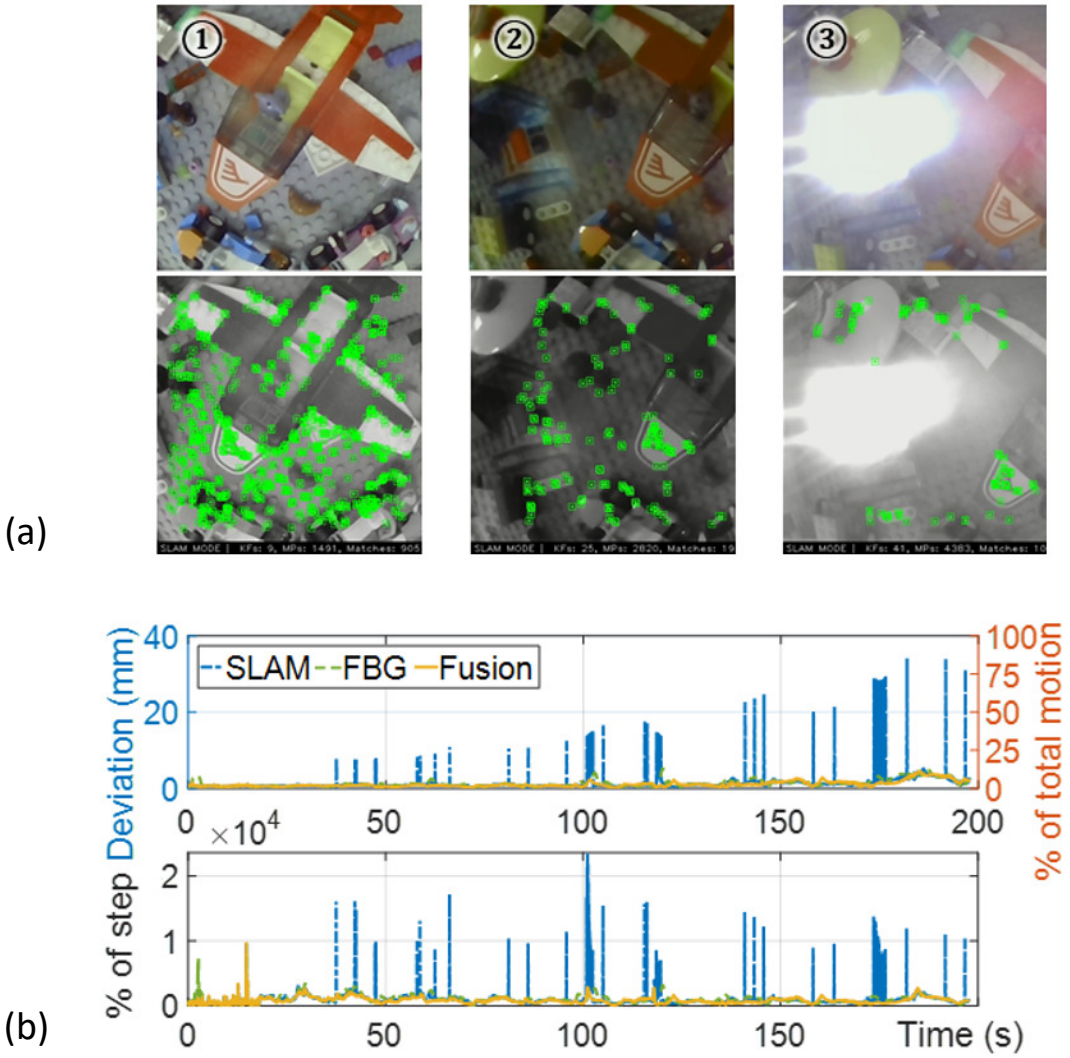

Fig. 3.8 Sensor fusion performance in different lighting conditions. **(a)** Camera views and corresponding feature points under ① usual lighting in laboratory, ② low intensity lighting, and ③ moving portable LED. **(b)** Deviations of fusion-, FBG- and SLAM-based pose estimation compared with EM tracking ground truth pose. Percentages of error with respect to total motion range and each-step motion are provided.

Lastly, scene reconstructions in form of 6D image stitching, as depicted in **Fig. 3.9a** and **Fig. 3.9b,** were also conducted for adverse conditions described in both **Section 3.3.3.1** and **3.3.3.2**. From fusion-based estimation, poses were obtained for 6D image stitching. On the whole, smoothly stitched image mosaics were acquired, owing to the stable and consistent pose estimation by our proposed fusion-based estimation method. Blurred regions marked with a dotted line on these mosaics were caused by obstacles and varying lights.

香港大學圖書館 HKU Libraries

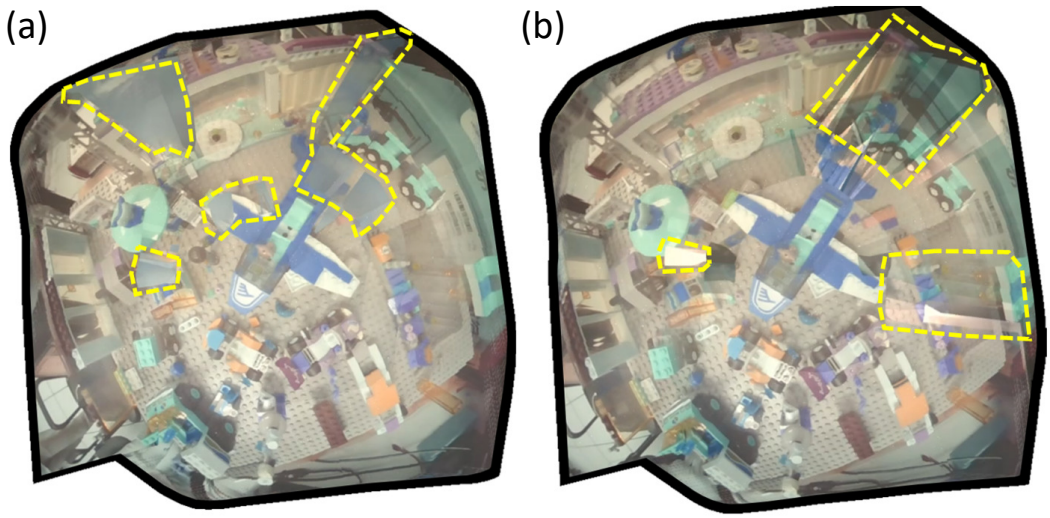

Fig. 3.9 Scenario reconstructions with disturbances of **(a)** moving obstacles and **(b)** varying lighting conditions. Several blurs due to moving obstacles or varying lighting are indicated by dotted outlines.

3.4 CONCLUSION AND FUTURE WORK

This study proposes an integrated soft robot control system with visual-strain fusion-based pose sensing. Prior data collection was not required as the data-driven model was trained online and could conduct prediction simultaneously. A single-core FBG fiber was wrapped on the surface of the robot to collect sparse strain measurements. At the same time, a monocular camera mounted at the end-effector estimated poses with a SLAM algorithm. SLAM estimated poses were then used to train an ELM model that maps FBG strain readings to poses. Next, FBG- and SLAM-derived poses were fused to give a reliable and robust pose feedback of the robot end-effector, enabling a smooth 6D image stitching even under adverse visual conditions such as the presence of obstacles and exaggerated lighting. The proposed method was proved immune to failures in SLAM that were due to poor quality of visual features. Mean estimation error was reduced from an RMSE of 3.116 mm to 1.324 mm when fusion was used instead of pure estimation by SLAM.

The proposed visual-strain fusion sensing modality could be extended to other robot designs, including multi-segment prototypes, although in this study we only validated it on a single-segment continuum robot. The application of single-core FBG fiber is not limited by the number of segments, as long as the adjacent segments are connected by a continuous joint that is smooth for wrapping the fiber. Notably, our learning-based FBG model

香港大學圖書館 HKU Libraries

incorporates sparse FBGs to predict robot pose based on its configuration, which has been proven capable of adapting to common local contacts. The more advanced multi-core fiber using OFDR technique can even eliminate the local/global interaction effect on the similar pose/configuration estimation, such as impulsive or continuous interaction-induced deformation. Examples are found in bronchoscopy (e.g., Ion endoluminal system, Intuitive Surgical, Inc.) and catheterization platforms [108, 109, 123]. Finally, it is possible to generate a new SLAM architecture by directly coupling FBG feedbacks with typical SLAM processes, such as estimator initialization, online loop detection, and re-localization. While our current approach is considered as "late-fusion", such kind of visual-FBG SLAM system is an "early-fusion", which is more sophisticated as FBG feedbacks are directly fed into the SLAM architecture. This preliminary work intends to provide a proof-of-concept of mapping FBG strain measurement to pose information by using a neural network, as well as how FBG feedback may assist SLAM in adverse visual conditions. Therefore, we employed a simpler approach in this work. Nonetheless, the "early-fusion" approach of visual-FBG SLAM remains to be one of our future research directions.

香港大學圖書館 HKU Libraries

CHAPTER 4
REAL-TO-VIRTUAL DOMAIN TRANSFER-BASED DEPTH ESTIMATION

4.1 INTRODUCTION AND RELATED WORK

Any surgical procedure involves the collaboration between different personnel like surgeons and nurses. Effective communication is paramount to ensuring a smooth surgical workflow. In particular, communication can be achieved by graphical annotations drawn on a display device when the use of an endoscope is involved. In this manner, any target structures inside the field of view can be annotated with information and instantly shared. Advantages brought by surgical annotation are not limited to within an operating theater. As it enables real-time graphical communication, beneficiaries include everyone involved in the procedure such as teachers, students and medical trainees. Examples of annotation include a multi-institutional cooperation during adrenalectomy through video conferencing [124], experimental illustrations of intention sharing by visualizing eye gazes of separated collaborators [125-129], and Da Vinci™ telestration that enables surgeons to draw sketches on live video streams [130]. These examples involved graphical annotations to facilitate effective communication. Nevertheless, they did not showcase the capability of allowing the annotations to keep their positions with respect to the anatomy upon camera movement.

香港大學圖書館 HKU Libraries

In this work, we aim to achieve 3D annotations on the endoscopic view. Annotations shall anchor to the target surface accurately and stably even during camera movement. We will achieve 3D annotation through monocular depth estimation, as well as endoscope pose tracking using an EM sensor. Not only should annotations anchor to the surgical scene during camera movement, accurate size change with respect to the endoscope's distance from the annotated target should also provide viewers with improved depth perception. To elaborate, achieving 3D annotation is essentially implementing AR. The 3D annotation is instantiated in a virtual 3D world and later registered to the real-world surgical field. By augmenting the exposed surgical view with intra- or pre-operatively obtained images or 3D models [131], AR applied in surgeries allows overlay of subsurface critical structures and pre-operatively planned trajectories that include depth information. Subsequently, it may reduce the risk of complications, increase surgical efficiency and aid with surgical training [21].

As a proof of concept, we selected nasal surgery for the implementation of 3D annotations. AR systems are the most useful when the target surgical sites have little deformation and movement [2], making the nasal cavity and paranasal sinuses a suitable candidate for AR implementation. Additionally, due to its proximity to the brain, many critical structures can be overlaid in the endoscopic view. To achieve AR in nasal surgeries, researchers and companies tend towards sensor-based approaches utilizing external equipment. The endoscope and the target anatomy are usually tracked by medical grade optical or EM sensors. The Scopis® Hybrid Navigation system (Stryker, USA) is a commercial example that combines optical and EM sensing to achieve AR. Pre-operative 3D models are usually obtained from CT scans, which are then registered to the sensor-based tracking system reference frame by rigid registration, enabling overlay of pre-operatively obtained models onto the real anatomy in the surgical scene.

Provided that we adopt the above approach to achieve 3D annotation, depth information observed by the camera would be based on the registered pre-operative 3D model. However, observed depth in this context may not be representative of the real surface during surgery, especially in the nasal cavity where soft mucosal linings are not clearly observable in CT scans. It is also notable that the quality of a 3D reconstruction from CT scans is highly dependent on scanning quality, reconstruction software, and human operation [2].

One of the possible alternatives to obtain depth is to resort to traditional vision-based approaches such as stereo or monocular visual simultaneous localization and mapping (vSLAM). vSLAM outputs camera trajectory and a 3D structure of an environment without

香港大學圖書館 HKU Libraries

any prior knowledge or the use of any active sensors. Using vSLAM, visual input can be taken advantage of to perform tracking and mapping. Depth can be obtained from vision in real-time, which may be more representative than depth based on a registered pre-operative model. Real-time stereo reconstruction has been performed previously for laparoscopic surgery [132], however, stereo vision is difficult to implement in nasal surgeries due to constraints on the endoscope size. Depth estimation through monocular endoscopes has also been demonstrated, for example by tracking and matching video frame feature points for both endoscope tracking and point cloud reconstruction in the nasal airway of a cadaver head [133], and through ORBSLAM [134] approaches for tracking laparoscope pose and mapping the surgical scene [135]. However, feature based tracking is prone to failure inside the nasal cavity owing to the lack of texture and the apparent repetition of patterns [136].

In view of the rapid development in deep learning-based monocular depth estimation, there lies a great opportunity in surgical AR to exploit vision-based depth. Novel examples giving promising estimation results include DORN [137] and DenseDepth [138]. Supervised learning methods such as DORN require an endoscopic image dataset with ground truth depth labels for training. During the application phase of the trained neural network, the estimated depth output is generated from color image input. Unfortunately, there does not exist a large, readily available labelled dataset for the nasal airway. It is also impractical to collect ground truth depth data inside the nasal airway using active sensors. To address this limitation, some researchers [139, 140] have attempted to train their depth estimation networks in a self-supervised manner such that no depth labelling is required prior to network training. Nonetheless, both employed a structure from motion (SfM) algorithm to obtain sparse depth before the training phase. Consequently, depth estimation output highly depends on the accuracy and quality of SfM output. Implicit domain adaptation that translates synthetic colon endoscopic images to depth maps by using pix2pix [141] has also recently been proposed. Other than paired simulated data, unlabeled real colon images were also involved in the training phase such that the trained model may produce more accurate depth predictions in patient data [142].

Adopting a similar approach used in prior art [143], we train a supervised depth estimation network in a virtual environment and utilize it to predict depth of real endoscopic images. Prior to depth estimation, a real-to-virtual image style transfer using Cycle Generative Adversarial Network (cycleGAN) [144] is performed. With adversarial learning, domain adaptation between the real domain and the synthetic domain is accomplished. Previously, cycleGAN-like architectures have been used to adapt real bronchoscopy images to virtual style images [145]. Real-to-virtual adaptation was also used for colonoscopy using a

香港大學圖書館 HKU Libraries

Generative Adversarial Network (GAN) architecture [146]. Through this approach, preparation of an unlimited amount of absolute ground truth depth becomes possible while depth prediction can be implemented on real-to-virtual-adapted real endoscopic images. Time and labor costs for data preparation through this approach would be minimal. In addition, real-to-virtual domain adaptation can remove patient-specific texture details that may vary widely between patients, potentially making the depth estimation network generalizable across patients [146].

Apart from aiming at generating depth that is more representative of a surface, we are also concerned with its stability. Therefore, a brief stability evaluation of our proposed system is included towards to end of this study. The major contributions of this study are listed below:

I. The application of monocular depth estimation is extended beyond offline 3D reconstruction of surgical scenes, into applications with real-time AR for surgical guidance.

II. A supervised depth estimation network is trained entirely in a virtual environment and used to predict depth from endoscopic images in real-time by implementing cycleGAN-based real-to-virtual style transfer.

III. Predicted depth is quantitatively evaluated against ground truth depth in a nasal airway phantom. Accuracy of augmented 3D annotations is evaluated while overall system stability is quantitatively assessed.

4.2 IMAGE DEPTH ESTIMATION AND 3D ANNOTATION

To perform 3D annotation, we propose to train a self-supervised monocular depth estimation network in a virtual environment. During application phases, an auxiliary real-to-virtual image style transfer network was adopted to first transform real endoscopic images to synthetic-like images. Framewise depth maps were then predicted by the monocular depth estimation network. By obtaining depth values of specific pixels and EM sensor-tracked camera poses, 3D annotation was ultimately achieved.

香港大學圖書館 HKU Libraries

4.2.1 Data Preparation for Deep Neural Network (DNN) Training

The goal of our method is to train a supervised depth estimation network in a virtual environment and utilize it to predict real endoscopic image depth. In doing so, synthetic endoscopic images and the corresponding ground truth depth maps were generated in a virtual world space using Unity3D. An overview of the data preparation process is illustrated in **Fig. 4.1**. A virtual camera was set up with intrinsic parameters obtained from camera calibration of an Olympus rhinolaryngoscope (ENF-VH). Not only did we match intrinsic parameters of the real and virtual endoscopes, but we also attached two point light sources near the virtual camera that exhibit realistic inverse square intensity fall-off in relation to distance from the source. Next, an anatomically accurate nasal airway model was imported into the virtual environment. Taking reference from Yang *et al.* [147] and Fan *et al.* [148] who fabricated patient-specific surgical plate and left atrium respectively from patient scans, we made our nasal airway model by obtaining CT scans of a cadaver head, followed by 3D segmentation and model editing. The model surface was assigned a uniform light-red color to emulate the nasal mucosal lining, however, noting that patient-specific textures such as vascular patterns were absent. Depth estimation network trained in this manner is hypothesized to have improved generalizability across patients.

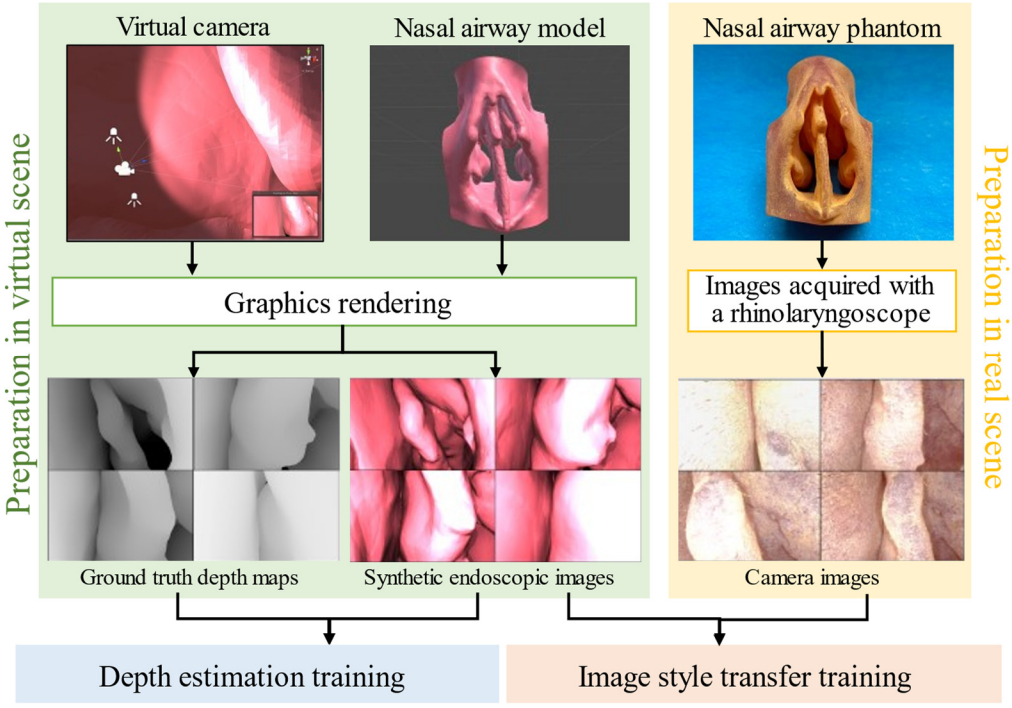

Fig. 4.1 Preparation of i) ground truth depth maps, ii) synthetic endoscopic images and iii) real endoscopic images for depth estimation and image style transfer training.

A dataset of 3,600 synthetic endoscopic images and the corresponding ground truth depth maps were captured from the virtual camera while the camera was moved inside the virtual nasal airway model along a pre-defined pathway. Ground truth depth maps were stored as greyscale images. Depth observed by the virtual camera was set to span a range of 0.01-25 mm. As we adopted a GAN-based unpaired image-to-image style transfer network, real endoscopic images and synthetic endoscopic images prepared do not necessarily correspond with one another. Therefore, 3,000 real endoscopic images were directly captured inside a 3D printed nasal airway phantom. The phantom was based on a segmented anatomically accurate nasal airway model and 3D-printed in a material with Shore hardness value of 70.

4.2.2 Real-to-virtual Image Style Transfer

Our aim for real-to-virtual image style transfer is to learn a mapping $G : X \rightarrow Y$, where the domain variance between real RGB image $x \in X$ and synthetic-style RGB image $y \in Y$ is bridged. As a result, a depth estimation network trained on synthetic endoscopic images can be deployed on real RGB endoscopic images. To obtain the mapping model, a GAN-based unpaired image-to-image translation method called CycleGAN [144] is applied. It consists of two translators G, F to learn the mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$. Two adversarial discriminators D_x and D_Y are trained to differentiate the style-transferred images from the domain images. The goal of the discriminator D_x is to distinguish $\{x\}$ and $\{F(y)\}$ which is style-transferred to style of X by translator F, and vice versa for D_Y and G. D_x and D_Y are both PatchGAN [141] classifiers. Transformer G is thus encouraged to translate X into outputs indistinguishable from domain Y. In other words, G is enforced by D_Y to produce synthetic-like images from real RGB images. The loss function combines adversarial loss [149] and cycle consistent loss [144].

香港大學圖書館 HKU Libraries

4.2.3 Image Depth Estimation

The architecture of our depth estimation model is an encoder-decoder network, which is the DenseDepth as presented in [138]. The encoder part is a pretrained DenseNet-161 network for extracting features from our RGB images and representing them as a feature

map. The decoder part contains blocks of convolutional and up-sampling layers to transform the feature map into the desired depth output. The same spatial shape of the encoder layers is skip-connected into the decoder to improve the prediction performance and produce sharper depth estimations. The loss function to train our network is a combination of depth loss and structural similarity (SSIM) loss. The model aims to learn the mapping between the synthetic endoscopic image dataset and the corresponding ground truth depth value such that it can predict the depth value in real endoscopic images.

4.2.4 3D Annotation

After the training phase of both depth estimation and image style transfer networks, system integration was performed to achieve 3D annotation using Unity3D as an interface for visualization. First, a 6-DoF EM sensor (Aurora, NDI, Canada) was attached to endoscope tip, as illustrated in **Fig. 2.26**. Using Tsai's method [64], hand-eye calibration was employed to find the transformation ${}^{c}\mathbf{T}_{s}$, a description of the sensor frame relative to the camera frame. To reduce error propagation in the case of imperfect hand-eye calibration, the EM sensor was attached to the endoscope tip at approximately 2 mm from the camera's optical axis. By directly streaming sensor pose in EM frame ${}^{em}\mathbf{T}_{s}$ into Unity3D, pose of the camera tip with respect to virtual world ${}^{w}\mathbf{T}_{c}$ was assigned as the virtual camera pose:

$$ {}^{w}\mathbf{T}_{c} = {}^{em}\mathbf{T}_{s} \cdot {}^{c}\mathbf{T}_{s}^{-1} \tag{4.1} $$

Next, RGB video frames were streamed from the endoscope during observation of the 3D printed nasal airway phantom. The phantom was static relative to the EM tracking field. Before passing video frames to i) Unity3D for visualization and ii) real-to-virtual image style transfer network, image un-distortion was applied as a data pre-processing step. Undistorted frames passed to (i) may then accurately be overlaid with virtual objects, which would be observed by a virtual camera that was distortion-free by default.

Style transferred image frames processed in (ii) were further relayed to the depth estimation network for generating framewise depth maps. Depth at each pixel was stored as a normalized float number in a range between 0 (far) and 1 (near), which was then converted into a depth range of 0.01-25mm in the virtual environment, matching the depth range of the image set used for depth estimation training. **Fig. 4.2** summarizes the procedures involved in estimating framewise depth maps.

香港大學圖書館 HKU Libraries

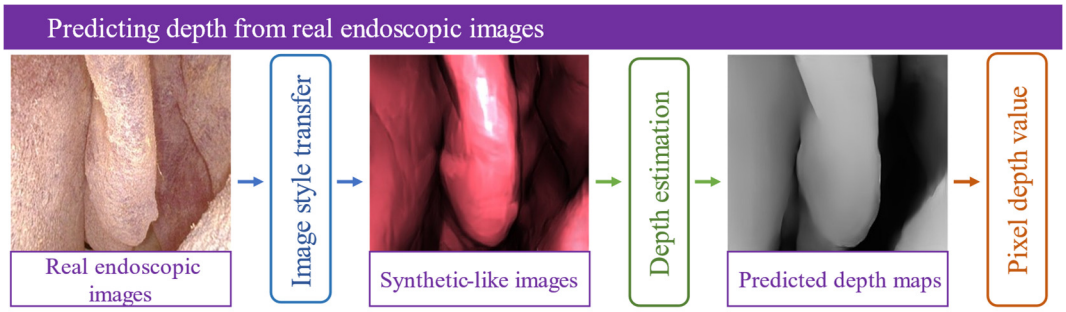

Fig. 4.2 Application of image style transfer and depth estimation networks for obtaining real-time framewise depth estimation from real endoscopic image inputs.

While RGB frames were displayed in the Unity3D view in real-time, a pixel (u, v) was selected by the cursor to begin annotation. Given the camera intrinsic matrix $\mathbf{K}$ and depth value d at (u, v) retrieved from a predicted depth map, an annotation element in the form of a simple spherical object with a position $\mathbf{p}_c = [x_c \quad y_c \quad z_c]^{\mathrm{T}}$ in camera coordinates was placed in virtual game world, where:

$$\mathbf{p}_c = \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \mathbf{K}^{-1} \cdot d \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{4.2}$$

which was further expressed as $\mathbf{p}_w$ in virtual world coordinates:

$$\begin{bmatrix} \mathbf{p}_w \\ 1 \end{bmatrix} = \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = {}^{w}\mathbf{T}_c \cdot \begin{bmatrix} \mathbf{p}_c \\ 1 \end{bmatrix} = {}^{em}\mathbf{T}_s \cdot {}^{c}\mathbf{T}_s^{-1} \cdot \begin{bmatrix} \mathbf{p}_c \\ 1 \end{bmatrix} \tag{4.3}$$

In summary, instantiating a 3D annotation requires i) predicted depth d at (u, v), ii) camera intrinsic matrix $\mathbf{K}$ and iii) camera pose from EM sensor.

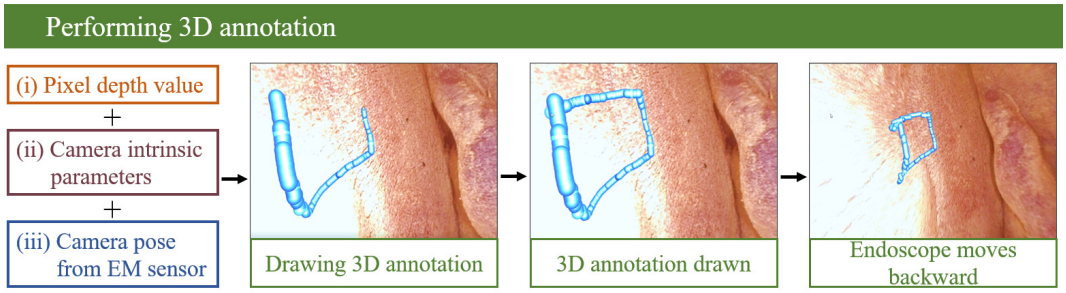

Fig. 4.3 Based on i) predicted depth, ii) camera intrinsic parameters and iii) camera pose from EM sensor, the annotations can be anchored onto the anatomical surface in a stable manner.

香港大學圖書館 HKU Libraries

4.3 ASSESSING DEPTH ACCURACY AND ANNOTATION STABILITY

Before the training phase of the image style transfer network, both the real endoscopic images and the synthetic endoscopic images were resized to 288×256 pixels. The translators consisted of 2 convolution layers with stride of 0.5, 9 residual blocks [150] and another convolution layer that outputs a feature map. For the discriminators, 70×70 PatchGANs [141] were employed. The entire network was trained for 200 epochs with a batch size of 1. Adam [151] optimizer with initial learning rate of 2×10^{-4} was applied. Weight parameter λ in [144] was set to be 10.

The depth estimation network was first trained with NYU Depth v2 [152] dataset as a pretraining step to obtain optimal layer weights for depth estimation. The network was trained with Adam [151] optimizer, initial learning rate 1×10^{-4} and batch size of 4 for 20 epochs. To train the network for our purpose, synthetic endoscopic images and the corresponding depth maps were used as subsequent fine-tune training with 50 epochs. Images were resized to 640×480 pixels prior to fine-tune training. Weight parameter λ in [138] was set to be 0.1. The proposed framework was implemented on a computer with an AMD Ryzen Threadripper 3960X CPU, 64GB RAM and two NVIDIA GTX 1080Ti GPU.

4.3.1 Endoscopic Image Dataset Preparation

To evaluate depth estimation accuracy, a testing dataset consisting of 2,400 RGB image frames captured by the endoscope during observation of the 3D printed nasal airway phantom was prepared. Simultaneously, corresponding ground truth depth maps were generated in the following manner:

I. By equation (4.1), the endoscope pose in the EM frame was assigned as the virtual camera pose in real-time.
II. Having the anatomically accurate nasal airway model, six non-co-planar anatomical positions $\{x_i\}$ on the model were recorded in model frame.

香港大學圖書館 HKU Libraries

III. Using a 6-DoF EM probe, the corresponding six anatomical positions $\{y_i\}$ on the 3D printed nasal airway phantom statically placed in EM tracking field were recorded.

IV. To find the transformation matrix ${}^{em}\mathbf{T}_m$ that registers the nasal airway model to the 3D printed phantom, where:

$$ {}^{em}\mathbf{T}_m = \begin{pmatrix} & \mathbf{R} & & \mathbf{t} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.4} $$

for $\mathbf{R}$ being the rotational matrix and $\mathbf{t}$ being the translational vector, $\mathbf{R}$ and $\mathbf{t}$ were solved by minimizing the following least-squares error:

$$ \sum_{i=1}^{N} || \mathbf{R}x_i + \mathbf{t} - y_i ||^2 \tag{4.5} $$

which in our case $N = 6$, solved using singular value decomposition (SVD) method proposed in [153].

V. When importing the nasal airway model into the virtual scene, ${}^{em}\mathbf{T}_m$ was applied to it. As both the phantom and endoscope were registered to the virtual world, ground truth depth maps could be collected while real RGB frames were being captured.

4.3.2 Annotation Stability Evaluation

Stability of an AR system can be described by the synchronization of movements by the virtual object and real object on a display. Two of the main factors influencing synchronization are i) discrepancies in sensors' latency and ii) intrinsic noise from the sensors. While intrinsic sensor noise is a limitation that could not be directly addressed, we attempted to observe and record data streaming latency from the EM sensor and the endoscope. In particular, depth estimation relies on endoscopic image streaming. Latency in obtaining depth is directly affected by latency in endoscopic image streaming. In this test, time spent in different segments of i) endoscopic image stream and ii) EM sensor data stream was approximated.

The second way to describe stability is consistency in the depth predicted. Prediction made by supervised monocular depth estimation often flickers due to independent per-frame processing [154]. In our system, depth inconsistency might seem irrelevant as depth is only

香港大學圖書館 HKU Libraries

assigned to a sphere annotation when the cursor is clicked on a pixel at a particular instance. However, consistency between frames is in fact still relevant when depth is continuously read and sphere annotations are consecutively made as the pressed cursor is dragged.

Through the registration method described in **Section 4.3.1**, the nasal airway model was registered to the phantom statically placed in the EM tracking volume, as shown in **Fig. 4.4a**. To evaluate stability in terms of depth consistency, a virtual sphere was placed on the airway wall at a point with known location. The endoscope was then directed at this sphere and moved in a forward-backward direction such that depth of the sphere with respect to the camera varies with time, as illustrated in **Fig. 4.4b-d**. This depth value can be directly obtained in the virtual world as this is the distance between virtual camera and virtual sphere, which we define as the "reference depth". Simultaneously, pixel coordinates of this sphere appearing in the Unity3D viewport were continuously captured and relayed to the depth estimation network. The corresponding predicted depth was obtained, which we define as the "predicted depth". The reference depth and predicted depth were captured and plotted against time. In total, five trials were conducted, each lasting for 30-60 seconds. The endoscope moving speed was maintained at around 3 mm/s and the data sampling rate was 50 Hz.

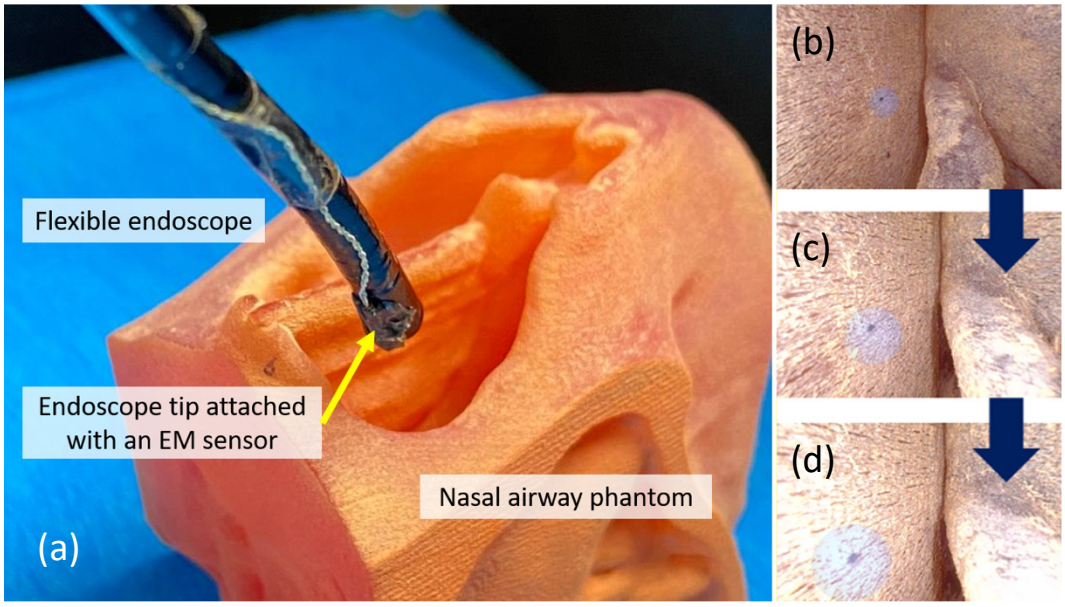

Fig. 4.4 (a) Depth consistency evaluation using a 3D-printed nasal airway phantom. Flexible rhinolaryngoscope (ENF-VH, Olympus) with a 6-DoF EM sensor (Aurora, NDI, Canada) attached at the tip was used in this experiment. **(b-d)** Manually inserting the endoscope tip into the nasal airway, a blue virtual sphere could be observed on the endoscopic view. Endoscope was moved in a forward-backward direction, with a moving speed maintained at around 3 mm/s and data sampling rate at 50 Hz.

香港大學圖書館 HKU Libraries

4.4 RESULTS AND DISCUSSION

Using depth values predicted by the proposed method, 3D annotation was implemented on the endoscopic view during an examination inside the nasal airway phantom. Next, system stability was quantitatively evaluated in two aspects, namely: i) latency discrepancy between different data streams, and ii) depth consistency. Both accuracy and stability evaluations demonstrated the feasibility of performing monocular depth estimation via style-transfer on endoscopic images, which enabled 3D annotation implementation. A supplementary video is available online [155], which illustrates the training data collection process, qualitative results of predicted depth, and 3D annotation implementation.

4.4.1 Depth Estimation Accuracy

Qualitative comparison between predicted depth maps and corresponding ground truth depth maps collected with the method described in **Section 4.3.1** is shown in **Fig. 4.5**. To quantify our depth estimation accuracy by comparing ground truth depth and predicted depth, normalized root-mean-square error (NRMSE):

$$\sqrt{\frac{\sum_i (x_i - y_i)^2}{n}} \left(x_{\max} - x_{\min} \right)^{-1} \tag{4.6}$$

and SSIM proposed in [156] were calculated. Similarity is indicated by a NRMSE close to 0 and a SSIM close to 1, where SSIM spans between -1 and 1. Depth prediction accuracy is shown in **Table 4.1**, which is juxtaposed to quoted prediction results of two existing methods that i) adopted a similar depth estimation workflow for colonoscopy [146] and ii) used dictionary learning (DiL) trained on CT colonoscopy images [157].

Table 4.1 Depth prediction result comparison with dictionary learning (DiL) [157] and unsupervised reverse domain adaptation [146].

	NRMSE	SSIM
DiL [157]	0.50	0.32
Mahmood *et al.* [146]	0.23	0.77
Our method	0.3224 ± 0.0773	0.8310 ± 0.0655

香港大學圖書館 HKU Libraries

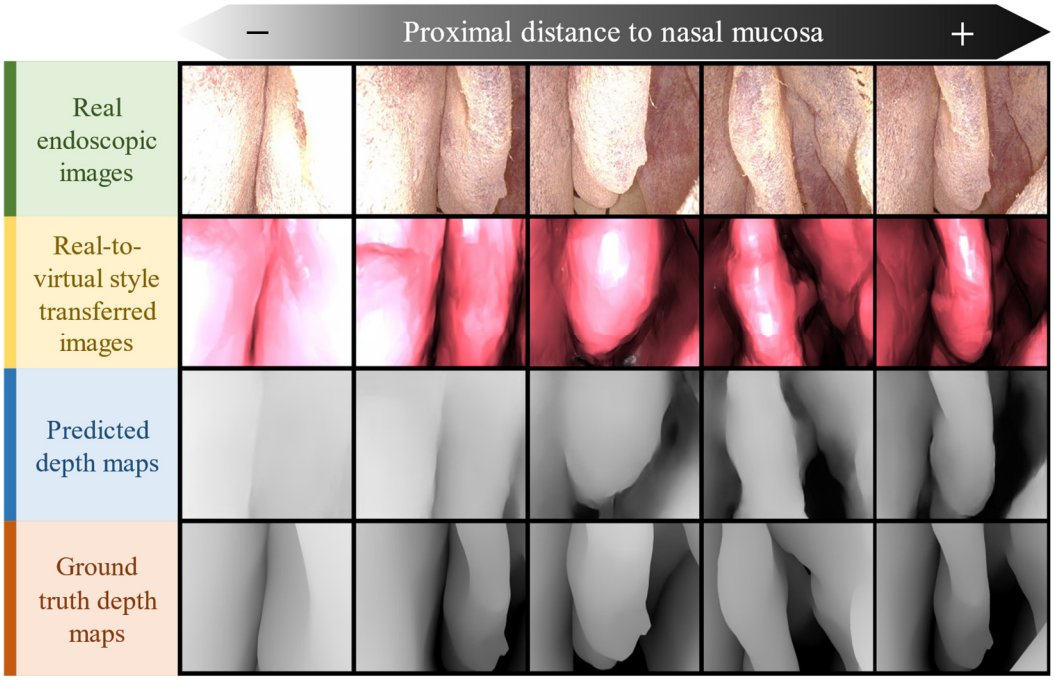

Fig. 4.5 Qualitative results of predicted depth (3rd row) in comparison with ground truth depth (4th row). Real endoscopic images (1st row) were first style-transferred to synthetic-like images (2nd row) before depth prediction by the supervised depth estimation network.

As compared to state-of-the-art shown in **Table 4.1**, Mahmood *et al.* [146] and our proposed method was significantly more accurate than the DiL implementation [157]. This could be attributed to the fact that their work did not incorporate a virtual camera model with point light sources exhibiting realistic inverse square intensity fall-off, which is believed to be a crucial element to consider in depth estimation. Despite achieving the best estimation accuracy in terms of SSIM, **Fig. 4.5** illustrates that our method will produce poorer depth estimation results when the endoscope is closer to the nasal mucosa. This is in part likely due to light intensity saturation when moving the endoscope closer to a surface, where edges and depth information of narrow passages tend to be lost, yielding an average depth biased towards a high proximity value. In real application of endoscopy, saturated lighting is usually avoided by adopting an auto-adjustment mode for light intensity. Unfortunately, as our model was trained with images in which the light intensity was fixed at a certain level, a fixed light intensity also had to be adopted during the application phase of the depth estimation model. Therefore, future work may include a thorough photometric calibration to characterize the lighting properties of an endoscope, such that our proposed method can be applied with auto-adjusted light intensity and auto-exposure.

香港大學圖書館 HKU Libraries

Another limitation is related to scale ambiguity. It is a common problem in every monocular pose and depth estimation solution because monocular cameras do not provide the details of scale between the physical objects and their images. The immediate depth output from the neural network is dimensionless, which is expressed in a normalized value in the range between 0 and 1. In our experimental setting, only one anatomically accurate 3D-printed nasal airway phantom was used for model training and testing. The 25 mm observable depth range of the virtual camera was set to suit the nasal cavity size of the phantom. At this point, we should be aware that the 25 mm range dictates depth values of all pixels because all normalized depth values were scaled by this 25 mm factor to give depth values in mm, so that scale ambiguity was in theory solved. However, when considering generalization of our proposed method on real patient data, it would not be practical to address scale ambiguity by simply scaling depth output with a 25 mm factor, given that size of the nasal cavity varies across different patients. To tackle this problem, an initialization step is necessary such that the system recognizes the actual scale prior to depth estimation. A remedial solution is to utilize an external tracking device for calibrating the scaling factor. For instance, the presented EM-based tracking of endoscope tip could be utilized for this initialization step.

Lastly, no noise was added to the testing dataset for evaluating the robustness of this depth estimation approach. Signal-to-noise ratio of images vary across different endoscopic systems, which is affected by factors such as the noise reduction algorithm employed [158] and the condition of the image sensor. Assuming a depth estimation network trained fully on noise-free images, its generalization ability on testing datasets that have varying noise levels is expected to be suboptimal. Future work shall include testing on datasets with noise added. In addition, adding noise in the training dataset is also a common data augmentation strategy, potentially preventing the neural network from being overfit.

4.4.2 Quantitative Results of System Stability

To our observation, the latency discrepancy between EM sensor data and endoscopic image frame display was the major factor contributing to the visual feeling of “instability”. EM sensor pose was equivalent to the pose of camera in the virtual game world, thus EM sensor’s motion is directly related to apparent movement of the virtual graphics overlaid onto the 2D endoscopic view. Therefore, latency discrepancy would lead to a “temporal misalignment” between the endoscopic image and the overlaid virtual graphics. Note that

香港大學圖書館 HKU Libraries

this issue is not a consequence of the anatomical differences across patients. Instead, it is solely caused by latency discrepancy. As shown in **Fig. 4.6**, a latency discrepancy of 10 ms was not significant in our system, stability in terms of temporal alignment between augmented objects and physical targets was satisfactory. If this latency discrepancy was large, synchronization could be achieved by manually adding delay to the data stream that had a lower latency, so that latency discrepancy between data 2 streams could be minimized. However, this method might not be a satisfactory solution when latency of any one of the data streams was high. For instance, if endoscopic image frame display experienced a 300 ms delay, the virtual camera pose (EM sensor pose value) that originally had a 100 ms latency needed to be delayed for an extra 200 ms to match with image frame display latency. As such, the surgeon might feel that the augmented endoscopic view becomes inconsistent with his hands-on feeling when moving the endoscope. Therefore, a more effective solution is to minimize latency of both the EM sensor data streaming and endoscopic image streaming as much as possible. This involves the use of hardware with optimized computation, tracking modalities with a lower latency, and communication methods with a larger data transmission rate.

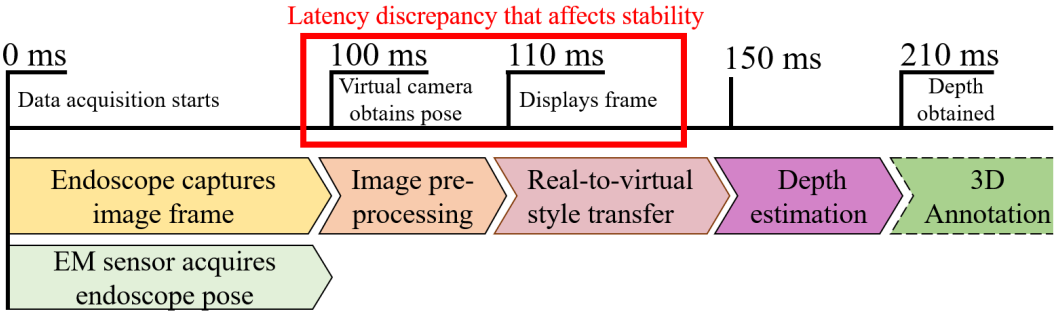

Fig. 4.6 Flowchart illustrating overall system latency. Temporal misalignment between virtual annotations and real objects in the endoscopic view contributes to the visual feeling of "instability". Main factor of this temporal misalignment is the latency discrepancy between i) the instance when virtual camera obtains a pose from the EM sensor and ii) the instance when an image frame is displayed. In this experiment, this latency discrepancy was only 10 ms, temporal misalignment was minimal.

Another noteworthy latency is the time required for depth estimation. As revealed in **Fig. 4.6**, real-to-virtual style transfer spent approximately 40 ms while depth estimation spent around 60 ms. Adding up the 110 ms duration between an image frame being taken and being displayed, a depth map was estimated around 210 ms after capturing of the image. It affects the position accuracy of the 3D annotation if the endoscope movement is rapid, because position of a 3D annotation at a pixel coordinate is derived from the depth value

香港大學圖書館 HKU Libraries

estimated at that pixel 210 ms ago. To avoid errors due to this factor, the endoscope was kept static when a 3D annotation was made. To fundamentally tackle this limitation so that an accurate 3D annotation can be made even when endoscope is moved rapidly, time required for real-to-virtual style transfer and depth estimation needs to be minimized. A possible solution is to tightly couple i) the style transfer network and ii) the depth estimation network together to form one network, so that computation time may be reduced. Currently, a style-transferred image is relayed from (i) to (ii) via a user datagram protocol (UDP) socket.

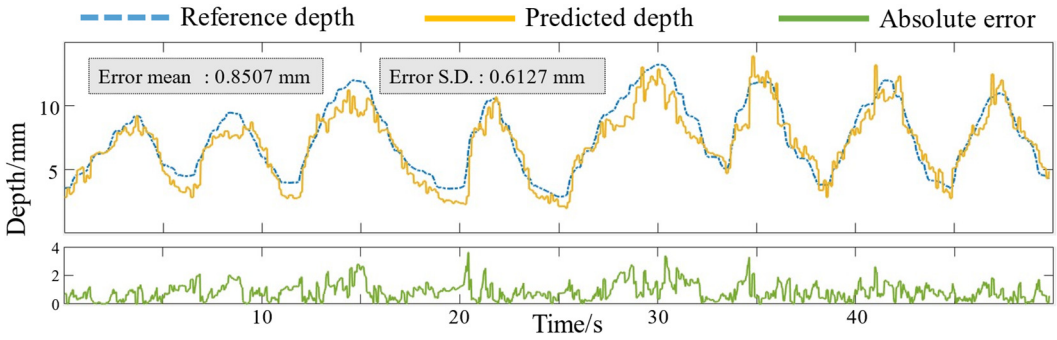

Fig. 4.7 Plot depicting one trial of the depth consistency evaluation. Reference depth and predicted depth were captured during forwards-backwards movement of the endoscope in the nasal phantom airway. Endoscope moving speed was maintained at around 3 mm/s and sampling rate was 50 Hz.

Next, stability in terms of geometric consistency was evaluated with the method described in **Section 4.3.2**. Average MAE between reference depth and predicted depth of 5 repeated trials was 1.1330 ± 0.9957 mm (or 5.5%-8.5% of the full 25 mm observable depth range of the virtual camera). While indicating a high accuracy, a low precision is revealed by a standard deviation being comparable to the MAE. The results showed rapid fluctuation of the predicted depth due to independent per-frame processing as described in Luo *et al.* [154]. This kind of geometric inconsistency in a temporal context can also be observed in **Fig. 4.7**, which depicts the datalogging process of one of the repeated trials. The predicted depth exhibited more fluctuations than the reference depth.

Although fluctuation in the predicted depth exists, the scale of predicted depth matches relatively well with the reference depth as shown in **Fig. 4.7.** A possible explanation is that the travel distance of the endoscope inside the nasal airway is relatively small compared to other anatomical sites like the colon, where scale drift is often observed in monocular depth estimation. In addition, our method design has indirectly minimized scale drift of the predicted depth. The depth prediction network we employed has a deterministic mapping

香港大學圖書館 HKU Libraries

which would likely output inaccurate results when inputs deviate from training data to a large extent. However, images are style-transferred to virtual-like images before being inputted into the depth estimation network. As long as virtual-like images resemble synthetic endoscopic images, predicted depth maps should be fairly coherent and consistent in terms of geometric scale.

Nonetheless, the maximum MAE recorded in this experiment was more than 3 mm. Compared with the "golden standard" 2 mm TRE for a navigation system to be considered safe to use [47], our results revealed that the proposed monocular depth estimation method was still not robust and accurate enough to be applied in an AR-assisted surgical guidance system. Especially for transnasal surgery, the tolerance for erroneous tracking and mapping is small due to the presence of narrow and delicate structures, as well as a high proximity to critical structures [38]. Finally, validation in this study did not involve cadaveric or human subjects. Accuracy and precision of this method were expected to be significantly affected due to adverse visual conditions such as presence of specular light, blood, and surgical instruments. Nevertheless, this study serves as a milestone and reference for our further development on vision-based tracking and mapping in endoscopic surgeries.

香港大學圖書館 HKU Libraries

4.5 CONCLUSION AND FUTURE WORK

In this work, we proposed a method to achieve real-time 3D annotation in a transnasal setting. Framewise depth is predicted from real-to-virtual domain transferred endoscopic images captured from within a nasal airway phantom, achieving a SSIM value of 0.8310 ± 0.0655. 3D annotation was achieved by integrating the EM-tracked endoscope pose with real-time predicted depth. Both the accuracy and stability evaluations demonstrated the feasibility and practicality of our proposed method.

It is worth noting that both the training and testing phases of our proposed method only involved data collected from a 3D-printed phantom. In our future work, we plan to collect video datasets of *ex-* and *in-vivo* nasal airway, in support to the generalizability of our method. Given that the image style transfer network would be trained with both *ex-* and *in-vivo* datasets, the accuracy and stability evaluations on unseen data should be comparable to the results obtained in this work. However, there is still an assumption that these unseen data have no adverse visual conditions such as surgical instrument obstructions and specular light reflection. We believe this preliminary work provides a proof-of-concept for further development towards a more generalizable system. Additionally, geometric inconsistency in the predicted depth will be addressed, potentially by adopting a self-supervised network that includes both depth and pose prediction, which may simultaneously be a more end-to-end estimation network with improved efficiency. With estimated poses, we also intend to explore the possibility of fusion between EM-acquired poses and poses estimated from monocular images, such that a more robust and stable tracking can be achieved. Thus, stability of a surgical AR system could subsequently be improved. Last but not least, our method currently applies to handheld endoscopes. When both pose and depth estimations are available, benefits of our method may extend beyond AR-based applications to robot-assisted MIS. Virtual fixtures and active constraints [159-161] are some example functions for enhancing surgeon's sensory feedback when robot control is involved.

香港大學圖書館 HKU Libraries

CHAPTER 5
CONCLUSION

Augmented reality is a technology that enables direct overlay of virtual images onto camera views, sparking a new opportunity to shape the future of the healthcare industry. Extra information such as subsurface critical structures, pre-operatively planned surgical paths, and surgical annotation can be augmented onto the endoscopic view, potentially increasing surgical ergonomics, efficiency, and safety. However, due to issues such as poor depth perception and visual cluttering, AR-assisted surgical guidance has not translated into mainstream clinical practice. In particular, accuracy is a major concern because erroneous augmentation may lead to visual disturbance and even major complications. Accuracy of an AR-assisted guidance system highly depends on tracking modalities, quality of patient's 3D anatomical models, and registration techniques. Regarding tracking modalities, limitations such as EM tracking interference and optical tracking line-of-sight issue may hinder system accuracy and robustness. Regarding patient's 3D anatomical models, their quality varies based on scanning quality, reconstruction software and human operation. Augmenting poorly segmented models onto the endoscopic view gives rise to an observed depth that is not representative of the real surface during a surgery. Therefore, this thesis aims to explore innovative sensing alternatives that can be incorporated in an endoscopic procedure,

香港大學圖書館 HKU Libraries

subsequently increases the stability, accuracy and ergonomics of an AR-assisted guidance system. Although transnasal endoscopy was the core focus in this work, the proposed sensing methods would also benefit other specialties that utilize endoscopes. For instance, arthroscopy, transoral endoscopy and otoscopy may be suitable candidates for applying our proposed methods. Similar to transnasal endoscopy, these procedures target at anatomies that involve low amounts of tissue deformation and have relatively rigid structures nearby for the use as registration references. Therefore, augmented guidance could be relatively well aligned with target anatomy on the image, given that our methods assume the surroundings can provide a sufficiently static frame of reference. For procedures that involve large-scale tissue deformation such as colonoscopies and gastroscopies, stereoscopic vision could be used to deduce relatively solid depth information through epipolar geometry.

In **Chapter 3**, a visual-strain fusion-based method for eye-in-hand camera pose estimation is introduced. Online learning-based pose estimation was performed by mapping FBG strain measurements to pose information. Sensing fusion between FBG-derived pose and SLAM-derived pose was then achieved, enabling smooth 6D image stitching even under adverse visual conditions such as the presence of obstacles, complete darkness, and intense lighting. Pose estimation was evaluated using a soft robot in this study. In fact, practical use of FBG on an endoscope is also viable, particularly for a flexible endoscope. Though, an extra external tracking modality will then be required to track the handle of a flexible endoscope because FBG-derived pose information assumes the FBG-wrapped object has a fixed reference frame. Nonetheless, this method is still beneficial when compared with the EM tracking method employed in **Chapter 4**. To elaborate, EM tracking requires attachment of an EM sensor at the tip of the flexible endoscope. With visual-strain fusion sensing, size of the endoscope tip and insertion tube can be minimized as no EM sensor is attached. In addition, while FBG and SLAM were only loosely coupled in this study, it is in fact possible to generate a new SLAM architecture by directly coupling FBG strain measurements with typical SLAM processes, such as estimator initialization, online loop detection, and re-localization. Such kind of visual-FBG SLAM system can potentially be applied for endoscopic navigation.

In **Chapter 4**, a monocular depth estimation method for achieving 3D annotations in a transnasal endoscopic surgery setting was introduced. A supervised depth estimation network was trained entirely in a virtual environment to predict depth from endoscopic images in real-time. During depth estimation, real endoscopic view was first style-transferred to a synthetic-like view using an adversarial network before being inputted into

香港大學圖書館 HKU Libraries

the depth estimation network. Both the accuracy and stability evaluations performed in an anatomically accurate nasal airway phantom demonstrated the feasibility and practicality of our proposed method. Finally, 3D annotation was achieved on the endoscopic view in the nasal airway phantom using the depth information obtained.

In theory, pose and depth estimation can be tightly coupled by deploying self-supervised learning. Though, as a preliminary proof-of-concept study, an EM sensor was employed for endoscope tracking to ensure more reliable tracking. By adopting self-supervised pose and depth estimation in our future work, an external sensing modality such as an EM sensor would not be necessary for pose measurement. In addition, compared with real-to-virtual domain transfer-based depth estimation that involves two separated neural networks, self-supervised depth estimation is a more compact alternative. Not only can network training efficiency be increased, computation speed for depth estimation may also be improved. Next, validation in this study did not involve cadaveric or human subjects. Future work on monocular endoscopic depth estimation would involve collection of more cadaver and patient nasal airway video datasets, alongside CT images for generation of virtual models. When a larger dataset is available for training and testing, we aspire to validate our method's generalizability across patients. In future work, to diversify our test cases with sufficient anatomical variation, we plan to collect video datasets from 10 patients and 10 cadaveric subjects. Each video clip should last for approximately 5 minutes at a framerate of 60 Hz. As there is no convention regarding the size of training dataset in the field of monocular depth estimation, we will take reference from Mahmood *et al.* [146], who used approximately 300k images in total and have adopted a similar depth estimation workflow as our own.

Although evaluations suggested feasibility of our proposed method, numerical errors revealed that both accuracy and stability are still not satisfactory for our method to be safely applied in clinical AR-assisted surgical guidance, especially for transnasal surgery where proximity to critical structures is high. In a highly dynamic surgical scene, adverse visual conditions such as presence of specular light reflection, blood, and surgical instruments are expected to severely lower the accuracy and precision of the proposed method. We believe that with the capability of monocular depth estimation networks as of today, a rough spatial clue can be rendered. However, a high level of accuracy that can meet the safety standard of clinical surgical navigation or AR-assisted guidance remains to be a challenging task. In the near future, technologies like stereo vision and structured light might still be the standard reliable methods for obtaining visual depth.

香港大學圖書館 HKU Libraries

REFERENCES

[1] Peters T.M., Linte C.A., Yaniv Z. and Williams J., *Mixed and augmented reality in medicine*. 2018: CRC Press.

[2] Vávra P., Roman J., Zonča P., Ihnát P., Němec M., Kumar J., Habib N. and El-Gendi A., *Recent development of augmented reality in surgery: a review.* Journal of healthcare engineering, 2017. **2017**.

[3] Wijsmuller A.R., Romagnolo L.G.C., Consten E., Melani A.E.F. and Marescaux J., *Navigation and Image-Guided Surgery*, in *Digital Surgery*. 2021, Springer. p. 137-144.

[4] Sukegawa S., Kanno T. and Furuki Y., *Application of computer-assisted navigation systems in oral and maxillofacial surgery.* Japanese Dental Science Review, 2018. **54**(3): p. 139-149.

[5] Waelkens P., Oosterom M.N.v., Berg N.S., Navab N. and van Leeuwen F.W., *Surgical navigation: an overview of the state-of-the-art clinical applications.* Radioguided Surgery, 2016: p. 57-73.

[6] Willems P., Van der Sprenkel J., Tulleken C., Viergever M. and Taphoorn M., *Neuronavigation and surgery of intracerebral tumours.* Journal of neurology, 2006. **253**(9): p. 1123-1136.

[7] Germano I.M., *The NeuroStation System for image-guided, frameless stereotaxy.* Neurosurgery, 1995. **37**(2): p. 348-350.

[8] Gumprecht H.K., Widenka D.C. and Lumenta C.B., *Brain Lab VectorVision neuronavigation system: technology and clinical experiences in 131 cases.* Neurosurgery, 1999. **44**(1): p. 97-104.

[9] ClaroNav. *Navient image guided navigation system receives Health Canada approval*. 2021 [cited 2022 5/6]; Available from: https://www.claronav.com/navient-image-guided-navigation-system-receives-health-canada-approval/.

[10] Citardi M.J., Yao W. and Luong A., *Next-generation surgical navigation systems in sinus and skull base surgery.* Otolaryngologic Clinics of North America, 2017. **50**(3): p. 617-632.

[11] Meola A., Cutolo F., Carbone M., Cagnazzo F., Ferrari M. and Ferrari V., *Augmented reality in neurosurgery: a systematic review.* Neurosurgical review, 2017. **40**(4): p. 537-548.

[12] Reardon E.J., *Navigational risks associated with sinus surgery and the clinical effects of implementing a navigational system for sinus surgery.* The Laryngoscope, 2002. **112**(S99): p. 1-19.

[13] Khan A., Meyers J.E., Yavorek S., O'Connor T.E., Siasios I., Mullin J.P. and Pollina J., *Comparing next-generation robotic technology with 3-dimensional computed tomography navigation technology for the insertion of posterior pedicle screws.* World Neurosurgery, 2019. **123**: p. e474-e481.

[14] Ravi B., Zahrai A. and Rampersaud R., *Clinical accuracy of computer-assisted two-dimensional fluoroscopy for the percutaneous placement of lumbosacral pedicle screws.* Spine, 2011. **36**(1): p. 84-91.

香港大學圖書館 HKU Libraries

[15] Atallah S., Martin-Perez B. and Larach S., *Image-guided real-time navigation for transanal total mesorectal excision: a pilot study.* Techniques in coloproctology, 2015. **19**(11): p. 679-684.

[16] Food and Drug Administration. *FDA Safety Communication: Navigational Accuracy Errors Associated with Frameless Stereotaxic Navigation Systems*. 2017 [cited 2022 20/2]; Available from: https://www.asahq.org/advocacy-and-asapac/fda-and-washington-alerts/fda-alerts/2017/06/frameless-stereotaxic-navigation-systems-fda-safety-communication.

[17] Chan J.Y., Holsinger F.C., Liu S., Sorger J.M., Azizian M. and Tsang R.K., *Augmented reality for image guidance in transoral robotic surgery.* Journal of Robotic Surgery, 2020. **14**(4): p. 579-583.

[18] Pratt P., Ives M., Lawton G., Simmons J., Radev N., Spyropoulou L. and Amiras D., *Through the HoloLens™ looking glass: augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels.* European radiology experimental, 2018. **2**(1): p. 1-7.

[19] Pessaux P., Diana M., Soler L., Piardi T., Mutter D. and Marescaux J., *Towards cybernetic surgery: robotic and augmented reality-assisted liver segmentectomy.* Langenbeck's archives of surgery, 2015. **400**(3): p. 381-385.

[20] Wong K., Yee H.M., Xavier B.A. and Grillone G.A., *Applications of augmented reality in otolaryngology: a systematic review.* Otolaryngology–Head and Neck Surgery, 2018. **159**(6): p. 956-967.

[21] Bernhardt S., Nicolau S.A., Soler L. and Doignon C., *The status of augmented reality in laparoscopic surgery as of 2016.* Medical image analysis, 2017. **37**: p. 66-90.

[22] Kim Y., Kim H. and Kim Y.O., *Virtual reality and augmented reality in plastic surgery: a review.* Archives of plastic surgery, 2017. **44**(3): p. 179.

[23] Tong H.-S., Ng Y.-L., Liu Z., Ho J.D., Chan P.-L., Chan J.Y. and Kwok K.-W., *Real-to-virtual domain transfer-based depth estimation for real-time 3D annotation in transnasal surgery: a study of annotation accuracy and stability.* International Journal of Computer Assisted Radiology and Surgery, 2021. **16**(5): p. 731-739.

[24] Chen L., Tang W., John N.W., Wan T.R. and Zhang J.J., *SLAM-based dense surface reconstruction in monocular Minimally Invasive Surgery and its application to Augmented Reality.* Computer methods and programs in biomedicine, 2018. **158**: p. 135-146.

[25] Li L., Yang J., Chu Y., Wu W., Xue J., Liang P. and Chen L., *A novel augmented reality navigation system for endoscopic sinus and skull base surgery: a feasibility study.* PLoS One, 2016. **11**(1): p. e0146996.

[26] Sielhorst T., Feuerstein M. and Navab N., *Advanced medical displays: A literature review of augmented reality.* Journal of Display Technology, 2008. **4**(4): p. 451-467.

[27] Kelly P.J., Alker Jr G.J. and Goerss S., *Computer-assisted stereotactic laser microsurgery for the treatment of intracranial neoplasms.* Neurosurgery, 1982. **10**(3): p. 324-331.

[28] Roberts D.W., Strohbehn J.W., Hatch J.F., Murray W. and Kettenberger H., *A frameless stereotaxic integration of computerized tomographic imaging and the operating microscope.* Journal of neurosurgery, 1986. **65**(4): p. 545-549.

[29] Qian L., Wu J.Y., DiMaio S.P., Navab N. and Kazanzides P., *A review of augmented reality in robotic-assisted surgery.* IEEE Transactions on Medical Robotics and Bionics, 2019. **2**(1): p. 1-16.

香港大學圖書館 HKU Libraries

[30] Bajura M., Fuchs H. and Ohbuchi R., *Merging virtual objects with the real world: Seeing ultrasound imagery within the patient.* ACM SIGGRAPH Computer Graphics, 1992. **26**(2): p. 203-210.

[31] Shahidi R., Wang B., Epitaux M., Grzeszczuk R. and Adler J., *Volumetric image guidance via a stereotactic endoscope.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 1998. Springer.

[32] Lapeer R., Chios P., Alusi G., Linney A.D., Davey M. and Tan A. *Computer assisted ENT surgery using augmented reality: preliminary results on the CAESAR project*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2000. Springer.

[33] Navab N., Blum T., Wang L., Okur A. and Wendler T., *First deployments of augmented reality in operating rooms.* Computer, 2012. **45**(7): p. 48-55.

[34] Sutherland I.E. *A head-mounted three dimensional display*. in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 1968.

[35] Francis C. *How Virtual Reality in Medicine and Healthcare Market May Turn as Winner after Covid-19 Scenario?* AMA Research & Media LLP. 2020 [cited 2022 20/2]; Available from: https://www.openpr.com/news/2021261/how-virtual-reality-in-medicine-and-healthcare-market-may-turn-as.

[36] Maltha J. *Philips showcases unique augmented reality concept for image-guided minimally invasive therapies developed with Microsoft*. Philips. 2019 [cited 2022 20/2]; Available from: https://www.philips.com/a-w/about/news/archive/standard/news/press/2019/20190224-philips-showcases-unique-augmented-reality-concept-for-image-guided-minimally-invasive-therapies-developed-with-microsoft.html.

[37] Edwards P., Chand M., Birlo M. and Stoyanov D., *The challenge of augmented reality in surgery.* Digital Surgery, 2021: p. 121-135.

[38] Castelnuovo P., Dallan I., Battaglia P. and Bignami M., *Endoscopic endonasal skull base surgery: past, present and future.* European Archives of Oto-Rhino-Laryngology, 2010. **267**(5): p. 649-663.

[39] Ramakrishnan V.R., Orlandi R.R., Citardi M.J., Smith T.L., Fried M.P. and Kingdom T.T. *The use of image-guided surgery in endoscopic sinus surgery: an evidence-based review with recommendations*. in *International forum of allergy & rhinology*. 2013. Wiley Online Library.

[40] Khor W.S., Baker B., Amin K., Chan A., Patel K. and Wong J., *Augmented and virtual reality in surgery—the digital surgical environment: applications, limitations and legal pitfalls.* Annals of translational medicine, 2016. **4**(23).

[41] Garcia J., Thoranaghatte R., Marti G., Zheng G., Caversaccio M. and González Ballester M.A., *Calibration of a surgical microscope with automated zoom lenses using an active optical tracker.* The International Journal of Medical Robotics and Computer Assisted Surgery, 2008. **4**(1): p. 87-93.

[42] Citardi M.J., Agbetoba A., Bigcas J.L. and Luong A. *Augmented reality for endoscopic sinus surgery with surgical navigation: a cadaver study*. in *International forum of allergy & rhinology*. 2016. Wiley Online Library.

[43] Liu W.P., Azizian M., Sorger J., Taylor R.H., Reilly B.K., Cleary K. and Preciado D., *Cadaveric feasibility study of da vinci si–assisted cochlear implant with augmented visual navigation for otologic surgery.* JAMA Otolaryngology–Head & Neck Surgery, 2014. **140**(3): p. 208-214.

香港大學圖書館 HKU Libraries

[44] Liu W.P., Richmon J.D., Sorger J.M., Azizian M. and Taylor R.H., *Augmented reality and cone beam CT guidance for transoral robotic surgery.* Journal of robotic surgery, 2015. **9**(3): p. 223-233.

[45] D'Agostino J., Wall J., Soler L., Vix M., Duh Q.-Y. and Marescaux J., *Virtual neck exploration for parathyroid adenomas: a first step toward minimally invasive image-guided surgery.* JAMA surgery, 2013. **148**(3): p. 232-238.

[46] Winne C., Khan M., Stopp F., Jank E. and Keeve E., *Overlay visualization in endoscopic ENT surgery.* International journal of computer assisted radiology and surgery, 2011. **6**(3): p. 401-406.

[47] Labadie R.F., Davis B.M. and Fitzpatrick J.M., *Image-guided surgery: what is the accuracy?* Current opinion in otolaryngology & head and neck surgery, 2005. **13**(1): p. 27-31.

[48] Samarakkody Z.M. and Abdullah B., *The use of image guided navigational tracking systems for endoscopic sinus surgery and skull base surgery: a review.* Egyptian Journal of Ear, Nose, Throat and Allied Sciences, 2016. **17**(3): p. 133-137.

[49] MathWorks. *What Is Camera Calibration?* 2022 [cited 2022 14/3]; Available from: https://www.mathworks.com/help/vision/ug/camera-calibration.html.

[50] Faculty of Physics, the Ludwig-Maximilians-Universität München. *Optical Tracking in Medical Physics*. 2021 [cited 2022 14/03]; Available from: https://www.med.physik.uni-muenchen.de/studium_lehre/advanced_laboratory_courses/p2/P2_manual.pdf.

[51] OpenCV. *Camera Calibration*. 2022 [cited 2022 14/3]; Available from: https://docs.opencv.org/3.4/dc/dbb/tutorial_py_calibration.html.

[52] Franz A.M., Haidegger T., Birkfellner W., Cleary K., Peters T.M. and Maier-Hein L., *Electromagnetic tracking in medicine—a review of technology, validation, and applications.* IEEE transactions on medical imaging, 2014. **33**(8): p. 1702-1725.

[53] Wu X. and Taylor R.H. *A framework for calibration of electromagnetic surgical navigation system*. in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*. 2003. IEEE.

[54] Schicho K., Figl M., Donat M., Birkfellner W., Seemann R., Wagner A., Bergmann H. and Ewers R., *Stability of miniature electromagnetic tracking systems.* Physics in Medicine & Biology, 2005. **50**(9): p. 2089.

[55] Wen J., *Electromagnetic tracking for medical imaging.* 2010.

[56] Pasku V., De Angelis A., De Angelis G., Arumugam D.D., Dionigi M., Carbone P., Moschitta A. and Ricketts D.S., *Magnetic field-based positioning systems.* IEEE Communications Surveys & Tutorials, 2017. **19**(3): p. 2003-2017.

[57] Jackson J.D., *Classical electrodynamics*. 1999, American Association of Physics Teachers.

[58] Schwartz M., *Principles of electrodynamics*. 2012: Courier Corporation.

[59] Balanis C.A., *Antenna theory: analysis and design*. 2015: John wiley & sons.

[60] Babic S.I. and Akyel C., *Calculating mutual inductance between circular coils with inclined axes in air.* IEEE Transactions on Magnetics, 2008. **44**(7): p. 1743-1750.

[61] Babic S., Sirois F., Akyel C. and Girardi C., *Mutual inductance calculation between circular filaments arbitrarily positioned in space: alternative to grover's formula.* IEEE transactions on magnetics, 2010. **46**(9): p. 3591-3600.

香港大學圖書館 HKU Libraries

[62] Yaniv Z. *Which pivot calibration?* in *Medical imaging 2015: Image-guided procedures, robotic interventions, and modeling*. 2015. International Society for Optics and Photonics.

[63] Feuerstein M. *Hand-Eye Calibration*. Computer Aided Medical Procedures & Augmented Reality, the Technical University of Munich. 2009 [cited 2022 15/3]; Available from: http://campar.in.tum.de/Chair/HandEyeCalibration.

[64] Tsai R.Y. and Lenz R.K., *A new technique for fully autonomous and efficient 3 D robotics hand/eye calibration.* IEEE Transactions on robotics and automation, 1989. **5**(3): p. 345-358.

[65] Horaud R. and Dornaika F., *Hand-eye calibration.* The international journal of robotics research, 1995. **14**(3): p. 195-210.

[66] Park F.C. and Martin B.J., *Robot sensor calibration: solving AX= XB on the Euclidean group.* IEEE Transactions on Robotics and Automation, 1994. **10**(5): p. 717-721.

[67] Daniilidis K., *Hand-eye calibration using dual quaternions.* The International Journal of Robotics Research, 1999. **18**(3): p. 286-298.

[68] Strobl K.H. and Hirzinger G. *Optimal hand-eye calibration*. in *2006 IEEE/RSJ international conference on intelligent robots and systems*. 2006. IEEE.

[69] Shen J., *Framework for ultrasonography-based augmented reality in robotic surgery: application to transoral surgery and gastrointestinal surgery*. 2019, Université Rennes 1.

[70] Talmadge J., Jiang Z.Y., Zebda D.A., Yao W.C., Luong A.U. and Citardi M.J., *Contour Map Point Distribution and Surgeon Experience Level Affect Accuracy of Surgical Navigation in a Pilot Study.* Annals of Otology, Rhinology & Laryngology, 2021: p. 00034894211005982.

[71] Grauvogel T.D., Engelskirchen P., Semper-Hogg W., Grauvogel J. and Laszig R., *Navigation accuracy after automatic-and hybrid-surface registration in sinus and skull base surgery.* PloS one, 2017. **12**(7): p. e0180975.

[72] Knott P.D., Batra P.S., Butler R.S. and Citardi M.J., *Contour and paired-point registration in a model for image-guided surgery.* The Laryngoscope, 2006. **116**(10): p. 1877-1881.

[73] Marques B., Plantefève R., Roy F., Haouchine N., Jeanvoine E., Peterlik I. and Cotin S. *Framework for augmented reality in Minimally Invasive laparoscopic surgery*. in *2015 17th International conference on E-health networking, application & services (HealthCom)*. 2015. IEEE.

[74] Mongen M.A. and Willems P.W., *Current accuracy of surface matching compared to adhesive markers in patient-to-image registration.* Acta Neurochirurgica, 2019. **161**(5): p. 865-870.

[75] Knott P.D., Batra P.S. and Citardi M.J., *Computer aided surgery: concepts and applications in rhinology.* Otolaryngologic Clinics of North America, 2006. **39**(3): p. 503-522.

[76] Danilchenko A. and Fitzpatrick J.M., *General approach to first-order error prediction in rigid point registration.* IEEE transactions on medical imaging, 2010. **30**(3): p. 679-693.

[77] Moghari M.H. and Abolmaesumi P., *Distribution of fiducial registration error in rigid-body point-based registration.* IEEE transactions on medical imaging, 2009. **28**(11): p. 1791-1801.

香港大學圖書館 HKU Libraries

[78] Sielhorst T., Bauer M., Wenisch O., Klinker G. and Navab N. *Online estimation of the target registration error for n-ocular optical tracking systems*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2007. Springer.

[79] Seginer A., *Rigid-body point-based registration: The distribution of the target registration error when the fiducial registration errors are given.* Medical image analysis, 2011. **15**(4): p. 397-413.

[80] Fitzpatrick J.M., West J.B. and Maurer C.R., *Predicting error in rigid-body point-based registration.* IEEE transactions on medical imaging, 1998. **17**(5): p. 694-702.

[81] Mandava V.R., Fitzpatrick J.M., Maurer Jr C.R., Maciunas R.J. and Allen G.S. *Registration of multimodal volume head images via attached markers*. in *Medical Imaging VI: Image Processing*. 1992. SPIE.

[82] Thompson S., Penney G., Dasgupta P. and Hawkes D., *Improved modelling of tool tracking errors by modelling dependent marker errors.* IEEE transactions on medical imaging, 2012. **32**(2): p. 165-177.

[83] Thompson S., Totz J., Song Y., Johnsen S., Stoyanov D., Ourselin S., Gurusamy K., Schneider C., Davidson B. and Hawkes D. *Accuracy validation of an image guided laparoscopy system for liver resection*. in *Medical imaging 2015: image-guided procedures, robotic interventions, and modeling*. 2015. SPIE.

[84] Thompson S., Schneider C., Bosi M., Gurusamy K., Ourselin S., Davidson B., Hawkes D. and Clarkson M.J., *In vivo estimation of target registration errors during augmented reality laparoscopic surgery.* International journal of computer assisted radiology and surgery, 2018. **13**(6): p. 865-874.

[85] Morvan Y., *Acquisition, compression and rendering of depth and texture for multi-view video.* 2009.

[86] Homberg B.S., Katzschmann R.K., Dogar M.R. and Rus D., *Robust proprioceptive grasping with a soft robot hand.* Autonomous Robots, 2019. **43**(3): p. 681-696.

[87] Hyatt P., Kraus D., Sherrod V., Rupert L., Day N. and Killpack M.D., *Configuration estimation for accurate position control of large-scale soft robots.* IEEE/ASME Transactions on Mechatronics, 2018. **24**(1): p. 88-99.

[88] Melekhov I., Ylioinas J., Kannala J. and Rahtu E., *Relative camera pose estimation using convolutional neural networks*, in *International Conference on Advanced Concepts for Intelligent Vision Systems*. 2017, Springer. p. 675-687.

[89] Maidi M., Ababsa F. and Mallem M., *Vision-inertial tracking system for robust fiducials registration in augmented reality*, in *2009 IEEE Symposium on Computational Intelligence for Multimedia Signal and Vision Processing*. 2009, IEEE. p. 83-90.

[90] Li Y., Snavely N., Huttenlocher D. and Fua P., *Worldwide pose estimation using 3d point clouds*, in *European Conference on Computer Vision*. 2012, Springer. p. 15-29.

[91] Nöll T., Pagani A. and Stricker D., *Markerless camera pose estimation-an overview*, in *Visualization of Large and Unstructured Data Sets-Applications in Geospatial Planning, Modeling and Engineering (IRTG 1131 Workshop)*. 2011, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[92] Shih C.-L. and Lee Y., *A Simple Robotic Eye-In-Hand Camera Positioning and Alignment Control Method Based on Parallelogram Features.* Robotics, 2018. **7**(2): p. 31.

[93] Yoshimi B.H. and Allen P.K., *Active, uncalibrated visual servoing*, in *IEEE International Conference on Robotics and Automation*. 1994, IEEE. p. 156-161.

香港大學圖書館 HKU Libraries

[94] Flandin G., Chaumette F. and Marchand E., *Eye-in-hand/eye-to-hand cooperation for visual servoing*, in *IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*. 2000, IEEE. p. 2741-2746.

[95] Gemeiner P., Einramhof P. and Vincze M., *Simultaneous motion and structure estimation by fusion of inertial and vision data.* The International Journal of Robotics Research, 2007. **26**(6): p. 591-605.

[96] Kendall A., Grimes M. and Cipolla R., *Posenet: A convolutional network for real-time 6-dof camera relocalization*, in *IEEE International Conference on Computer Vision*. 2015. p. 2938-2946.

[97] Rambach J.R., Tewari A., Pagani A. and Stricker D., *Learning to fuse: A deep learning approach to visual-inertial camera pose estimation*, in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2016, IEEE. p. 71-76.

[98] Eckenhoff K., Geneva P. and Huang G., *Sensor-failure-resilient multi-imu visual-inertial navigation*, in *International Conference on Robotics and Automation*. 2019, IEEE. p. 3542-3548.

[99] Mirzaei F.M. and Roumeliotis S.I., *A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation.* IEEE Transactions on Robotics, 2008. **24**(5): p. 1143-1156.

[100] Qin T., Li P. and Shen S., *Vins-mono: A robust and versatile monocular visual-inertial state estimator.* IEEE Transactions on Robotics, 2018. **34**(4): p. 1004-1020.

[101] Medgadget. *Flex Robotic System, a Snake to Navigate Colon for Transanal Endoscopic Procedures*. 2017 [cited 2022 20/4] Available from: https://www.medgadget.com/2017/05/flex-robotic-system-snake-navigate-colon-transanal-endoscopic-procedures.html.

[102] Ryu S.C. and Dupont P.E., *FBG-based shape sensing tubes for continuum robots*, in *IEEE International Conference on Robotics and Automation*. 2014, IEEE. p. 3531-3537.

[103] Liu H., Farvardin A., Grupp R., Murphy R.J., Taylor R.H., Iordachita I. and Armand M., *Shape tracking of a dexterous continuum manipulator utilizing two large deflection shape sensors.* IEEE Sensors Journal, 2015. **15**(10): p. 5494-5503.

[104] Zhuang W., Sun G., Li H., Lou X., Dong M. and Zhu L., *FBG based shape sensing of a silicone octopus tentacle model for soft robotics.* Optik, 2018. **165**: p. 7-15.

[105] Roesthuis R.J., Kemp M., van den Dobbelsteen J.J. and Misra S., *Three-dimensional needle shape reconstruction using an array of fiber bragg grating sensors.* IEEE/ASME transactions on mechatronics, 2013. **19**(4): p. 1115-1126.

[106] Seifabadi R., Gomez E.E., Aalamifar F., Fichtinger G. and Iordachita I., *Real-time tracking of a bevel-tip needle with varying insertion depth: Toward teleoperated MRI-guided needle steering*, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2013, IEEE. p. 469-476.

[107] Shi C., Luo X., Qi P., Li T., Song S., Najdovski Z., Fukuda T. and Ren H., *Shape sensing techniques for continuum robots in minimally invasive surgery: A survey.* IEEE Transactions on Biomedical Engineering, 2016. **64**(8): p. 1665-1678.

[108] Reisenauer J., Simoff M.J., Pritchett M.A., Ost D.E., Majid A., Keyes C., Casal R.F., Parikh M.S., Diaz-Mendoza J. and Fernandez-Bussy S., *Ion: technology and techniques for shape-sensing robotic-assisted bronchoscopy.* The Annals of Thoracic Surgery, 2022. **113**(1): p. 308-315.

香港大學圖書館 HKU Libraries

[109] Ha X.T., Ourak M., Al-Ahmad O., Wu D., Borghesan G., Menciassi A. and Vander Poorten E., *Robust catheter tracking by fusing electromagnetic tracking, fiber bragg grating and sparse fluoroscopic images.* IEEE Sensors Journal, 2021. **21**(20): p. 23422-23434.

[110] Xu R., Yurkewich A. and Patel R.V., *Curvature, torsion, and force sensing in continuum robots using helically wrapped FBG sensors.* IEEE Robotics Automation Letters, 2016. **1**(2): p. 1052-1059.

[111] Sefati S., Hegeman R., Alambeigi F., Iordachita I. and Armand M., *FBG-based position estimation of highly deformable continuum manipulators: Model-dependent vs. data-driven approaches*, in *International Symposium on Medical Robotics*. 2019, IEEE. p. 1-6.

[112] Saccomandi P., Oddo C.M., Zollo L., Formica D., Romeo R.A., Massaroni C., Caponero M.A., Vitiello N., Guglielmelli E. and Silvestri S., *Feedforward neural network for force coding of an MRI-compatible tactile sensor array based on fiber Bragg grating.* Journal of Sensors, 2015. **2015**.

[113] Lun T.L.T., Wang K., Ho J.D., Lee K.-H., Sze K.Y. and Kwok K.-W., *Real-time surface shape sensing for soft and flexible structures using fiber Bragg gratings.* IEEE Robotics and Automation Letters, 2019. **4**(2): p. 1454-1461.

[114] Xiong W., Cai C. and Kong X., *Instrumentation design for bridge scour monitoring using fiber Bragg grating sensors.* Applied Optics, 2012. **51**(5): p. 547-557.

[115] Wang X., Fang G., Wang K., Xie X., Lee K.-H., Ho J.D., Tang W.L., Lam J. and Kwok K.-W., *Eye-in-Hand Visual Servoing Enhanced With Sparse Strain Measurement for Soft Continuum Robots.* IEEE Robotics Automation Letters, 2020. **5**(2): p. 2161-2168.

[116] Alambeigi F., Pedram S.A., Speyer J.L., Rosen J., Iordachita I., Taylor R.H. and Armand M., *SCADE: Simultaneous Sensor Calibration and Deformation Estimation of FBG-Equipped Unmodeled Continuum Manipulators.* IEEE Transactions on Robotics, 2019. **36**(1): p. 222-239.

[117] Huang G.-B., Zhu Q.-Y. and Siew C.-K., *Extreme learning machine: a new learning scheme of feedforward neural networks*, in *IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*. 2004, IEEE. p. 985-990.

[118] Huang G.-B., Zhou H., Ding X. and Zhang R., *Extreme learning machine for regression and multiclass classification.* IEEE Transactions on Systems, Man, Cybernetics, Part B, 2011. **42**(2): p. 513-529.

[119] Huang G.-B., Liang N.-Y., Rong H.-J., Saratchandran P. and Sundararajan N., *On-line sequential extreme learning machine.* Computational Intelligence, 2005. **2005**: p. 232-237.

[120] Li T., *Research on soft-sensing methods for the size of melt pool in MgO single crystal furnace*. 2013, Dalian University of Technology.

[121] Fu H.-C., Ho J.D., Lee K.-H., Hu Y.C., Au S.K., Cho K.-J., Sze K.Y. and Kwok K.-W., *Interfacing soft and hard: a spring reinforced actuator.* Soft Robotics, 2020. **7**(1): p. 44-58.

[122] Fang G., Chow M.C., Ho J.D., He Z., Wang K., Ng T., Tsoi J.K., Chan P.-L., Chang H.-C. and Chan D.T.-M., *Soft robotic manipulator for intraoperative MRI-guided transoral laser microsurgery.* Science Robotics, 2021. **6**(57): p. eabg5575.

[123] Dong Z., Wang X., Fang G., He Z., Ho J.D.L., Cheung C.-L., Tang W.L., Xie X., Liang L., Chang H.-C., Ching C.K. and Kwok K.-W., *Shape Tracking and Feedback Control of Cardiac Catheter Using MRI-guided Robotic Platform – Validation with Pulmonary Vein Isolation Simulator in MRI.* IEEE Transactions on Robotics, 2022 (Accepted).

香港大學圖書館 HKU Libraries

[124] Treter S., Perrier N., Sosa J.A. and Roman S., *Telementoring: a multi-institutional experience with the introduction of a novel surgical approach for adrenalectomy.* Annals of surgical oncology, 2013. **20**(8): p. 2754-2758.

[125] Kwok K.-W., Sun L.-W., Mylonas G.P., James D.R., Orihuela-Espina F. and Yang G.-Z., *Collaborative gaze channelling for improved cooperation during robotic assisted surgery.* Annals of biomedical engineering, 2012. **40**(10): p. 2156-2167.

[126] Yang G.-Z., Mylonas G.P., Kwok K.-W. and Chung A. *Perceptual docking for robotic control*. in *International Workshop on Medical imaging and virtual reality*. 2008. Springer.

[127] Vitiello V., Kwok K.-W. and Yang G.-Z., *Introduction to robot-assisted minimally invasive surgery (MIS)*, in *Medical Robotics*. 2012, Elsevier. p. 1-P1.

[128] Kwok K.W., Sun L.W., Vitiello V., James D.R., Mylonas G.P., Darzi A. and Yang G.-Z. *Perceptually docked control environment for multiple microbots: application to the gastric wall biopsy*. in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2009. IEEE.

[129] Chetwood A.S., Kwok K.-W., Sun L.-W., Mylonas G.P., Clark J., Darzi A. and Yang G.-Z., *Collaborative eye tracking: a potential training tool in laparoscopic surgery.* Surgical endoscopy, 2012. **26**(7): p. 2003-2009.

[130] Bogen E.M., Augestad K.M., Patel H.R. and Lindsetmo R.-O., *Telementoring in education of laparoscopic surgeons: An emerging technology.* World journal of gastrointestinal endoscopy, 2014. **6**(5): p. 148.

[131] Lee S.-L., Lerotic M., Vitiello V., Giannarou S., Kwok K.-W., Visentini-Scarzanella M. and Yang G.-Z., *From medical images to minimally invasive intervention: Computer assistance for robotic surgery.* Computerized Medical Imaging and Graphics, 2010. **34**(1): p. 33-45.

[132] Stoyanov D., Scarzanella M.V., Pratt P. and Yang G.-Z. *Real-time stereo reconstruction in robotically assisted minimally invasive surgery*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2010. Springer.

[133] Mirota D.J., Wang H., Taylor R.H., Ishii M., Gallia G.L. and Hager G.D., *A system for video-based navigation for endoscopic endonasal skull base surgery.* IEEE Transactions on Medical Imaging, 2011. **31**(4): p. 963-976.

[134] Mur-Artal R., Montiel J.M.M. and Tardos J.D., *ORB-SLAM: a versatile and accurate monocular SLAM system.* IEEE transactions on robotics, 2015. **31**(5): p. 1147-1163.

[135] Mahmoud N., Cirauqui I., Hostettler A., Doignon C., Soler L., Marescaux J. and Montiel J. *ORBSLAM-based endoscope tracking and 3D reconstruction*. in *International workshop on computer-assisted and robotic endoscopy*. 2016. Springer.

[136] Ma R., Wang R., Pizer S., Rosenman J., McGill S.K. and Frahm J.-M. *Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019. Springer.

[137] Fu H., Gong M., Wang C., Batmanghelich K. and Tao D. *Deep ordinal regression network for monocular depth estimation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[138] Alhashim I. and Wonka P., *High quality monocular depth estimation via transfer learning*. arXiv preprint arXiv:1812.11941, 2018.

香港大學圖書館 HKU Libraries

[139] Liu X., Sinha A., Ishii M., Hager G.D., Reiter A., Taylor R.H. and Unberath M., *Dense depth estimation in monocular endoscopy with self-supervised learning methods.* IEEE transactions on medical imaging, 2019. **39**(5): p. 1438-1447.

[140] Reiter A., Léonard S., Sinha A., Ishii M., Taylor R.H. and Hager G.D. *Endoscopic-CT: learning-based photometric reconstruction for endoscopic sinus surgery*. in *Medical Imaging 2016: Image Processing*. 2016. International Society for Optics and Photonics.

[141] Isola P., Zhu J.-Y., Zhou T. and Efros A.A. *Image-to-image translation with conditional adversarial networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[142] Rau A., Edwards P.E., Ahmad O.F., Riordan P., Janatka M., Lovat L.B. and Stoyanov D., *Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy.* International journal of computer assisted radiology and surgery, 2019. **14**(7): p. 1167-1176.

[143] Atapour-Abarghouei A. and Breckon T.P. *Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[144] Zhu J.-Y., Park T., Isola P. and Efros A.A. *Unpaired image-to-image translation using cycle-consistent adversarial networks*. in *Proceedings of the IEEE international conference on computer vision*. 2017.

[145] Zhao C., Shen M., Sun L. and Yang G.-Z., *Generative Localization With Uncertainty Estimation Through Video-CT Data for Bronchoscopic Biopsy.* IEEE Robotics and Automation Letters, 2019. **5**(1): p. 258-265.

[146] Mahmood F., Chen R. and Durr N.J., *Unsupervised reverse domain adaptation for synthetic medical images via adversarial training.* IEEE transactions on medical imaging, 2018. **37**(12): p. 2572-2581.

[147] Yang W.-f., Choi W.S., Wong M.C.-M., Powcharoen W., Zhu W.-y., Tsoi J.K.-H., Chow M., Kwok K.-W. and Su Y.-x., *Three-dimensionally printed patient-specific surgical plates increase accuracy of oncologic head and neck reconstruction versus conventional surgical plates: a comparative study.* Annals of surgical oncology, 2021. **28**(1): p. 363-375.

[148] Fan Y., Yang F., Cheung G.S.-H., Chan A.K.-Y., Wang D.D., Lam Y.-Y., Chow M.C.-K., Leong M.C.-W., Kam K.K.-H. and So K.C.-Y., *Device sizing guided by echocardiography-based three-dimensional printing is associated with superior outcome after percutaneous left atrial appendage occlusion.* Journal of the American Society of Echocardiography, 2019. **32**(6): p. 708-719. e1.

[149] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. and Bengio Y. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.

[150] He K., Zhang X., Ren S. and Sun J. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[151] Kingma D.P. and Ba J., *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

[152] Wang L., Shen X., Zhang J., Wang O., Lin Z., Hsieh C.-Y., Kong S. and Lu H., *DeepLens: shallow depth of field from a single image.* arXiv preprint arXiv:1810.08100, 2018.

[153] Arun K.S., Huang T.S. and Blostein S.D., *Least-squares fitting of two 3-D point sets.* IEEE Transactions on pattern analysis and machine intelligence, 1987(5): p. 698-700.

香港大學圖書館 HKU Libraries

[154] Luo X., Huang J.-B., Szeliski R., Matzen K. and Kopf J., *Consistent video depth estimation.* arXiv preprint arXiv:2004.15021, 2020.

[155] Tong H.-S. *Depth Estimation for Real-time 3D Annotation in Transnasal Surgery*. 2021; Available from: https://youtu.be/kCi1ux-Q1FQ.

[156] Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P., *Image quality assessment: from error visibility to structural similarity.* IEEE transactions on image processing, 2004. **13**(4): p. 600-612.

[157] Nadeem S. and Kaufman A. *Computer-aided detection of polyps in optical colonoscopy images*. in *Medical Imaging 2016: Computer-Aided Diagnosis*. 2016. International Society for Optics and Photonics.

[158] Obukhova N., Motyko A., Pozdeev A. and Timofeev B. *Review of noise reduction methods and estimation of their effectiveness for medical endoscopic images processing*. in *2018 22nd Conference of Open Innovations Association (FRUCT)*. 2018. IEEE.

[159] Kwok K.-W., Mylonas G.P., Sun L.W., Lerotic M., Clark J., Athanasiou T., Darzi A. and Yang G.-Z. *Dynamic active constraints for hyper-redundant flexible robots*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2009. Springer.

[160] Kwok K.-W., Vitiello V. and Yang G.-Z. *Control of articulated snake robot under dynamic active constraints*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2010. Springer.

[161] Payne C.J., Kwok K.-W. and Yang G.-Z. *An ungrounded hand-held surgical device incorporating active constraints with force-feedback*. in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2013. IEEE.

香港大學圖書館 HKU Libraries