Imperial College London

A Gaze-contingent Framework for Perceptually-Enabled Applications in Healthcare

Alexandros Kogkas

Supervised by:
George Mylonas, Ph.D. and Ara Darzi, M.D., F.R.C.S.

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Clinical Medicine Research (Surgery and Cancer)
of Imperial College London and the Diploma of Imperial College

Department of Surgery and Cancer
Imperial College London

August 2019

I hereby declare that this thesis and the work reported herein was composed and originated by myself. Information derived from the published and unpublished work of others was appropriately referenced and credited.

Alexandros Kogkas (2019)

Abstract

Patient safety and quality of care remain the focus of the smart operating room of the future. Some of the most influential factors with a detrimental effect are related to suboptimal communication among the staff, poor flow of information, staff workload and fatigue, ergonomics and sterility in the operating room. While technological developments constantly transform the operating room layout and the interaction between surgical staff and machinery, a vast array of opportunities arise for the design of systems and approaches, that can enhance patient safety and improve workflow and efficiency.

The aim of this research is to develop a real-time gaze-contingent framework towards a "smart" operating suite, that will enhance operator's ergonomics by allowing perceptually-enabled, touchless and natural interaction with the environment. The main feature of the proposed framework is the ability to acquire and utilise the plethora of information provided by the human visual system to allow touchless interaction with medical devices in the operating room. In this thesis, a gaze-guided robotic scrub nurse, a gaze-controlled robotised flexible endoscope and a gaze-guided assistive robotic system are proposed. Firstly, the gaze-guided robotic scrub nurse is presented; surgical teams performed a simulated surgical task with the assistance of a robot scrub nurse, which complements the human scrub nurse in delivery of surgical instruments, following gaze selection by the surgeon. Then, the gaze-controlled robotised flexible endoscope is introduced; experienced endoscopists and novice users performed a simulated examination of the upper gastrointestinal tract using predominately their natural gaze. Finally, a gaze-guided assistive robotic

system is presented, which aims to facilitate activities of daily living. The results of this work provide valuable insights into the feasibility of integrating the developed gaze-contingent framework into clinical practice without significant workflow disruptions.

Acknowledgements

I would like to start expressing my special thanks to my supervisor Dr George Mylonas. Without his guidance and constant support, this work would not have been possible. His problem-solving skills and dedication, being able to provide support at any time of the day, have been invaluable and a source of inspiration. I am undoubtedly grateful for the skills that I have developed and the opportunities that I was granted during this PhD and I could not have hoped for a better supervision. Thank you for believing in me and giving me the opportunity to work in such an impactful project within in a great and multi-disciplinary environment.

I would like to thank my second supervisor Professor Lord Ara Darzi. I am really grateful that I had the opportunity to work alongside a personality with such a unique vision in healthcare. I would also like to acknowledge the crucial role of the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC), who supported this research work.

Furthermore, I would like to thank Dr Steve McAteer and Dr Karen Kerr for their availability and effort. Special thanks to Dr Stefan Leutenegger for his technical guidance, Dr Nisha Patel for her clinical feedback and Ahmed Ezzat for his collaboration during my PhD.

Throughout my PhD journey I had the opportunity to work alongside brilliant researchers and friends; Fernando, Joric, Ming, Mark, Eftychios, James, Giovanni, Juana and Mike. It has been a pleasure to share this journey with you. I would also like to thank Efthymios for his invaluable friendly support.

The last years would not have been the same without Maira. Thank

you for the moments and your unfaltering support, especially during the last challenging year.

Last but not least, I would like to express my enormous gratitude to my family, especially my parents Andreas and Giota, my siblings Souzana and Stefanos and my grandparents Christos and Despoina. You have always been supporting, encouraging and inspiring me to whatever I pursue, and I would not be where I am without you. You are the greatest motivation to me. Lastly, I want to dedicate this PhD thesis to my beloved uncle Tasos.

Contents

List of Tables

List of Figures

Abbreviations

ADL	Activities of Daily Living.
ALS	amyotrophic lateral sclerosis.
DOF	degrees of freedom.
EEG	Electroencephalography.
EF	Effort.
ETG	eye-tracking glasses.
F/T sensor	Force/Torque sensor.
FOV	field of view.
FR	Frustration.
GC	Gaze-contingent Control.
GUI	graphical user interface.
HC	Hand Control.
HSN	Human scrub nurse.
HSNt	Human scrub nurse only task.
MCS	motion capture system.
MCS CS	motion capture system coordinate frame.
MD	Mental Demand.
OP	Own Performance.
PD	Physical Demand.
PoR	point of regard.

R&HSNt	Robot and human scrub nurse task.
RCS	robot coordinate system.
RSN	Robotic scrub nurse.
SLAM	Simultaneous Localisation and Mapping.
SPHt	Spherical cavity task.
ST	Surgical trainee (surgeon).
TD	Temporal Demand.
UGIt	Upper Gastrointestinal tract task.
WCS	world coordinate system.
WMRM	wheelchair-mounted robotic manipulators.

Chapter 1

Introduction

A safe operating room (OR) has to be constantly adapted to the introduction of new technologies and increasingly complex surgical procedures. New technologies may add complexity to the surgical workflow, but at the same time provide new opportunities for the design of systems and approaches that can enhance patient safety and improve workflow and efficiency. Several studies have been carried out to establish the requirements of the OR of the future, focusing on surgical workflow optimisation, system integration and standardisation, particularly in image guided surgery [38]. Seagull et al. in [39] identified four strategic areas where solutions to problems would be of paramount importance for the evolution of the OR of the future: cognitive simulation, informatics, "smart image" and ergonomics/human factors. Similarly, Bharathan et al. in [40] identify ergonomics, imaging, navigation, medical informatics, training and simulation as the key innovation areas.

It is anticipated that all equipment in the OR of the future will be fully integrated and networked into a smart operating suite. For this purpose, fully integrated OR suites are being provided by companies, such as Karl Storz's OR1 NEO™ [41], where the entire surgical environment (e.g. endoscopic devices, video/data sources, surgical table, ceiling lights) can be tailored to and by the user and can be controlled from a central location within a sterile area. Similarly, the OR.net project [42]

provides a framework to connect medical devices and IT systems under a safe standard protocol. Such operating rooms, where a large amount of information is made available through a unique integrated system, offer tremendous opportunities for implementing novel human-computer interfaces, context-aware systems, automated procedures and augmented visualisation features.

In recent decades, the advent of minimally invasive surgery (MIS) has transformed the OR towards a technology-centred space. Therefore, developments in human-computer interaction research need to be adopted in the surgical workflow, to allow perceptually enabled interactions with medical technologies.

By approaching the eyes as the only visible part of the brain, we can consider them not just in the conventional sense as receptors of visual stimuli, but also as perception- and cognition-rich actors within the operating theatre. This approach could allow harnessing the power of the underlying mental processes that lead from perception to cognition to action, and seamlessly endow intelligence to implemented human-computer interfaces. To this end, *eye-tracking* can be used to measure ocular movement and the point of gaze.

Eye-tracking has been used in various areas, to improve driving safety [43], convey perceptual skills [44], evaluate cognitive activity [45], monitor the situation awareness by understanding the behaviour of expert and novice pilots [46], as well as for human-computer interaction [47]. McMullen et al. in [48] used eye-tracking combined with EEG as a brain-machine interface to control a robotic prosthetic limb. In clinical settings, eye-tracking has been successfully used to enhance collaboration by sharing gaze information between supervisor and trainee [21, 49] and to distinguish between novice and expert surgeons [50, 51]. Eye-tracking has also been used in objective measurement of surgical skills [52], for enhancing training skills [53], for revealing opportunities to improve performance [54] by facilitating surgical skill acquisition, as well as for gaze control of surgical instruments [24]

1.1 Research Objectives

The full potential of eye-tracking, is significantly undermined when semantic and contextual correlation between the captured visual attention and the environment cannot be established. When it comes to eye-tracking in the OR, knowledge of the dynamic interactions between the theatre attendants and the environment is limited without information on the objects being fixated at or manipulated. Reconstruction and segmentation of the theatre space along with recognition and tracking of objects will allow invaluable information to be acquired on surgical workflow and on the attendants' behaviour. Real-time object recognition is a challenging task widely used in several domains. One such example is the rapidly evolving field of self-driving cars, where the safety of driver and pedestrians is paramount. Robust pedestrian and obstacle detection through reconstruction and segmentation of the environment is a fundamental requirement for acquiring perceptual information and for ensuring safety [55]. Numerous other examples are available, for providing artificial vision to blind people [56, 57] and to mobile robots [58], capturing traffic information [59] and facilitating public surveillance [60]. Within the operating room, Allan et al. in [61] introduced micro-scale detection and localisation of surgical instruments during minimal invasive surgery.

Moreover, a significant body of research has explored "perceptually enabled" interactions in the sterile environment using technologies like 3D cameras, voice commands or eye-tracking [2]. This way the surgeon can be kept in the loop of decision-making and task-execution in a seamless way that is likely to help improving overall operational performance and reducing communication errors. Eye-tracking has the potential to provide a "third hand" and a seamless way to allow perceptually enabled interactions within the surgical environment.

The aim of this PhD research is to develop a real-time gaze-contingent framework that enhances the operator's ergonomics applied to several clinical contexts. Eye-tracking, object recognition/tracking and robotic

assistance are employed to allow touchless and natural interaction and integration with the environment.

The research questions of the thesis are summarised as follows:

- How can the 3D point of regard of a user be estimated in a free-view fashion?
- Can the free-view 3D gaze framework be used in the operating theatre in a simulated surgical task (ex vivo) with clinical teams and how does a gaze-controlled robotic scrub nurse affect the performance and ergonomics of the team?
- Can the free-view 3D gaze framework be used in simulated flexible endoscopy and how does a gaze-controlled flexible endoscope affect the performance and ergonomics of experienced endoscopists and novices?
- How can the free-view 3D gaze framework be used for robotic assistance in activities of daily living?

1.2 Original Contributions

The key original contributions of this thesis include:

- Development of a novel framework for real-time, free-view, global 3D fixation localisation, through the use of unrestricted wearable eye-tracking, dynamic spatial 3D reconstruction and camera pose estimation techniques.
- Use of the framework to allow real-time, free-view fixation-guided recognition and tracking of objects.
- Semantic and contextual correlation between visual attention and environment.

- Use of the framework to allow simultaneous global macro- (theatre-wide) and micro-scale (patient/screen-wise) fixation localisation in the operating theatre.
- Development and clinical evaluation of a free-view, gaze-controlled robotic scrub nurse for the operating theatre.
- Development and clinical evaluation of an intuitive, fully motorised gaze-controlled system for non-restricting, free-view flexible endoscopy.
- Development and evaluation of a free-view, 3D gaze-guided assistive robotic system for activities of daily living.

1.3 Publications

The work presented in the thesis has resulted a number of peer-reviewed conference papers and journal publications:

- **Kogkas AA**, Ezzat A, Thakkar R, Darzi A, Mylonas GP (2019), "Free-view, 3D Gaze-Guided Robotic Scrub Nurse", International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). **Single track oral talk presentation**
  Chapter 4 is based on this article.
- **Kogkas AA**, Glover B, Patel N, Darzi A, Mylonas GP (2019), "Gaze-contingent Robotic Flexible Endoscopy", Hamlyn Symposium on Medical Robotics
  Parts of Chapter 5 are based on this article.
- **Kogkas AA**, Ezzat A, Darzi A, Mylonas GP (2018), "Free-view Gaze Controlled Image Navigation; One application of a Perceptually-enabled Smart Operating Room", 8th Joint workshop on new technologies for computer/robot assisted surgery (CRAS)

- Wang MY*, **Kogkas AA***, Darzi A, Mylonas GP (2018), "Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). **Joint first author**
  Chapter 6 is based on this article.

- Oude Vrielink TJC, Gonzalez-Bueno Puyal J, **Kogkas AA**, Mylonas GP (2018), "Intuitive Gaze-Control of a Robotized Flexible Endoscope", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
  Parts of Chapter 5 are based on this article.

- **Kogkas AA**, Darzi A, Mylonas GP (2017), "Gaze-contingent perceptually enabled interactions in the operating theatre", International Journal of Computer Assisted Radiology and Surgery
  Parts of Chapter 3 are based on this article.

- **Kogkas AA**, Darzi A, Mylonas GP (2016), "Gaze-driven human-robot interaction in the operating theatre", 6th Joint workshop on new technologies for computer/robot assisted surgery (CRAS)

- **Kogkas AA**, Sodergren MH, Darzi A, Mylonas GP (2016), "Macro- and micro- scale 3D gaze tracking in the operating theatre", Hamlyn Symposium on Medical Robotics

1.4 Thesis Overview

The outline of the thesis is as follows:

Chapter 2 provides an overview of the safety risks which lurk in the OR first. Then, an overview of touchless interaction modalities with emphasis in gaze-contingent systems and applications is presented. Finally, key aspects of the future OR that have been identified in the literature are reviewed. The aim is to emphasise the value and potential of the gaze-contingent interaction modality the proposed framework introduces in the

context of surgery, and the perceptually enabled data it can contribute to the recently emerged field of Surgical Data Science towards Surgery 4.0.

Chapter 3 proposes a 3D fixation localisation framework developed with the synergy of conventional wearable eye-tracking glasses, motion capture system and RGB-D cameras. The core functionalities are presented, including 3D spatial reconstruction, head pose estimation, 2D fixation classification, 2D to 3D fixation mapping, micro-scale fixation localisation. Other components of the framework are described in this chapter, such as the equipment employed, the coordinate frames registrations performed, the data acquisition method followed and the hardware agnostic software architecture implemented. Finally, the accuracy and performance of the framework are evaluated in a simulated surgical setup to demonstrate its usability in surgical applications.

Chapter 4 presents a novel 3D gaze-guided robotic scrub nurse. The platform is evaluated in simulated surgery to determine usability and acceptability with clinical teams. 10 teams of surgical trainees and trained scrub nurses performed an ex vivo task on pig colon. Surgeons used gaze via wearable eye-tracking glasses to select surgical instruments on a screen, in turn initiating the robot to deliver the desired instrument. Real-time gaze-screen interaction is based on the framework presented in chapter 2 using optical trackers for the head pose estimation. Comparison is done between human- and robot-assisted tasks showing no significant difference in overall task load of the surgeon. Quantitative and qualitative feedback is positive. There is no significant difference in task workflow (interruptions) or operative time.

Chapter 5 introduces a fully motorised gaze-controlled system for non-restricting, free-view flexible endoscopy. It is based on a robotised system, which allows hands-free control of the endoscopic view in an intuitive fashion, using the natural gaze of the user to steer the endoscope tip. Real-time gaze-screen interaction is based on the framework presented in chapter 2 using optical trackers for the head pose estimation. The feasibility and comparison against traditional hand control are assessed

among 8 experienced endoscopists and 8 novice users, in a simulated examination task of the upper gastrointestinal (GI) tract. The results show that gaze controlled endoscopy is a feasible concept. Novice users are significantly faster with gaze control, while expert endoscopist are significantly faster with hand control.

Chapter 6 proposes an assistive robotic system with an intuitive free-view gaze interface, which is designed for use outside the operating theatre; for people with motor disabilities. The user's point of regard is estimated in 3D space while allowing free head movement and is combined with object recognition and trajectory planning. Real-time 3D gaze estimation is based on the framework presented in chapter 2 using the Perspective-n-Point (PnP) approach for the head pose estimation. This system allows the user to interact with objects using fixations. Two operational modes are implemented to cater for different eventualities. The automatic mode performs a pre-defined task associated with a gaze-selected object, while the manual mode allows gaze control of the robot's end-effector position on the user's frame of reference. User studies report effortless operation in automatic mode. A manual pick and place task achieves a success rate of 100% on the users' first attempt.

Chapter 7 summarises the key contributions and results of this thesis and discusses current limitations and future potentials deriving from this work.

Chapter 2

Perceptually-Enabled, Smart Operating Room

2.1 Introduction

Improved surgical outcome and patient safety in the operating theatre is a constant challenge and has been extensively discussed in the medical literature [62]. Arguably, the most influential factors with a detrimental effect on these two areas are related to suboptimal communication among the staff, poor flow of information, staff workload and fatigue, ergonomics and the sterility of the operating theatre [62–66]. The integration of new technologies in the OR has played a significant role in these factors; it has been reported that 36% of communication failures are related to equipment use [67]. However, while new technologies may add complexity to the surgical workflow, at the same time they provide new opportunities for the design of systems and approaches that can enhance patient safety and improve workflow and efficiency [62]. A number of initiatives have assessed the state-of-the-art in technological developments and identified key areas where future innovative solutions could be used to optimise the operating environment [39, 40, 62, 68–70]. This chapter presents an overview of the safety risks which lurk in the surgical procedure first. Then, an overview of touchless interaction modalities with emphasisf

on gaze-contingent systems and applications is presented. Finally, key aspects of the future OR that have been identified in the literature are reviewed. The aim is to emphasise the value and potential of the proposed gaze-contingent interaction modality in the context of surgery, and the perceptually enabled data it can contribute to the recently emerged field of Surgical Data Science towards Surgery 4.0.

2.2 Patient safety risk factors in the OR

2.2.1 Suboptimal communication

Velasquez et al. [71] analogises the surgical procedure to a concert by a philharmonic orchestra. They identify the relationship of hundreds of iterations of the same routine by the OR team to the countless rehearsals and concerts of the same piece performed by the musicians [71]. Both result to further knowledge and expertise acquisition [71]. The surgical team distinguish signs and phases in the operation that enable them to anticipate the following step [71]. Precise timings of actions compliant with the safety protocol are decided by the surgeon, as an orchestra director [71].

Lingard et al. [72] investigated the nature of communications among OR team members from surgery, nursing, and anaesthesia. The purpose was to find communication patterns and sites of tension in order observe their effect on novices [72]. Their results revealed a variety of communicative events, from jokes and social chats to commands and silences, which were clustered into prominent themes: time, resources, roles and relationships, safety and sterility, and situation control [72]. These themes resulted in communicative tension, affecting predominantly the novices, who reacted by either mimicking the senior surgeon or withdrawing from the communication, with negative consequences for the relationships of team members [72].

Moss et al. [73] explored the communication patterns in the OR, as a

prerequisite for designing technological applications which will enhance OR coordination and patient safety [73]. The majority of communications was found to have occurred face to face, with equipment being the most common purpose amongst them [73]. They conclude that trivial tasks such as patients' preparation for surgery and surgical equipment management, occupy a significant portion of communications among the staff [73]. Therefore, by automating them, information exchange can be reduced, in turn minimising the chances of workflow interruptions and adverse events in the OR [73].

2.2.2 Communication failures

The operating theatre is reportedly the environment where unintentional patient harm is most likely to happen [62, 74, 75]. Human error is natural to occur [76, 77], however there are factors which increase the risk. Some of the most influential factors are related to suboptimal communication among the staff, poor flow of information, staff workload and fatigue in the operating theatre [62, 78].

Ineffective communication and teamwork in the OR is a common source of errors leading to unintentional patient harm and is often irrelevant to individual clinical skills [79–81].

An analysis of the characteristics of communication failures in the OR, by recording and analysing 90 hours of 48 surgical procedures, highlighted the key effects of the communication errors as inefficiency, tension, delay, workaround, resource waste, patient inconvenience and procedural error [82]. They found that 30.6% of all team exchanges in the operating room are classified as failures and one third of them resulted to immediate effects that imperiled patient safety [82].

The roots of such failures can be also found in equipment use [67], team instability and lack of coordination (i.e. inexperienced surgeons, different scrub nurse for each case) [83] and diverse linguistic and cultural backgrounds [84]. Nonetheless, the complexity of medical care leads even experienced individuals to commit errors [79].

2.2.3 Ineffective collaboration

Individual clinical skills are not sufficient to achieve a safe, error-free surgical operation [85]. Communication breakdowns have been reported as key sources of wrong-site operations and other sentinel events [86]. Effective communication and teamwork are essential components for a successful and safe operation [86, 87]. However, perception of effective teamwork has been assessed with various perspectives [86]; in relation to patients' death rates in intensive care units [88], nurse turnover in the OR [89], error reduction in aviation [90] and job satisfaction [86, 91]. To this end, nurses' high rates of dissatisfaction are related to insufficient teamwork and ultimately to shortages in nursing personnel, becoming a vicious circle for the quality of patient care and patient outcomes [86, 92].

Makary et al. in [86] attempted to measure teamwork in the surgical setting. They found significant divergences in perceptions of teamwork in the OR; surgeons and anaesthesiologists appear more satisfied with the collaboration than the nurses [86]. This difference in teamwork perception may originate by authority, gender, training and patient-care responsibilities or by different perspective on the definition of effective collaboration; nurses often associate it with the physician respecting their input, whereas physicians with anticipation of the next step by the nurse [86].

Ineffective collaboration and communication between physicians and nurses may have negative impact on patient outcomes [93, 94]. More recently, Matthys et al. [94] conducted a systematic review to investigate the impact of collaboration between physicians and nurses on patient outcomes. They reported patient outcomes related to blood pressure, satisfaction and hospitalisation to improve with physicians-nurses collaboration, whereas colorectal screening, hospital length of stay and health-related quality of life are outcomes that appeared neutral to this collaboration [94].

Staplers et al. [95] highlights the importance of open communication between physicians and nurses for effective collaboration. Other collabo-

ration dependencies reported, include trust, respect, shared leadership, recognition of unique contribution and collegiality [95, 96]. Deficiencies in collaboration seem to be related to time pressure, unclear roles, lack of organisational support, poor leadership, different standards and professional values, varying aims and priorities and vertical management structures [94, 97–100].

Several studies have focused on how collaboration in medical tasks is affected by spatial organisation of the OR and information systems [85, 94, 101, 102]. The positioning of artefacts (i.e. schedule whiteboard) affects the attention of the surgical staff and the formulation of shared common spaces, leading to better information flow and improved collaboration [85, 103].

2.2.4 Poor ergonomics

The term "ergonomics" has been described as "the concept of designing the working environment to fit the worker, instead of forcing the worker to fit the working environment" [104]. Seagull et al. [39] identified ergonomics and human factor as one of the four pillars for a smart, safe operating room of the future.

Palmer et al. [105] studied flow disruptions in the cardiac OR and they found that one third of the overall disturbances were caused by the physical layout of the room [105].

Scupelli et al. [103] used the concept of the "information hotspots" to investigate how the physical layout affects coordination. They found that positioning artefacts (i.e. schedule whiteboard) in information hotspots, facilitated the formulation of common information spaces and improved information flow [85, 103].

Although the physical layout shares similar principles between open surgery and MIS, the additional equipment required for the latter, adds complexity to the spatial arrangement of the OR [106]. This leads to increased workflow disruptions [106], while in the long term it may also cause work-related musculoskeletal disorders (WMSDs) due to suboptimal body posture [107].

Matern et al. [108] explored the workplace conditions in the OR in German hospitals. 97% of the surgeons agreed that enhancing ergonomic factors is necessary; working posture, difficulties using the OR lights and handling of devices are among these factors [108]. Most importantly, they observed that the effects of poor ergonomics go beyond workflow disruptions to safety risks for the personnel and the patients [108].

Compared to open surgery, MIS has been found to increase the risk for the surgeon to suffer from neck pain, upper and lower extremity pain, numbness, and fatigue [109]. A common cause is the suboptimal physical layout (i.e. positioning of display monitors, table height) [110]. Of particular interest is the fact that surgeons' chronic health issues are not the only effects of their suboptimal postures in the OR. In some cases (30%), surgeons would consider their WMSDs while recommending a surgical approach to their patients [109]. Additionally, WMSDs appear to be the cause for frequent (sick) leaves amongst surgeons [111].

Seagull [110] identifies the reasons why widely adopted standards in ergonomics have not been implemented in surgery, in four points: the workspace relies on human anatomy (organic workspace), each patient has unique anatomy, time pressure and strict regulatory requirements for instruments design.

2.2.5 Sterility

Maintaining a sterile surgical environment (area, people, devices) is an important and challenging ritual [112] and has major effects in communication patterns [72]. Usually, OR attendants in the aseptic area are not allowed to leave it and come in direct contact with medical equipment, to avoid contamination [85]. Therefore, most of the medical devices (i.e. monitors, computers) are positioned outside the aseptic area and OR members need to be available to assist upon request [85].

In some cases, OR members need to access a workstation outside the aseptic zone [113, 114]. For example, radiologists may move from a sterile to a non-sterile zone to view and manipulate the radiation screen

themselves [113]. While in a non-sterile zone, they remove their gloves to handle the equipment and then re-scrub to return in the aseptic zone [113]. However, this comes with significant cost in time. To deal with time constraints, they may use their gown between their sterile gloved hands and the equipment (i.e. computer mouse) [113]. Both strategies jeopardise patient safety [113].

Consequently, sterility restrictions have a significant effect on the interaction with medical technologies [85]. Therefore, touchless interaction with equipment and machinery will play a key role in the future sterile operating theatres.

2.3 Touchless interaction in the OR

Technologies in the operating theatre are an integral component of the surgical workflow. Each individual surgical procedure relies on different ensembles of medical devices. The advent of MIS has transformed the OR setup in such way, that various imaging devices are necessary to assist the surgical team pre- and intra-operatively (i.e. endoscopic cameras monitors) [85]. Moreover, access to imaging data, such as Computer Tomography (CT), Magnetic Resonance Imagery (MRI) etc., is a frequent requisiteness by the surgeons during the operation. Therefore, the surgical team is highly depended on these medical technologies and interaction with them is fundamental for a seamless and safe operation.

Traditionally, the interaction with medical devices is being performed with keyboard, mouse, joysticks, touchscreens or control panels. However, these touch-based interaction modalities lurk the risk of contamination [115]. Wipeable surfaces or sterile covers may enable direct use of interaction devices, but this would jeopardise sterility in the OR when used by non-sterile staff [85].

A number of solutions have been adopted to overcome these limitations. A common practice by the surgeons is to ask for the help of other surgical team members outside the sterile area [113, 116]. However, this can

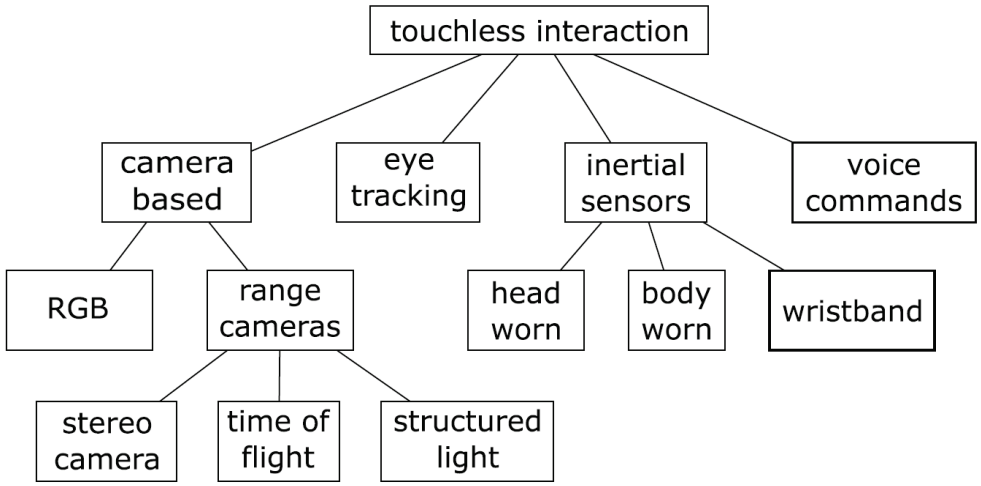


Figure 2.1: Overview of touchless interaction methods and devices as presented in [2], reproduced with permission from Springer.

be a slow and ineffective task [117], as the surgeons need to instruct personnel who are often unfamiliar with the interface [85]. Additionally, the task context is usually highly dependent on clinical knowledge and interpretation [113, 118]. Therefore, surgeons occasionally walk away from the patient [119], raising safety concerns due to sterility, workflow disruptions and overall efficiency [118].

Touchless interfaces have been developed to address the restrictions which traditional interaction means engender [2, 113, 118, 120]. Recently, Mewes et al. [2] did a systematic literature review on systems that focus on touchless human-computer interaction in ORs and interventional radiology suites. They classify the results based on the application (control of medical image viewers, laparoscopic assistance, telerobotic assistance, OR control, robotic OR assistance and intraoperative registration) and the sensors used (Fig. 2.1) [2]. Here, an overview of the systems is presented by the means of interaction (Table 2.1). We assume touchless interaction when media that can cause cross infection (i.e. hands) have no physical contact with non-sterile areas (i.e. computer peripherals).

Table 2.1: Interaction modalities with medical technologies, based on whether means that can cause cross infection (i.e. hands), have physical contact with non-sterile areas (i.e. computer peripherals).

Touch-based interaction	Touchless interaction
Manual/hand control	Foot pedal
Keyboard	Vocal commands
Mouse	Hand gestures
Joystick	Electromyography (EMG)
Touchscreen	Electroencephalography (EEG)
Control panel	Body posture
	Head pose
	Finger tracking
	Gaze-contingent commands

2.3.1 Foot pedal

Food pedal is a commonly used interface in the OR. The da Vinci® (Intuitive Surgical, Inc.) surgical robot is using foot pedals to allow the surgeons to switch modalities between camera and instrument control, decouple the master from instruments control, swap instrument arms and perform electrosurgical tasks.

The AESOP® robot is designed to hold and move the laparoscope and is controlled through foot pedals (Fig. 2.2(a-c)). Hand and voice control interfaces are also integrated in the system (Fig. 2.2(c-d)). Allaf et al. [5] compared the foot and voice control modalities of AESOP® (Fig. 2.2(d)) and found that foot pedal control is faster and less disruptive than voice, despite the ergonomic superiority of the latter.

Veelen et al. [121] conducted a study on the ergonomic acceptability of foot pedals among 45 laparoscopic surgeons. The results showed that

only 16% of them are satisfied with the current design of foot pedals and reported incidents such as pushing the wrong switch (75%) [121]. Based on the guidelines arisen from the study, they designed an ergonomically improved foot pedal.

2.3.2 Voice control

The operating theatre is an environment with significant traffic, communications and noise. Therefore, capturing and interpreting vocal signals is a fairly challenging process [2]. However, it is a modality which allows surgeons natural interaction with medical equipment, without the need to look away from the surgical field [5]. Nevertheless, other modalities such as foot pedals have been reported to perform faster [5].

Nathan et al. [3] developed a voice-controlled robotic interface to hold and position the endoscope. They evaluated the system with 10 cadaver heads and found no significant difference in overall duration of an endonasal task [3]. However, the most significant contribution of such systems is the need of only one surgeon to perform the task. In a similar interface [122], voice control is allegedly more effective and precise than humans in complex tasks.

Carpintero et al. [6] developed a robotic scrub nurse system that delivers surgical instruments from one instrument tray to another (Fig. 2.2(e)). The instruments are detected by a visual recognition module, which is triggered by a human scrub nurse using vocal commands. The visual recognition module was evaluated over 7 instruments with 98.1% accuracy and the voice recognition module over 20 commands with 93.5% success rate [6].

Perrakis et al. [123] compared two voice recognition systems included in commercial integrated ORs (Siemens Integrated OR System SIOS and Karl Storz OR1 NEO™) [2]. Evaluating through various features the systems provide (e.g. OR table adjustment, gas pressure control, video controller activation), they found that although SIOS vocal control was significantly faster and more reliable than OR1, manual control was faster

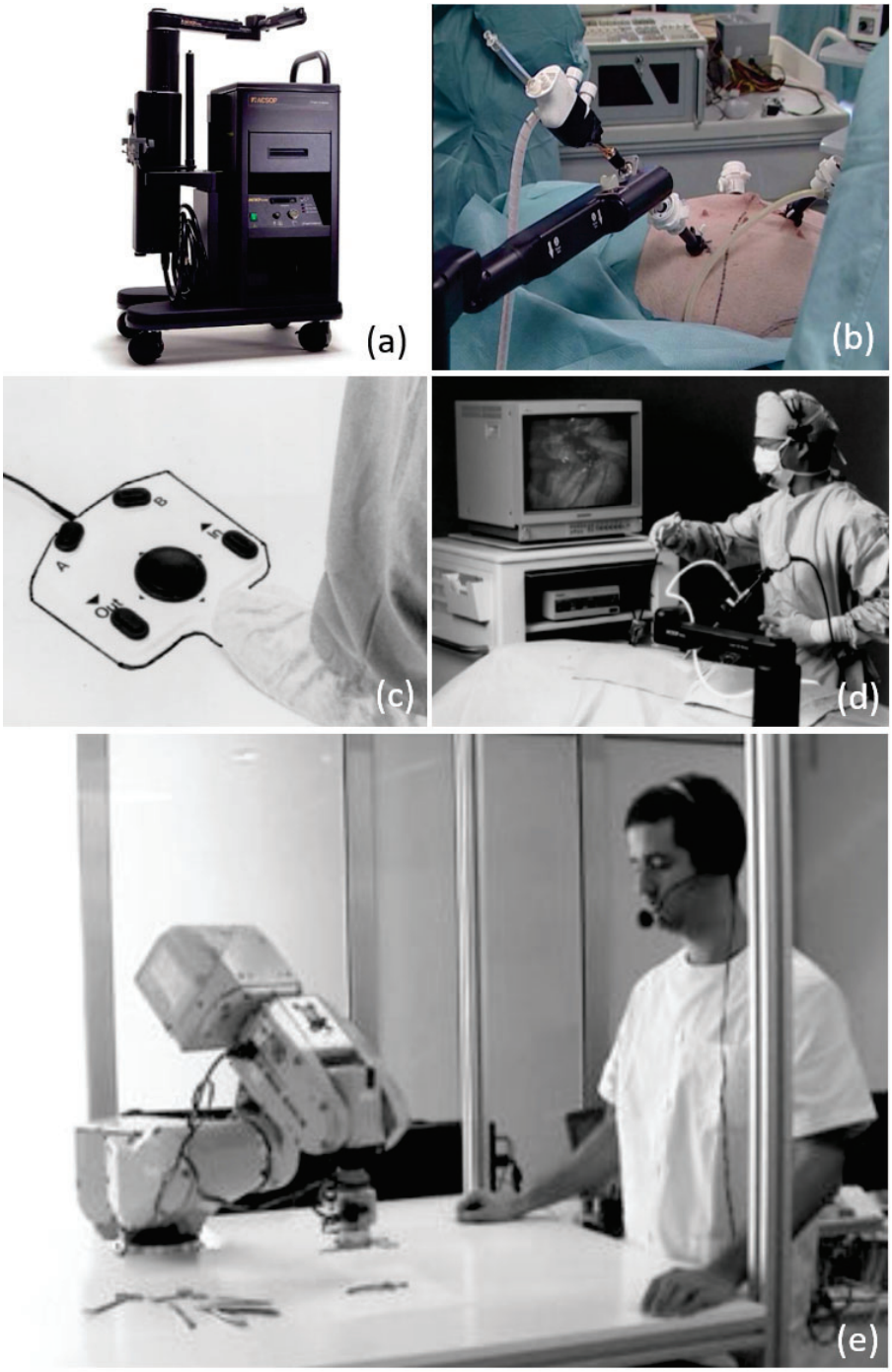


Figure 2.2: (a) The AESOP® laparoscope holder, ©Georg Thieme Verlag KG. reproduced with permission [3]. (b) AESOP® in the OR, reproduced with permission from Springer, adapted from [4]. (c) Foot pedal interaction with AESOP®, reproduced with permission from Springer, adapted from [5]. (d) Voice control of AESOP®, reproduced with permission from Springer, adapted from [5]. (e) Voice controlled robotic scrub nurse, ©2010 IEEE, adapted from [6].

than voice control in both systems. Moreover, the authors highlight the fact that despite SIOS producing no errors in this study, such systems have not been integrated to surgical workflow due to various flaws, such as poor interface design and complicated voice commands [123].

2.3.3 Body movement gestures

A large amount of literature has studied the feasibility of sensors which track human body movements to control medical technologies in the OR. The mass production of Microsoft Kinect (Fig. 2.3(a)) as a gaming console and its low cost subsequently, revealed the opportunity to the research community to exploit the ensemble of sensors it was provided with (structured light depth sensor, RGB camera, voice input, integrated skeleton tracking) to develop applications further than gaming [2, 124].

Hand gestures are commonly used for interaction with medical technologies (Fig. 2.4(b,c,e,f)). Wachs et al. [8] used Kinect to develop a medical image navigation interface (Fig. 2.4(b)). With a similar hand-gesture detection/recognition approach, Collumeau et al. [125] developed a remote interface to control a virtual surgical lighting arm . However, gesture-based interfaces are prone to accidental triggered routines due to continuous tracking by the sensor [7]. The system developed in [7] (Fig. 2.4(a)) is comprised of 10 gestures and takes into account the body orientation to detect the intention of the user to manipulate the display. The detection rate is 92.26% and reliability 89.97%, with 99.55% true positive and 1.3% false positive rate on intention recognition [7].

Imaging systems are frequently used in the OR to plan needle insertion or for intraoperative registration of anatomical images. Gong et al. [126] used Kinect skeleton and depth data to classify hand gestures and align a 3D model to X-ray images. Wen et al. [11] used a Kinect to control a surgical robot for needle insertion based on hand gestures, while an augmented reality projection on the patient provided needle guidance, with less than 2 mm error (Fig. 2.4(e-f)). Herniczek et al. [127] placed an orientation sensor under sterile glove to guide needle insertion during

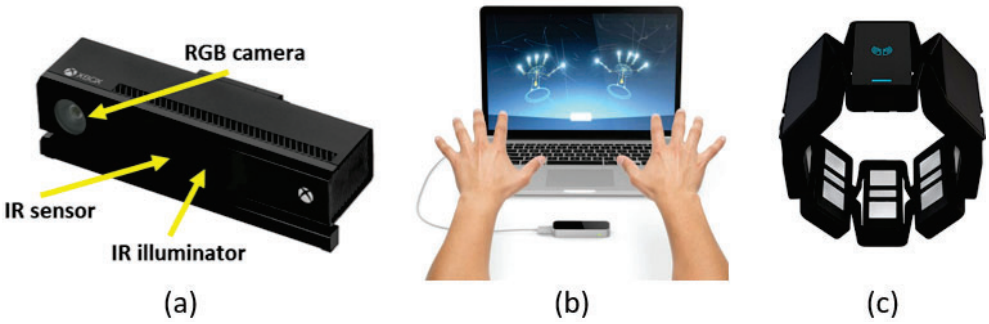


Figure 2.3: Sensors used for body movement gestures detection. (a) Microsoft Kinect v2 RGB-D camera[1]. (b) Leap Motion™ controller[2]. (c) Myo Gesture Control Armband (Thalmic Labs)[3].

ultrasound-guided nephrostomy, based on a set of 4 gestures.

In contrast to other application specific interfaces, Graetzel et al. [117] used a stereo camera to develop a universal interface, which maps hand gestures to standard computer mouse functionalities (pointer movement, click). Although clicking was robust, the mouse pointer often jittered and the system was not robust to rapid gestures, causing confusion to the users [117].

Leap Motion™ Controller (Fig. 2.3(b)) is a hand gesture sensor, which relies on a stereo camera and infrared emitters, that facilitate the hand segmentation when it is positioned over the sensor [2]. It has been used frequently for interfaces that allow the OR team to interact with medical image visualisation environments. Rosa and Elizondo [13] used Leap Motion™ Controller to navigate through dental images intra-operatively with hand and finger gestures (Fig. 2.6(b)).

Nishikawa et al. [128] developed a camera-based system, that allows laparoscopic camera control. It relies on the user's head gestures (head movements), derived by face features tracking [128]. They reported high accuracy, but fatigue in the users' neck was indicated. A similar interface

[1]Source of original picture: https://commons.wikimedia.org/wiki/File:Xbox-One-Kinect.jpg
[2]Source: https://leapmotion.com
[3]Source: MYO

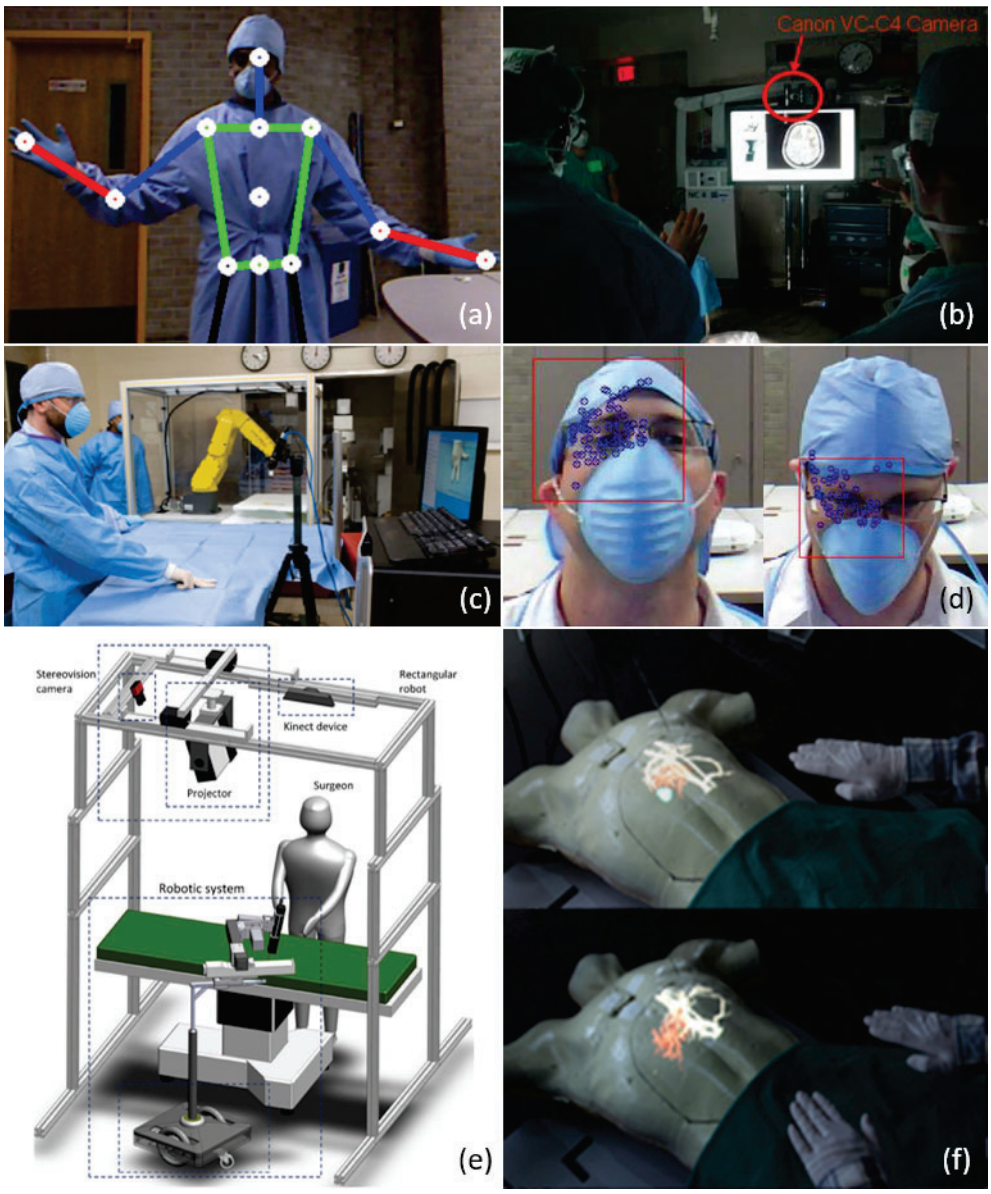


Figure 2.4: (a) Using Microsoft Kinect skeleton tracking to assist hand gesture-driven image navigation in the OR, reproduced with permission from Elsevier, adapted from [7]. (b) Hand gestures for browsing medical images, ©2007 TSI® Press, adapted from [8] and (c) controlling a robotic scrub nurse, ©2012 IEEE, adapted from [9]. (d) Facial orientation to control endoscopic views of tissue depth, ©2010 IEEE, adapted from [10]. (e,f) Hand gesture guided needle guidance system with augmented reality projection on the patient, reproduced with permission from Elsevier, adapted from [11].

Figure 2.5: Example of hand gestures lexicon using the Myo armband, ©2015 The Eurographics Association, adapted from [12].

was developed by Wachs et al. [10] (Fig. 2.4(d)), but in this case the camera movement was triggered when the head pose exceeded an angular threshold. Evaluation with 4 users in simulated larynx biopsy showed the system was faster to learn than keyboard, but slower overall [10].

Myo armband (Fig. 2.3(c)) is a commercial sensor which uses 8 electromyographic (EMG) sensors to sense electrical signals from the forearm muscles [2]. Hettig et al. [12] used Myo to navigate through medical images (Fig. 2.5, 2.6(a)). Results from two user studies and one clinical test revealed positive feedback by users, but also robustness issues [12], which make this sensor insufficient for applications in clinical environments [2].

Johnson et al. [113] explored the implications of touchless interfaces in for the ORs. A common issue, which has already been mentioned in this section, is the unintentional activation of the system due to permanent gesture-tracking by the sensors [113]. Other concerns include the spatial arrangement in such a cluttered and dynamic environment, cognitive demands and the design of a user-friendly and efficient gestural vocabulary [113].

2.3.4 Gaze-contingent interfaces

Ergonomic limitations and poor efficacy of the aforementioned modalities for touchless interactions in the OR, hinder their adoption in the surgical workflow [129].

While foot pedals, body gesture- and voice-controlled interfaces require

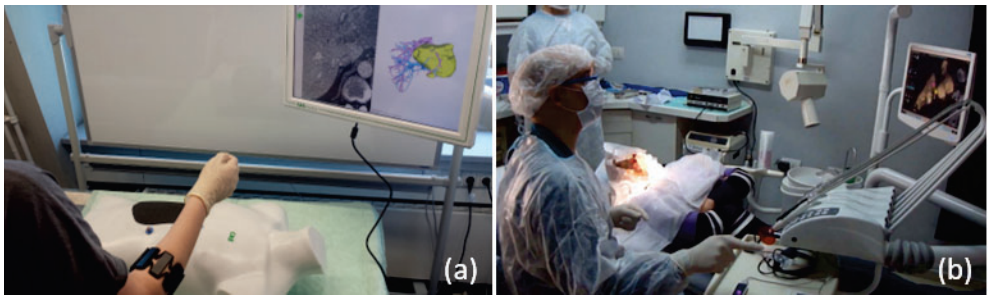


Figure 2.6: (a) Using the Myo armband for exploration of 3D medical image data, ©2015 The Eurographics Association, adapted from [12]. (b) Image navigation in dental surgery with the Leap Motion™ controller, licensed under CC BY, adapted from [13].

explicit actions to prompt human-computer interaction (feet / hand / body motions, vocal commands), gaze-contingent interfaces allow natural eye behaviour. To this end, *eye-tracking* can be used to measure user's ocular movement and point of gaze. By approaching the eyes as the only visible part of the brain, we can consider them not just in the conventional sense as receptors of visual stimuli, but also as perception- and cognition-rich actors within the operating theatre [62]. This approach could allow harnessing the power of the underlying mental processes that lead from perception to cognition to action, and seamlessly endow intelligence to implemented human-computer interfaces [62]. However, gaze-based interaction systems require careful design, as the distinction between unintentional looking and intentional gaze commands can be challenging due to natural eye behaviour, known as the Midas touch problem [130]. A number of solutions have been proposed to deal with this problem [131,132], such as fixations' dwell time [133,134] and intention recognition through machine learning techniques [23, 135–138].

Although tracking the eye position and movement appeared in the literature in the previous centuries, the technological advancements of the latest decades have changed significantly the way eye movements are observed [139, 140]. An eye-tracking technique uses contact lenses with mirrors or magnetic search coils, in order to measure electromagnetic

Figure 2.7: Common eye-tracking techniques. (a) Contact lenses with magnetic search coils, left: republished with permission from the authors, from [14], right: ©Chronos Vision [15]. (b) Electro-oculography (EOG), ©2011 IEEE, adapted from [16]. (c) Video-oculography (VOG) with remote eye-tracker, left: ©Tobii Technology [17], right: ©The Eye Tribe [18]. (d) Eye-tracking glasses, ©SensoMotoric Instruments (SMI) [19].

variations caused by the user's eye movements [141] (Fig. 2.7(a)). This is intrusive, with limited time span, but is also very accurate [141, 142]. Another method is electro-oculography (EOG), which uses electrodes to measure the electric potential difference of the skin around the eye, which varies as the eye rotates [143] (Fig. 2.7(b)). It is cumbersome, uncomfortable and not suitable for mobile applications; therefore it is used widely for ophthalmological studies [144]. A more comfortable and less intrusive eye-tracking technique is video-oculography (VOG) (Fig. 2.7(c-d)). VOG relies on recording eye movements with digital cameras and determining the eye positions and movements by processing the recorded eye images. In particular, usually IR light sources are employed to produce reflections on the boundaries of the lens and cornea [145]. By processing the resulting images (Purkinje images), the pupil and the corneal reflection (glint) are identified, which in turn provide the gaze direction [145].

To analyse eye gaze behavioural patterns and correlate gaze direction to visual stimuli, a reference plane is necessary. Eye movements need to be calibrated on these physical or virtual planes to obtain the user's accurate 2D point of regard on the specific plane. The most common configurations are remote eye-trackers (Fig. 2.7(c)) and head-mounted/wearable eye-tracking glasses (Fig. 2.7(d)). In the first setup, an eye-tracker is mounted on a screen, providing accurate 2D gaze information on the screen plane. However, this setup restricts the user's head to a fixed position and the workspace to the screen where the eye-tracker is calibrated to. Wearable eye-trackers provide 2D gaze information on the user's head frame of reference without any restriction on the user's movement. However, gaze information on fixed planes, such as monitors, cannot be retrieved directly from a wearable eye-tracking setup.

Duchowski in [146] classifies eye-tracking applications into two categories: diagnostic and interactive. In diagnostic applications, the user's visual attention is recorded to analyse behavioural patterns on a given stimulus or to assess interfaces [146–149]. In interactive applications, the

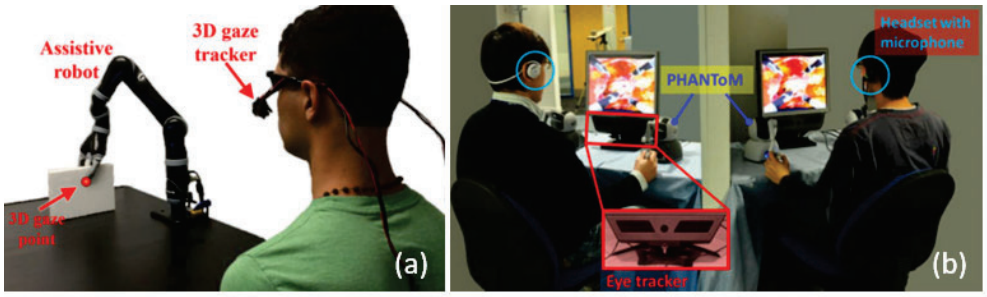


Figure 2.8: (a) Eye-gaze based system for Activities of Daily Living (ADL), ©2017 IEEE, adapted from [20]. (b) Collaborative eye-tracking paradigm during robotic assisted surgery, reproduced from Springer open access, adapted from [21].

gaze direction provided by the eye-tracker serves as a control modality (i.e. computer mouse pointer, robot control) [23, 62, 147, 150].

As a diagnostic tool, eye-tracking has been used in various areas, to improve driving safety [43], convey perceptual skills [44], evaluate cognitive activity [45, 151], improve efficiency in collaborative tasks through gaze awareness [152], as well as monitor the situation awareness by understanding the behaviour of expert and novice pilots [46]. In clinical settings, eye-tracking has been successfully used to enhance collaboration by sharing gaze information between supervisor and trainee [21, 49] (Fig. 2.8(b)) and to distinguish between novice and expert surgeons [50, 51]. Eye-tracking has also been used in objective measurement of surgical skills [52], for enhancing training skills [53] and for revealing opportunities to improve performance [54] by facilitating surgical skill acquisition.

As an interaction means, eye-tracking has been an alternative to hand control, mostly when hands are not available [85, 153–156]. In the healthcare domain, it has been used as an assistive solution for disabled people [20, 150, 157–160] (Fig. 2.8(a)). For Activities of Daily Living (ADL) eye-tracking can be used to provide valuable perceptual information and eventually improve the quality of the robotic assistance. For example, Li et al. [20] use ocular vergence to determine the 3D point of regard, followed by neural networks to improve the accuracy of gaze mapping

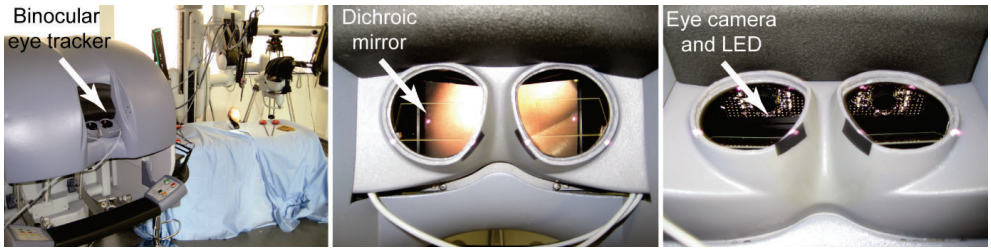


Figure 2.9: Eye-tracking integration with the da Vinci® surgical robot, reproduced with permission from Springer, adapted from [22].

(Fig. 2.8(a)). Using 3D gaze the user can define the contour of a target object to be grasped by the robot. However, the lack of a world frame of reference restricts the capabilities of the system to predefined and calibrated spaces. Moreover, a long calibration procedure is required and a head stand to prohibit head movement [150].

Within the operating theatre, eye-tracking has been used to enhance collaboration in laparoscopic and robotic setups by sharing the visual attention of multiple collaborators [21, 49] (Fig. 2.8(b)). Although robotic surgery introduced more precise and less invasive operations, the control interfaces are not equally ergonomic [2]. Therefore, eye-trackers have been integrated within the console of robotic systems, such as the da Vinci® [22, 145, 161–166] (Fig. 2.9). In such settings, tissue surface reconstruction is used to enable dynamic active constraints, motion stabilisation and image guidance. To this end, binocular eye-trackers have been integrated to surgical consoles to estimate the 3D point of regard and allow motion stabilisation [145]. Stoyanov et al. [22] used binocular eye-tracking and nonparametric clustering to optimise ablation paths on a phantom heart model (Fig. 2.9). Visentini-Scarzanella et al. [165] used the same approach to 3D reconstruct a deformable silicon heart phantom. Tong et al. [161] used 3D fixations to guide users through haptic feedback. Similarly, Noonan et al. [163] used the 3D fixations as commands to guide a robotic probe to the intended locations. Mylonas et al. [164] introduced fixation-guided virtual constraints through haptic feedback, in order to prevent the surgical instruments from unwanted

movements. Clancy et al. [166] developed a gaze-controlled autofocus system for the da Vinci® endoscope, using an eye-tracker and liquid lens. This approach was faster, more ergonomic and natural comparing to the default foot-pedal-based mechanical focus system [166]. Lastly, Li et al. [162] proposed a gaze-controlled ultrasound interface with a novel head motion compensation algorithm.

A common application of gaze-contingent interfaces in surgery is the control of the laparoscopic cameras (Fig. 2.10). Frequent adjustments of the camera occur in the surgical workflow due to the limited field of view which they provide [23]. Therefore, an assistant is usually requested to manoeuvre the camera under the guidance of the surgeon [23]. However, communication failures between the surgeon and the assistant are frequent and may lead even to patient harm, as discussed in previous section. Thus, gaze-based interaction with the laparoscopic screen has been proposed in the literature. To achieve this, gaze-contingent closed-loop controllers are used to control the laparoscopic camera, which rely on the distance of the user's visual attention (fixation point) to the centre of the laparoscopic screen (visual feedback).

Such a closed-loop controller is used in [24]. Gaze commands are generated to control a single joint of a 5 DOF mechatronic laparoscope, either by selecting which joint to activate or through automatic selection [24] (Fig. 2.10(b)). In [167], the laparoscopic camera held by the AESOP® medical robot is controlled based on eye-tracking data. Fujii et al. [168] used gaze gestures to control a laparoscope mounted on an articulated robotic arm with a velocity controller. Staub et al. [25] used wearable eye-tracking to control endoscope positioning (Fig. 2.10(c)). Custom eye-tracking glasses and head tracking are combined to estimate the point of regard on the screen, using a stereo display with IR LEDs and a wide angle camera on the glasses [25]. Finally, the TransEnterix Senhance [169] is a commercial robotic system to control the laparoscope camera based on eye-gaze.

However, current gaze-contingent laparoscope camera control solutions

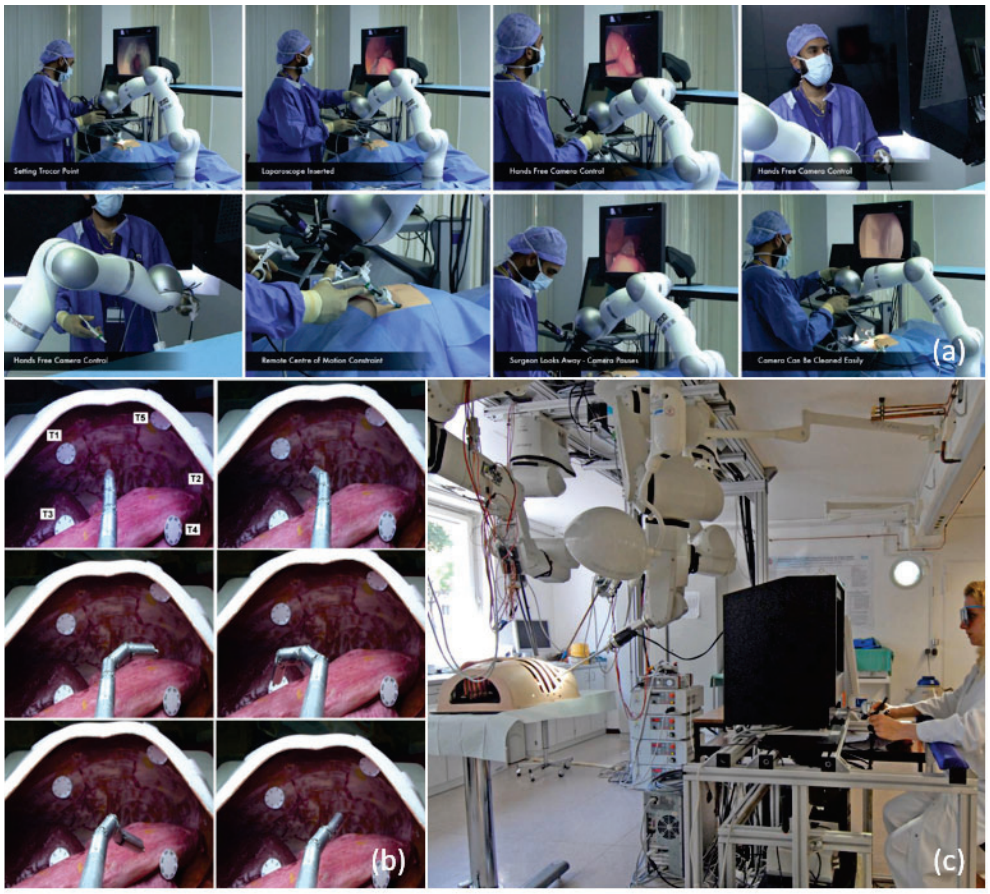


Figure 2.10: (a) Gaze gesture based control of a laparoscopic camera, mounted on a robotic arm, licensed under CC BY, adapted from [23]. (b) Gaze-control of a mechatronic laparoscope, ©2010 IEEE, adapted from [24]. (c) The *ARAMIS* system for gaze-control of robotic endoscope or surgical tool, using eye-tracking glasses and a stereo display, ©2012 IEEE, adapted from [25].

Figure 2.11: (a) The *GazeTap* system using gaze and feet as input channels to allow hand-free interaction with medical equipment in the OR, republished with permission from the authors [26]. (b) A wearable eye-tracking system designed for the scrub nurse, reproduced with permission from Springer, adapted from [27].

restrict the surgeon from roaming freely in the operating theatre and interacting with multiple medical devices in a hand-free fashion.

The constant integration of new technologies in the OR engenders the need of ergonomic and safe human-computer interfaces in the sterile environment. Unger et al. [170] described the design and evaluation of a wearable eye-tracking system designed for the scrub nurse with features to relieve the surgeon (Fig. 2.11(b)). They investigated three use cases using eye gaze for interaction in the OR: making a video call, labelling surgical instruments and changing the light conditions in the OR. Each modality is triggered when the user is fixating on the corresponding barcode markers, which were printed and placed in relevant locations in the OR. Hatcher et al. [26] combined gaze and feet as input channels to allow hands-free interaction with medical equipment in the OR (Fig. 2.11(a)). The gaze on screen is estimated using wearable eye-tracking glasses and fiducial markers on the display. They evaluate the system on an image selection and manipulation task. However, both systems lack clinical validation to investigate their impact and robustness on the surgical workflow.

2.4 Operating room of the future

The operating theatre is reportedly the environment where unintentional patient harm is most likely to happen [62,74]. Some of the most influential factors are related to *suboptimal communication among the staff, poor flow of information, staff workload and fatigue and the sterility of the operating theatre* [63]. While new technologies may add complexity to the surgical workflow, at the same time they provide new opportunities for the design of systems and approaches that can enhance patient safety and improve workflow and efficiency [62]. A number of initiatives have assessed the state-of-the-art in technological developments and identified key areas where future innovative solutions could be used to optimise the operating environment, such as *cognitive simulation, informatics, "smart" imaging, "smart" environments, ergonomics/human factors and group-based communication technologies* [39].

In the spirit of the Internet of Things (IoT) and the recent explosion of data-driven sciences, it is anticipated that equipment, surgical instruments, consumables and staff will be fully integrated and networked within a "smart" operating suite [62]. This could happen in a number of ways, such as electronically, using computer vision, RFID markers or other technologies [61, 171]. Partially integrated operating suites are already being provided by companies, such as the Karl Storz's OR1 NEO™ [41], where components of the surgical environment (e.g., endoscopic devices, video/data sources, surgical table, ceiling lights) can be tailored to and by the user and can be controlled from a central location within the sterile area (Fig. 2.12) [62]. Such operating suites, where a large amount of information can be made available through a unique integrated system, offer tremendous opportunities for implementing novel human-computer interfaces, context-aware systems, automated procedures and augmented visualisation features [62].

Moreover, a significant body of research has explored "perceptually enabled" interactions in the sterile environment using technologies like 3D

Figure 2.12: Integrated operating suit. KARL STORZ OR1 NEO®, ©KARL STORZ - Endoskope, Germany, reproduced with permission [28]

cameras, voice commands or eye-tracking [2]. This way the surgeon can be kept in the loop of decision-making and task-execution in a seamless way that is likely to help improving overall operational performance and reducing communication errors [62].

2.4.1 OR of the future in literature

Back in 2003, Rattner et al. [68] explored the state of the art technology at the time and identified the desired technology in the OR of the future: smart instruments; image-guided augmentation; collection, analysis and intelligent display and storage of data; algorithms to extract relevant information; development of a plug-and-play environment [68]. They highlight the prospects of a "new technique called machine learning to try to pull information out of complex data" [68]. Trends and patterns, which would allow to predict adverse events, could be identified by applying machine learning to data from multiple surgical operations [68]. Moreover, the new machinery introduced with the advent of MIS could not reveal its full potential without their communication through common interfaces and their integration in a network for control, data capture and safety [68].

Cleary et al. [69] identified the clinical and technical requirements of an effective and vital OR of the future in 5 areas: operational efficiency and workflow; systems integration and technical standards; telecollaboration; surgical robotics; intraoperative diagnosis and imaging; and surgical informatics [69]. They refer to the need of standardised interfaces among devices, as well as a plug-and-play system as a reference point for communication and control of multiple devices [69]. The nascent area of surgical informatics is also highlighted, as the patient data collection, storage, retrieval, sharing and rendering, which could provide the surgeon with valuable decision support systems intra-operatively [69].

Seagull et al. [39] explored the smart OR of the future focusing on areas which affect patient safety and operation efficiency. They discuss about key aspects of the future OR, such as cognitive simulation, informatics, "smart" imaging, "smart" environments, ergonomics/human factors, operational glitch analysis and group-based communication technologies [39]. The contributions of these areas are classified in 4 "pillars": surgical simulation, smart image, informatics and ergonomics/human factors [39]. Information systems can be benefited from integrating human factors to them, while enhanced ergonomics in the OR can minimise the risk of infection, shorten operation duration and reduce surgeon and staff fatigue [39]. By fusing multi-modal data (pre-/intra-/post-operative, from other processes, etc.) efficiency and safety patterns can be extracted, which would allow to backtrace, understand or even predict adverse events [39]. To this end, the significance of perceptual data in healthcare is signified: they can optimise information systems, in turn minimising medical error, improving efficiency, minimising risk to patients and caregivers, and reducing costs [39].

More recently, Bharathan et al. [40] discussed about the OR of the future where patient remains the focus. As the modern surgical care is highly dependent on safety, efficacy and cost effectiveness, this future OR concept has adopted practices from the aviation and petroleum industries to enhance patient safety, patient and staff satisfaction and minimise costs

[40]. Areas of innovations for the future OR include ergonomics, imaging, navigation, medical informatics, training and simulation [40]. The authors also emphasise the importance of reliable data management for the modern surgical care in the information age, where an integrated OR is essential for managing the resources [40]. Eye-tracking is mentioned as a source of invaluable perceptual information which can lead to surgical workflow predictions, such as errors [40]. To achieve this, standard behavioural patterns can be formulated by collecting and analysing surgeon- and procedure-specific data from several surgical procedures [40]. Deviations from these patterns can raise alerts before an adverse event occurs [40].

Lastly, Kenngott et al. [70] referred to the OR of the future as the means to realise "cognition-guided surgery". The principle of cognitive surgery is the use of technology to bridge IT infrastructures, medical equipment, staff and patients in the "Intelligent Hospital" or "Hospital 4.0" [70]. This is achieved by capturing perceptual information, interpreting it through a surgical knowledge base to generate context-aware actions (alerts on potential risk, camera guidance, etc.) and then closing the loop by appending experience in the knowledge base [70]. When this system provides the appropriate information at the right time to the relevant people, the clinical procedures can be optimised and patient safety enhanced [70].

2.4.2 Surgery 4.0 and Surgical Data Science

Since 2011, the term "Industry 4.0" has appeared to name the fourth industrial revolution [172]. The first industrial revolution (late 18th – early 19th century) concerns the exploitation of steam and water power that shifted production from hand methods to machines [173]. The second industrial revolution (late 19th - early 20th century) refers to the use of technology, electricity and subsequent infrastructures, that increased productivity and mass production [173]. The third industrial revolution (1970s) is defined by the integration of the most recent technological developments (electronics, IT, telecommunications) and infrastructures

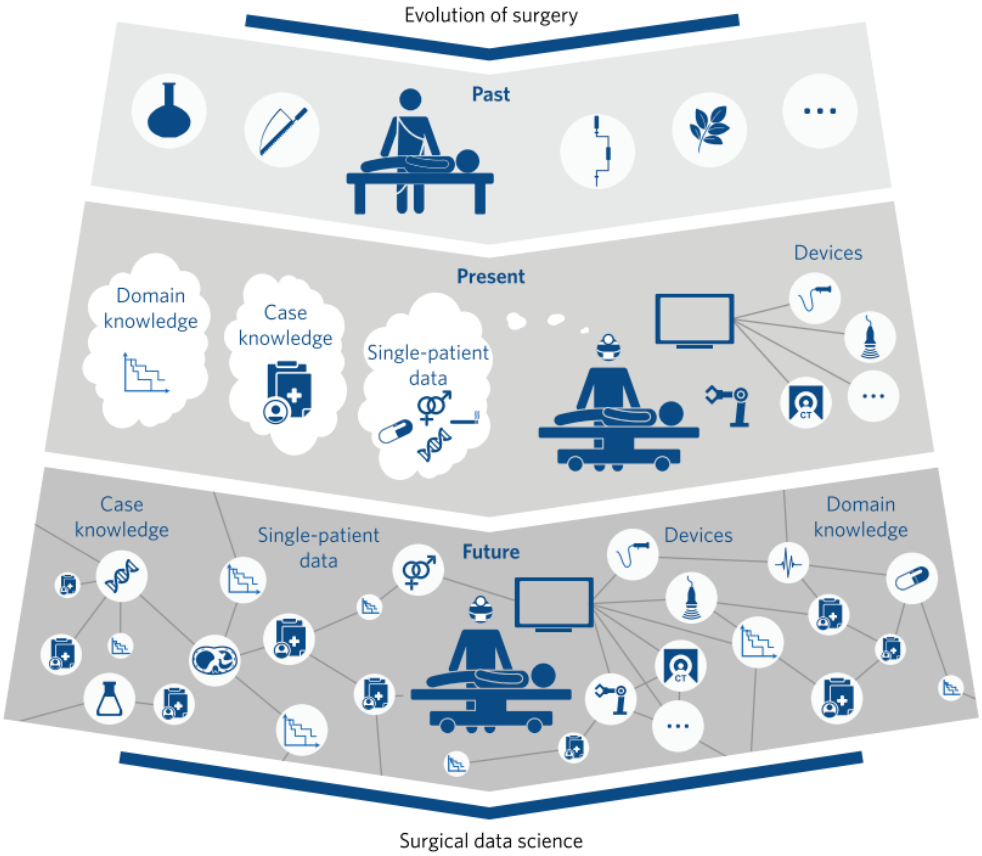


Figure 2.13: The evolution of surgery, reproduced with permission from Springer [29].

(railroad networks) in production that further augmented mass production [172, 173]. The fourth revolution refers to integration of information technologies and data science into the production line. Technologies such as the Internet of Things (IoT) can be used to capture large amount of data in real time and then analyse these data to determine current process status in order to optimise product value [172].

The advent of technologies like the IoT has led to the creation of similar terms and processes in the healthcare domain and especially in surgery. In this case, Surgery 1.0 stands for open surgery, Surgery 2.0 laparoscopic surgery, Surgery 3.0 robotic-assisted surgery with remote control, and Surgery 4.0 is information-based robotic surgery [174].

As the new paradigm of industrial production relies on the retrieval and fusion of a large amount of data through interconnected devices to generate a knowledge base, improved surgical outcome and Surgery 4.0 require plethora of multi-modal data provided by various sensing modalities in the operating theatre. Such modalities include patient biological signals, cameras, endoscopic video streams and other sensory data deriving from the theatre, the OR staff or the patients [175]. However, the lack of structure of this rich information, impedes their usability with artificial intelligence algorithms to build a knowledge base and integrate it into the surgical workflow [29, 175]. To acquire this amount of data in a safe, structured and dynamic manner, the need for a centralised tool emerges [40]; the OR.net project is such an attempt [42, 176, 177].

Surgical data science aims to harness these data under a common framework, to improve surgical outcome and quality of care through diagnosis, prognosis or treatment [29] (Fig. 2.13). To accomplish these, the synergy of these data with artificial intelligence, can facilitate recognising patterns, predicting events, providing guidance to the OR team for optimal decision making, improving ergonomics, automating tasks (i.e. in robotic-assisted surgery) and others [29].

To reach this evolution, allowing perceptually-enabled interactions in the OR and harnessing the perceptual data engendered by these interactions, is key.

2.5 Conclusion

The design of the operating theatre of the future is inspired by the unique opportunities revealed by recent technological advancements. Mimicking the integration of these technologies in industries such as aviation and autonomous driving, the OR can become an intelligent space where the surgeon will be always in the loop of the decision-making process. Nevertheless, the focus for designing novel applications for the OR must be patient safety. In this chapter the main factors which jeopardise patient

Table 2.2: Safety risk factors in the OR.

Communication failures	Procedural error
Information flow	Attention / situation awareness
Team coordination	Staff workload
Workflow disruptions	Fatigue
Performance	Staff dissatisfaction
Task efficiency	Language/cultural gap
Tension	Staff shortages
Delays	Equipment physical layout
Time pressure	

safety where summarised (Table 2.2). The aim of this thesis is to develop and translate state-of-the-art technologies in the clinical setup, to solve problems signified in the medical literature. Ultimate goal is the work presented here to contribute towards methods and applications which will enhance operator ergonomics, patient safety, team collaboration and staff training (Fig. 2.14).

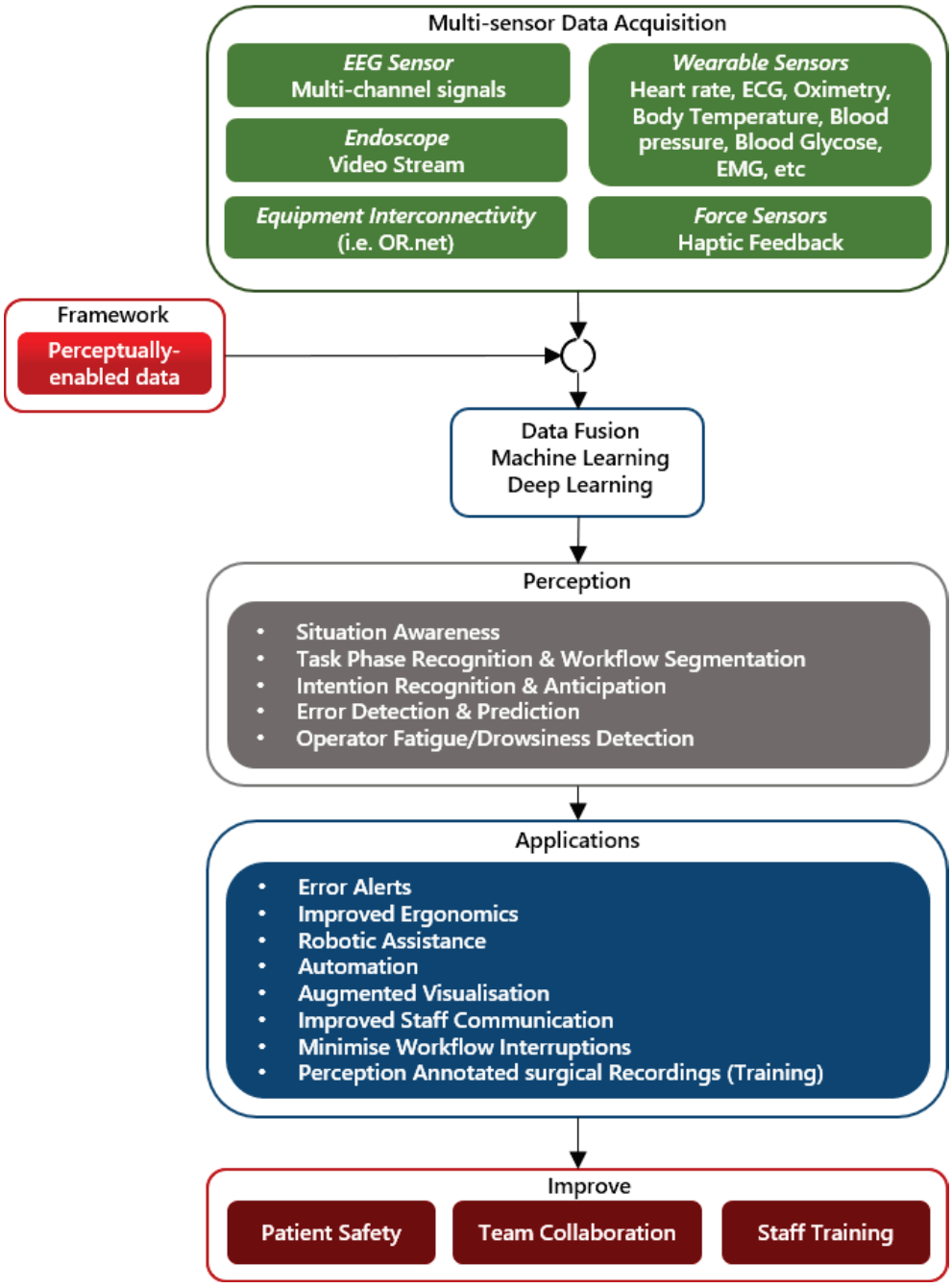


Figure 2.14: Potential applications of the perceptually enabled data deriving from this thesis, in conjunction with multi-sensor data, towards the improvement of patient safety, team collaboration and staff training.

Chapter 3

3D Gaze Localisation Framework [1]

3.1 Introduction

By approaching the eyes as the only visible part of the brain, we can consider them not just in the conventional sense as receptors of visual stimuli, but also as perception- and cognition-rich actors within the operating theatre. This approach could allow harnessing the power of the underlying mental processes that lead from perception to cognition to action, and seamlessly endow intelligence to implemented human-computer interfaces. Eye-tracking has been proposed as an input method when it is not possible to operate a system with human hands [85], thus providing the user with a "third hand". However, most of the systems available provide gaze information on a 2D plane, hence either a fixed surface such as a screen for remote eye-trackers, or a fixed virtual plane with respect to the user's head for head-mounted eye-trackers. To this end, 3D coordinates of the point of regard can provide semantic and

[1] Content from this chapter was published as:
Gaze-contingent perceptually enabled interactions in the operating theatre. Kogkas A., Darzi A., Mylonas G. International Journal of Computer Assisted Radiology and Surgery, 12, 1131–1140 (2017), doi: 10.1007/s11548-017-1580-y.

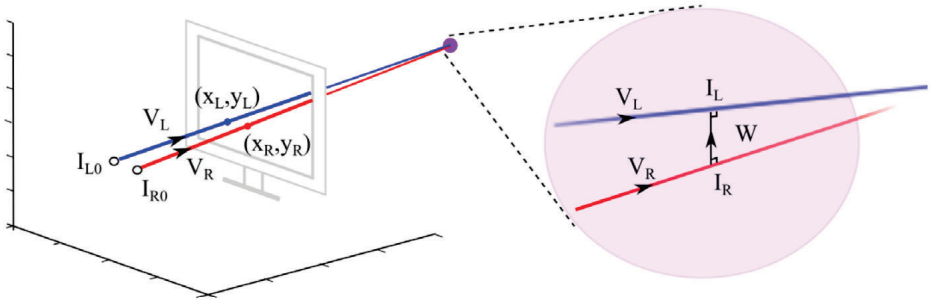


Figure 3.1: The visual axes intersection method and error, ©IOP Publishing. Reproduced with permission. All rights reserved, adapted from [30].

contextual correlation between the captured visual attention and the environment.

Only during the last decade research has explored the potential of 3D gaze tracking in virtual or real environments. Most of the approaches estimate the visual attention in the 3D space as an extension of traditional 2D gaze tracking techniques [178]. Kar et al. [179] provided an overview of the gaze estimation systems and algorithms and Larrazabal et al. [180] reviewed these techniques and assessed them focusing on their potential in clinical applications. Li in [178] classified the 3D gaze estimation methods into 3 categories: direct mapping method, visual axes intersection method, and depth plane method.

The direct mapping method is similar to the regression method used for 2D gaze tracking, hence a mapping function is used to map eye movements to 3D gaze locations [178]. The mapping function derives from a user calibration routine, where the user fixates at targets with known 3D coordinates and eye features are recorded [178]. However, due to the complexity of mapping eye movements to a 3D environment, accurate 3D gaze tracking is very challenging to be achieved [178]. Such implementations are presented in [181–184].

The visual axes intersection method relies on the assumption that the visual axes from the two eyes intersect at the 3D location where the user is fixating at, thus extracting the 3D PoR [178]. However, practically this

assumption is usually not affirmed, as the visual axes may not intersect in the 3D space (Fig. 3.1) [178]. Therefore, one point on each axis are defined, so that they have the shortest distance from the other axis and the middle point of the line connecting them is the estimated 3D PoR [178]. However, this method does not yield accurate results even in small workspaces, as the error in the visual axes propagate in the 3D gaze estimation pipeline [178]. Such implementations are proposed in [30, 181, 182, 185–188].

The depth plane method is similar to the geometric method used for 2D gaze tracking, hence intersecting a visual axis to a screen [178]. The assumption in this approach is that the 3D PoR is on a virtual plane perpendicular to the user's fixation axis [178]. First the visual axis is estimated and then the depth, as the vertical distance between the user and the virtual plane [178]. Such examples are proposed in [189–193]. Results have shown low accuracy in relatively small workspaces.

Another family of 3D gaze estimation techniques is the appearance based method (Fig. 3.2(b)). This method is based on learning a mapping function directly from correlating eye images to gaze directions. This function does not derive from a specific model, rather than is trained with eye images of known gaze directions. Although this method allows natural movements and can perform better than others when low resolution eye images are provided, it yields low accuracy results [32, 34, 194–199].

Another approach for 3D gaze tracking is based on head-mounted / wearable eye-trackers, thus allowing free head movement. It relies on the localisation of the eye-tracker's scene camera pose in space and then mapping the gaze vector from the head frame of reference to a known 3D scene. In [200–202] the 3D environment is reconstructed using the scene camera pose information and computer vision techniques, such as structure from motion. Other approaches use external hardware, such as RGB-D cameras to provide 3D spatial information [31, 33, 203–205] (Fig. 3.2(c)). Finally, motion capture systems have been added to improve the camera pose estimation accuracy and enable accurate 3D gaze data which

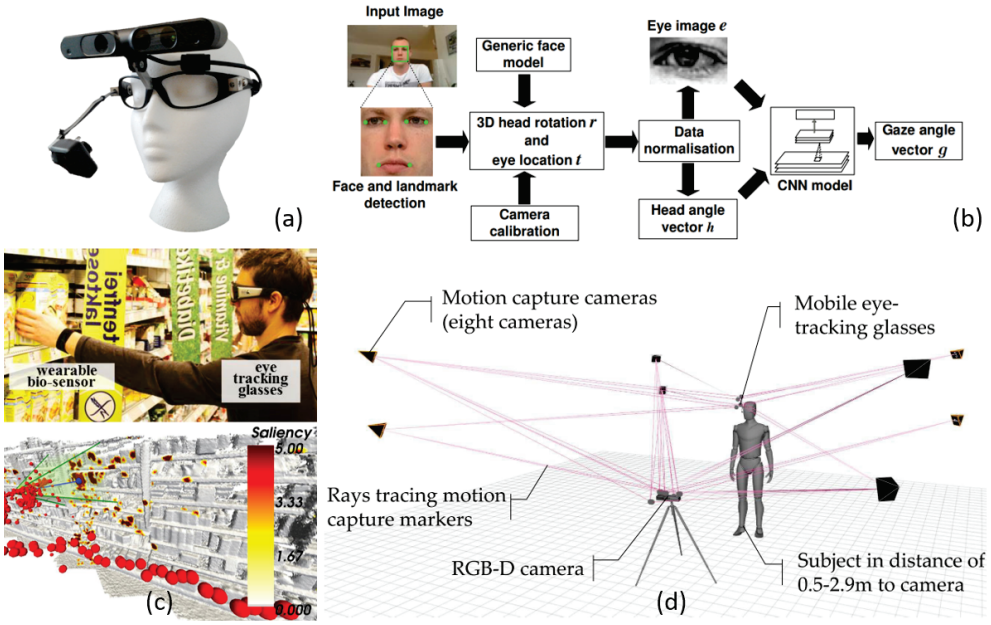


Figure 3.2: Gaze estimation approaches allowing natural head movements. (a) RGB-D camera in conjunction with head-mounted eye-tracker, reproduced with permissions by the authors [31]. (b) Appearance based method, ©2015 IEEE [32]. (c) ETG camera pose estimation in 3D space approach, reproduced/adapted with permission from Springer [33]. (d) Motion capture system approach for camera pose estimation, reproduced with permission from Springer [34].

can be used for generating gaze datasets (Fig. 3.2(d)) [34].

This chapter introduces a novel real-time framework, for free-view, global 3D fixation localisation, through the use of unrestricted wearable eye-tracking, dynamic spatial 3D reconstruction and camera pose estimation techniques. This framework allows simultaneous global macro- (theatre-wide) and micro-scale (patient/screen-wise) fixation localisation in the operating theatre. The accuracy of the core feature of the framework, the 3D fixation localisation, is assessed.

Overall, the work presented here is fundamentally driven by the need to keep the surgeons and their physical interactions with the environment tightly integrated into the decision-making process. Core functionalities of this multi-sensor framework presented in this chapter include: real-time

free-viewing 3D fixation localisation, spatial reconstruction and modelling of the operating theatre, micro-scale fixation localisation. One or more wearable eye-tracking devices can be used in combination with RGB-D cameras and advanced computer vision techniques. The ultimate goal is to develop functionalities, methodologies, open-source software and a low cost generic hardware framework that can be adapted to any operating theatre with minor modifications and effort.

3.2 Framework Overview

A core aspect of the proposed framework is its capability to calculate and display the 3D fixation of one or more theatre attendants (Fig. 3.3). Wearable eye-tracking glasses (ETG) and their integrated scene camera can be used to provide 2D gaze information and a scene video on the head frame-of-reference of a user. After a short calibration routine, gaze vectors can be mapped to unique 2D gaze points on a virtual plane attached to the scene camera of the ETG. This plane is also fixed to and rotates with the user's head. Consequently, there is no direct quantitative correlation between 2D fixations and 3D positions of objects in space. To overcome this limitation, localisation of 3D fixations is achieved through the combined use of conventional wearable eye-tracking, fixed in space RGB-D cameras for 3D reconstruction of the environment and (occasionally) a motion capture system (MCS) for the head pose estimation. The framework relies on the ability to provide an accurate estimate of one's head pose (equivalent to the ETG's scene camera pose) on a world coordinate system (WCS) fixed with respect to the operating theatre. The pose is then used to map the 2D gaze information reported by the eye-tracker to a unique 3D fixation in the world frame-of-reference. Then, the 3D fixation can be translated into screen 2D fixation information when the user gazes on a screen in space, allowing simultaneous macro- (theatre-wise) and micro-scale (patient/screen-wise) fixation localisation (Figs. 3.3-3.4).

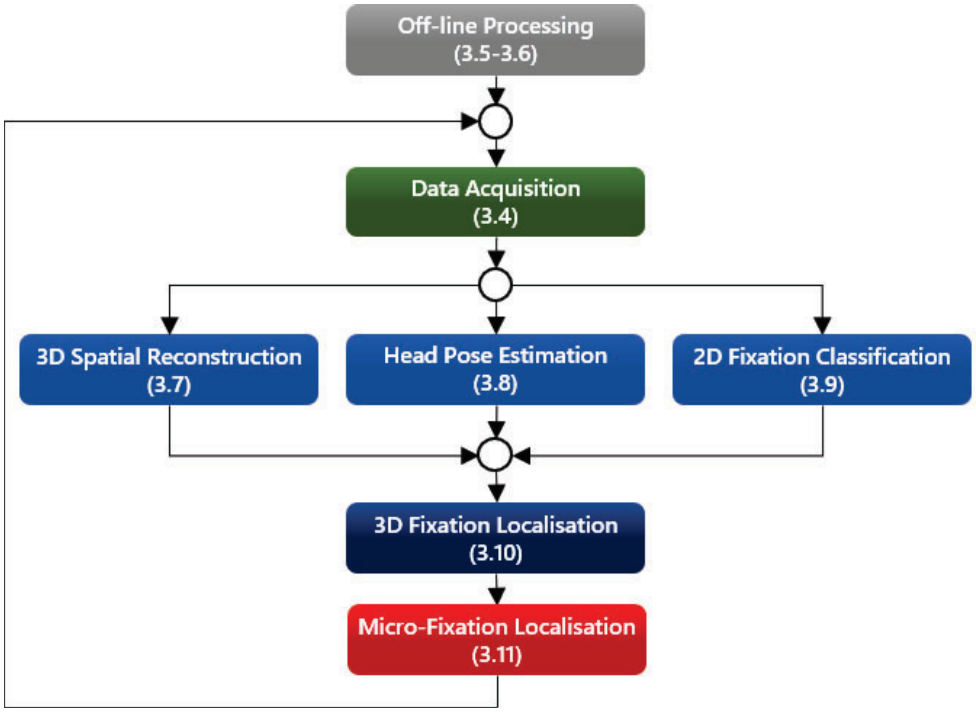


Figure 3.3: Framework overview flowchart. For each component, the corresponding section where it is described is reported.

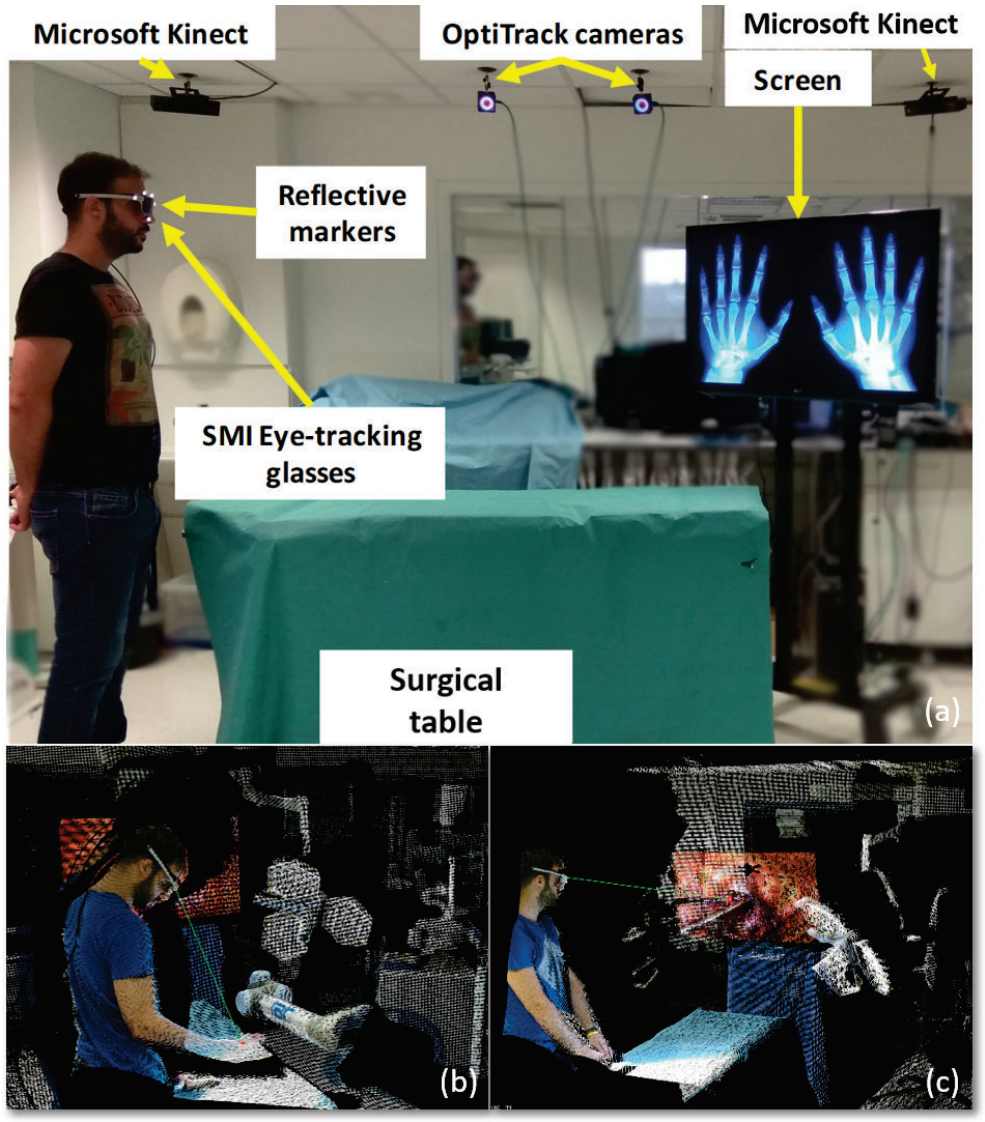


Figure 3.4: (a) The setup of the proposed framework. The user wears the eye-tracking glasses (ETG), where spherical reflective markers are mounted to form an asymmetric 3D structure. RGB-D cameras 3D reconstruct the theatre and the motion capture system (MCS) tracks the user's head pose (equivalent to the ETG's scene camera pose). The framework estimates the user's fixation theatre-wide/macro-scale (b) and patient-wise/micro-scale (c).

3.3 Equipment

3.3.1 Eye-tracking

For eye-tracking the SMI [19] Eye-tracking Glasses 2 Wireless (SensoMotoric Instruments GmbH) are used (Fig. 3.5(a)). By tracking the position of the pupil and/or artificially generated features using near-infrared light sources and miniature cameras on the glass frame, the gaze direction of the user on the scene camera's frame of reference can be determined. The glasses operate as a fully mobile gaze-tracking device with $60Hz$ sampling rate. An RGB scene camera with a resolution of 1280×960 pixels records an egocentric video at 24 frames per second. The field of view (FOV) of the scene camera is 80°(horizontal) and 60°(vertical). Scene video and eye-tracking data are streamed in real-time to a PC. The output of the system is a 2D gaze point on the image plane of the scene camera with a stated accuracy of 0.5°of visual angle.

3.3.2 RGB-D sensing

For RGB-D sensing, the Microsoft Kinect v2 is used for capturing depth and colour images concurrently (Fig. 3.5(b)). The Kinect uses an RGB camera with a resolution of 1920×1080 pixels at $30Hz$, an infrared emitter and an infrared camera with resolution of 512×424 at $30Hz$. It has $30ms$ latency and 2–$4mm$ average depth accuracy error [206]. The FOV of the depth sensing is 70°(horizontal) and 60°(vertical) and it operates at distances between $50cm$ and $\sim 4.5m$. For depth estimation, the time-of-flight method is used [207].

3.3.3 Motion Capture System (MCS)

For head pose tracking the OptiTrack MCS (NaturalPoint, Inc.) [208] is used, with four Prime 13 cameras with 240 fps and FOV 42°$\times$ 56°, stating sub-millimetre accuracy (Fig. 3.5(c)). Spherical reflective markers

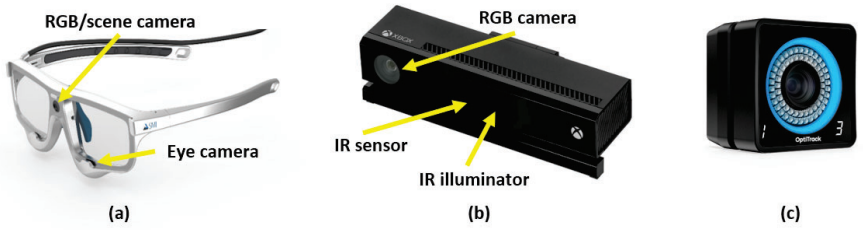


Figure 3.5: (a) SMI eye-tracking glasses (ETG), ©SensoMotoric Instruments (SMI) [19]. (b) Microsoft Kinect v2 RGB-D sensor[1]. (c) OptiTrack™ Prime 13, ©NaturalPoint Inc. [209].

are employed to define a rigid body geometry, which is tracked by the OptiTrack software.

3.3.4 Workstation

A Windows 10 PC is used for acquiring and streaming the ETG and MCS data and a Linux PC with Ubuntu 14.04 is used for all other modules. The Linux PC runs on Intel Xeon Processor, NVIDIA GTX 580 1.5 GB, 16 GB RAM.

3.4 Data Acquisition

The ETG's API by SMI provides the scene video and eye related data. The MCS provides with spherical marker's positions and the 6 DOF pose of a user defined rigid body's geometry in the motion capture system coordinate frame (MCS CS). The RGB frames and the 2D gaze information by the ETG and the 6 DOF rigid body pose of the ETG are streamed timestamped through UDP to the Linux PC, where they are decoded. For RGB-D camera the Kinect bridge [210] is used to acquire sensor data and convert them into ROS compatible messages.

[1]Source of original picture: https://commons.wikimedia.org/wiki/File:Xbox-One-Kinect.jpg

3.5 Calibration

The accuracy of the calibration process is of paramount importance. Four types of calibrations are performed:

- Camera calibration for the ETG's RGB scene camera
- User-specific eye-tracking calibration of the ETG
- RGB-depth calibration for the Microsoft Kinect sensor
- MCS cameras extrinsic calibration

3.5.1 ETG's RGB Scene Camera

Camera calibration refers to the estimation of the parameters of a lens and image sensor of an image or video camera. These parameters can be used for the correction of lens distortion, measurement of the real size of objects or localisation of the camera in the scene. Camera parameters include intrinsic / extrinsic parameters and distortion coefficients and are estimated by 3D-2D correspondences in world and image coordinates respectively. Usually, these correspondences are acquired by multiple images of a calibration pattern, such as a chessboard. The pinhole camera model represents a simple camera with a single small aperture instead of a lens. Its parameters are represented by the intrinsic (camera optical centre and focal length) and extrinsic (location in the 3D space) parameters, which map the 3D space into the image plane. With the extrinsic parameters, the world points are transformed to camera coordinates. With the intrinsic parameters the camera coordinates are mapped into the image plane. The intrinsic and extrinsic camera parameters of the eye-tracker (ETG) scene camera are calibrated using a chessboard and the camera calibration toolbox of OpenCV 3.2 library [211].

3.5.2 Eye-tracking

Mapping eye fixations to specific points in the image plane of the video sequence, provided by the RGB/scene camera, requires a calibration procedure. During this procedure, users are asked to fixate on a certain amount of predefined points in their FOV, keeping their head pose fixed. Using the API provided by SMI, the parameters of a generic physiological 3D eye model are refined and the model is used to calculate the gaze vector. The model is a combination of shapes, light refraction and reflection properties of the different parts of the eyes. This process is not transparent and is dealt with internally by SMI algorithms. Moreover, the SMI API allows only 1-3-point calibration. A 9-point calibration method using polynomial regression was implemented, to achieve higher accuracy in 2D fixations [212].

3.5.3 Microsoft Kinect Sensor

Although the RGB camera and the depth sensor of the Kinect are placed closely and capture similar planes, their slight spatial divergence may cause significant inaccuracies. The RGB and IR cameras intrinsic parameters and their rigid transformation were estimated using the calibration process provided by [210].

3.5.4 Motion Capture System

The spatial correlation of the four MCS cameras in the motion capture system coordinate frame (MCS CS) is calibrated through the OptiTrack Motive software package. It involves moving a calibration wand consisted of spherical reflective markers of known geometry while cameras are recording the sequence. After the cameras' calibration, the ETG rigid body is defined by six spherical reflective markers with fixed asymmetric geometry mounted on the ETG.

3.5.5 Multiple-Kinect Setup

The WCS in defined by the synergy of two or more RGB-D cameras fixed in respect to the operating theatre. Their spatial correlation is defined by capturing simultaneous RGB frames with chessboard observations in their common FOV and solving the non-linear least squares problem using the Levenberg-Marquardt algorithm [213, 214].

3.6 Coordinate Frames Registration

In the proposed system, we use the Kinect's coordinate system as the word frame of reference (WCS). To align multiple local coordinate systems to the global one, coordinate frame transformation is necessary.

3.6.1 SLAM to Kinect

One approach for head pose estimation used in this thesis is the Simultaneous Localisation and Mapping (SLAM) method [215]. This, relies on the localisation of a monocular camera within a local map, which is extracted during the initialisation of the method. Using a monocular camera for SLAM initialisation results to a scaled map, which is useful for tracking the camera pose in the 3D space. However, knowing the camera extrinsic parameters in relation to the world coordinates is desirable, as the 3D fixation in world coordinates is necessary for gaze-guided tasks, such as robotic arm manipulation or object recognition.

In [216] the initial camera pose is estimated by the correspondences of the two initial keyframes. The first keyframe is the frame of reference of the map. The initial pose is used to triangulate the map and full global bundle adjustment to refine the initial map. For the registration, fiducial markers (for 2D-3D correspondences) and the EPnP algorithm [37] are employed to estimate the pose of the two first keyframes in the Kinect's frame of reference (WCS). Defining the pose of the reference frame and the initial pose in the Kinect's coordinates result to the extraction of the

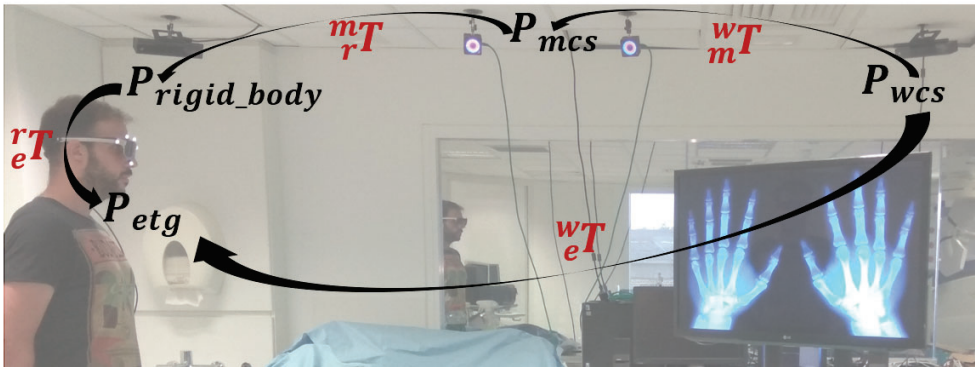


Figure 3.6: The transformations among the coordinate systems when a motion capture system (MCS) is employed to track the head pose.

initial map in the world coordinate system (WCS).

3.6.2 OptiTrack to Kinect (WCS – MCS CS)

The transformations shown in (Fig. 3.6) are described by the following equation:

$$^{w}_{e}T =^{w}_{m} T *^{m}_{r} T *^{r}_{e} T \tag{3.1}$$

Where:
$^{w}_{e}T$ is the 6 DOF ETG's scene camera pose in the WCS (P_{wcs}),
$^{w}_{m}T$ is the rigid transformation between the WCS (P_{wcs}) and the MCS CS (P_{mcs}),
$^{m}_{r}T$ is the 6 DOF pose of the rigid body (formed by reflective markers mounted on the ETG asymmetrically) in the MCS CS (P_{mcs}) and
$^{r}_{e}T$ is the rigid transformation between the rigid body (P_{rigid_body}) and the ETG scene camera (P_{etg}).

To estimate the 6 DOF ETG's scene camera pose in the WCS, the $^{w}_{m}T$ and $^{r}_{e}T$ rigid transformations need to be defined. Therefore, the problem is formulated into the hand-eye calibration problem $AX = YB$, where Y corresponds to $^{r}_{e}T$ and X to $^{w}_{m}T^{-1}$. The method by Shah et al [35] is used for this purpose. It involves capturing 6 DOF poses both of the

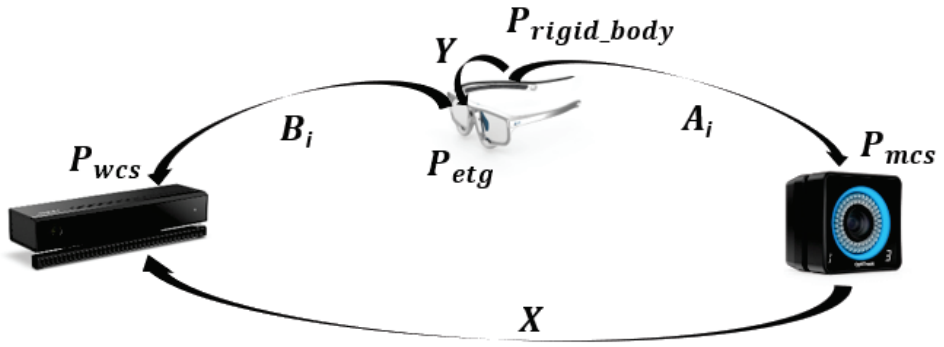


Figure 3.7: The hand-eye calibration problem formulation ($AX = YB$) [35].

ETG rigid body in the MCS CS (A or ${}^{m}_{r}T^{-1}$) and the ETG scene camera in the WCS (B or ${}^{w}_{e}T^{-1}$), simultaneously. The first is provided by the MCS API. The latter is calculated employing the EPnP algorithm [217], given the 2D-3D correspondences of an asymmetric checkerboard. The 2D correspondences are observed by the ETG scene camera and the 3D by the RGB-D sensor's RGB and IR camera (Fig. 3.7).

3.7 3D Spatial Reconstruction

An essential part of the proposed framework is the real-time continuous spatial reconstruction of the environment. The Microsoft Kinect v2 is employed to acquire the depth information of the scene as a depth image. Then, the depth image of each camera is converted into a point cloud using its intrinsic camera calibration parameters and *PCL library* [218]. The multiple-Kinect calibration result is used to produce a single point cloud of the environment.

After the scene is 3D reconstructed, processing of the point cloud is performed. At first we remove the outliers using statistical analysis techniques. Assuming the distribution of the points to their neighbours is *Gaussian*, all points with mean distance outside the interval defined by the mean distance and standard deviation of all points to all their

neighbours, are removed [219]. Then, the point cloud is compressed using octree representation [220] with the OctoMap library [221] and PCL [218]. This allows faster data transmission and (gaze) ray casting. A 3D voxel grid is created and the points are limited to the centroids of each voxel.

3.8 Head Pose Estimation

The estimation of the head pose in the 3D reconstructed space is the most essential part of the proposed framework. For the case of the wearable eye-trackers, the scene camera moves along with user's head movement, thus the estimation of the camera pose (extrinsic parameters) is equivalent to the head pose.

There are multiple approaches to address this process. Optical trackers are used to determine the camera pose in a world coordinate system, with high accuracy and high cost. Instead, the synergy of advanced computer vision techniques (i.e. PnP, SLAM) and 3D spatial reconstruction can be used to estimate the scene camera pose, with less accuracy but significantly lower hardware cost.

3.8.1 Perspective-n-Point (PnP)

In the computer vision literature, *Perspective-n-Point (PnP)* is the problem of estimating the pose of a camera given its intrinsic parameters and a set of n 3D-2D correspondences (points in the world – projections on the image plane). The camera pose has 6 DOF, consisted of the rotation and the translation of the camera to the WCS.

Lepetit et al in [37] proposed a closed form solution to the perspective-n-point problem, that uses n (for $n \geq 3$) 3D-to-2D point correspondences. According to EPnP, each point is expressed as a weighted sum of four virtual control points [37]. These points become the unknowns, so the problem is reduced to the estimation of the coordinates of four virtual control points in the camera referential [37].

Significant part of the pose estimation using EPnP, is defining the 3D-2D correspondences. In our implementation, this is achieved incorporating the ETG's scene video and the Kinect's RGB and depth images. For each 2D feature in the camera referential (ETG video) the 2D correspondence in the Kinect RGB image is obtained and then mapped to the Kinect depth image to provide the 3D correspondence. This approach relies on the accurate alignment of Kinect's RGB and depth images, acquired after the Kinect calibration procedure.

At first, the *Binary robust invariant scalable keypoints (BRISK)* algorithm [36] is used to detect features in both RGB images and extract the respective descriptors. Then, *Brute-Force matching* is used to match the corresponding features based on the Hamming distance of their descriptors. Finally, EPnP with RANSAC and Gauss-Newton Optimisation [37] provide the ETG's scene camera pose in space (Fig. 3.8).

There are significant limitations when incorporating a PnP algorithm to our system. This is because the feature matching algorithms are reliable under specific circumstances, such as for similar view angles between the corresponding planes, which means a high amount of outliers are present when PnP is applied on extreme frame angles, resulting to inaccurate head pose estimation. For sufficient results, the user's head pose would be restricted by the Kinect's positioning in the theatre. Consequently, the usability of the proposed framework would restrict the surgeon's movements and prevent an unrestricted free-viewing 3D eye-tracking experience. However, this method was deemed satisfactory in applications where the user does not require significant mobility (i.e. motion impaired patients).

3.8.2 Simultaneous Localisation and Mapping (SLAM)

Simultaneous Localisation and Mapping (SLAM) [222] is a method of building a map of an unknown environment by a mobile robot and estimating its pose within it. It consists of multiple phases, each of which can be computed in multiple ways: landmark extraction, data association,

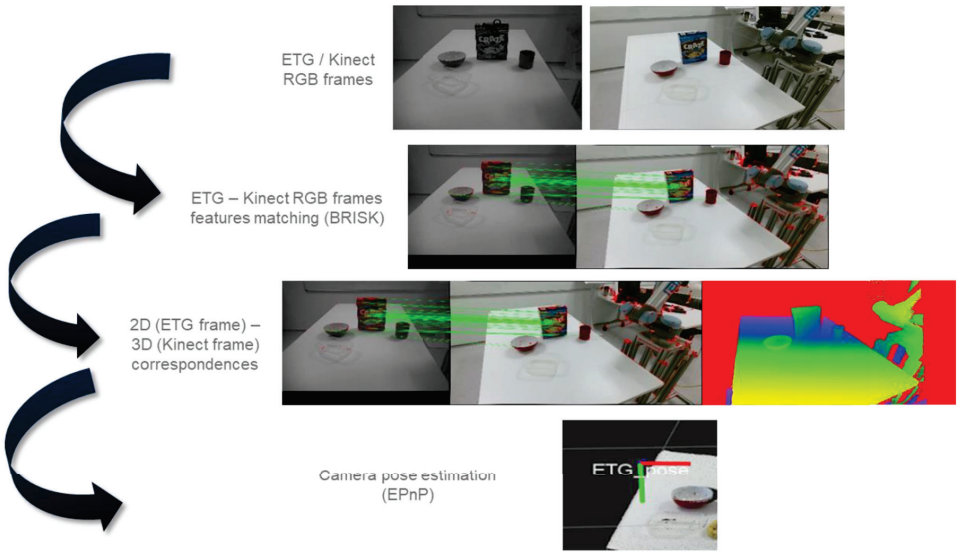


Figure 3.8: The head pose estimation approach using PnP to estimate the ETG camera pose. At first, the *BRISK* algorithm [36] is used to detect features in the RGB images (ETG scene camera and Kinect) and extract the respective descriptors. Then, *Brute-Force matching* is used to match the corresponding features. Finally, EPnP with RANSAC and Gauss-Newton Optimisation [37] provide the ETG's scene camera pose in space.

state estimation, state update and landmark update.

The ORB-SLAM algorithm [216] is used with a monocular camera to estimate its pose in a 3D environment and map features of video frames. This method is robust to severe motion clutter, allows wide baseline loop closing and re-localisation, and includes full automatic initialisation. The main tasks performed are: tracking, mapping, re-localisation, and loop closing. Tracking refers to the estimation of the relative position of the camera to the scene objects in real time. Mapping refers to the construction of a 3D map of the environment in which the camera moves. Using a short video sequence, ORB-SLAM generates an initial map using ORB features and a homography assuming a planar scene, or a fundamental matrix assuming a non-planar scene. Then, it builds/updates the keyframe-based map and tracks the camera pose

(extrinsic parameters) related to it. ORB-SLAM uses *Bundle Adjustment* for the map initialisation, local mapping and loop closing.

Whilst ORB-SLAM is a method to track a camera in an unknown environment, it shows a significant drift after extensive usage in a clinical scenario. The head pose component is of paramount importance to our framework, as a small offset would result to significant error in the 3D gaze estimation. Moreover, our implementation was not deemed satisfactory for clinical trials, as head/camera abrupt rotational movements cause robustness issues, such as losing tracking in space.

3.8.3 Optical Tracking

Optical tracking is a method to identify the pose of a tracked object by observing light on it [223].

An optical tracking system can be referred either as outside-in or inside-out tracking [223]. In an outside-in setup, the cameras are fixed in the environment and observe the tracked object which is moving freely. In an inside-out setup the camera is moving and tracks a reference frame [223].

The observed light on the tracked object can be either transmitted (active tracking) through LEDs, or reflected (passive tracking) [223]. For the latter, markers on the object are used which are coated with an infrared light retroflective material [223]. The light is transmitted by an infrared light source, reflected by the markers and observed by two or more cameras, which are rigidly positioned in space [223]. The markers are located at fixed positions on the tracked object, to form an asymmetric geometry.

Since active tracking relies on cables running to the LEDs, it is not widely used in clinical scenarios. Therefore, we use a passive tracking, outside-in setup. The OptiTrack motion capture system is used to estimate the ETG camera pose. Six spherical markers are mounted on the ETG to form an asymmetric rigid body and allow OptiTrack to provide its unique 6 DOF pose in space.

The benefits of employing such a setup include high measurement

accuracy and robustness to any interference by metallic or conducting objects [223]. However, passive optical tracking is sensitive to line of sight occlusions (markers invisible by the cameras) and scattered light from different sources (i.e. the sun, other infrared cameras) [223].

This head pose estimation approach is used to prove the feasibility and the benefits of the proposed gaze-contingent framework in clinical settings, where a vision based approach is not robust enough yet to be used during long surgical procedures.

3.9 2D Fixation Classification

Conventional wearable eye-trackers allow determination of a user's gaze direction. After the eye-tracking calibration procedure, the 2D PoR on the image plane of the video is estimated.

Saccadic and micro-saccadic movements can occur and need to be handled, as they constitute undesirable control commands. Saccades are very high speed "ballistic" eye movements occurring in between fixations, while micro-saccades are small amplitude and low frequency drift movements occurring during fixations [224, 225]. A filter is therefore implemented to discard non-fixational gaze-data. A fixation is classified using the method presented in [226], meaning that a set of consecutive eye movements that maintain a relatively constant velocity in the visual field of view are classified as fixation. Fast saccadic movements are filtered out with a velocity threshold of 36 deg/s and dwell time is set to 0.2 s. Further filtering of the fixation data is performed, applying a median filter to eliminate noisy data derived by micro-saccadic eye movements.

3.10 2D to 3D Fixation Localisation

The mapping of 2D fixations to 3D world coordinates is based on calculating the gaze direction vector and its intersection on the 3D reconstructed model. The ray casting feature by OctoMap library [221] and PCL [218] is

used to backproject the gaze ray from the octree compressed scene model on the estimated camera pose's origin, allowing real-time 3D fixation localisation.

Definition of the gaze ray direction vector requires the calculation of two points, the 2D fixation in world coordinates and the camera centre of projection. The ray is defined by the line connecting these two points. First, the 2D point X_c is transformed in the camera coordinate system:

$$X_c = K^{-1} * p_c \tag{3.2}$$

then the point is transformed in the world coordinate system:

$$X_w = R^{-1} * (X_c - T) \tag{3.3}$$

and finally the centre of projection C_{op} is calculated:

$$C_{op} = -R^{-1} * T \tag{3.4}$$

Where:
$p_c = \begin{bmatrix} u & v & 1 \end{bmatrix}^T$ is the homogeneous coordinates of the image point,
K is the matrix of intrinsic camera parameters and
$\begin{bmatrix} R & T \end{bmatrix}$ are the rotation and translation of the camera (extrinsic parameters).

3.11 Micro-scale Fixation Localisation

Among the assets of the proposed framework is its ability to combine the benefits of both wearable and remote eye-trackers, hence mobility of the user in space and gaze localisation on a screen respectively, by using only wearable eye-tracking glasses. We define this attribute as *hybrid macro- and micro-scale fixation localisation. Macro-fixation* refers to the 3D fixation in the 3D reconstructed environment and *micro-fixation* is

the 2D PoR on a screen positioned in space.

In the previous sections we described how the 3D gaze information is estimated. For the micro-scale requirements, the gaze ray vector is defined (as in the previous section) and the four corners of the screen in the WCS is either manually defined offline or dynamically tracked through the RGB-D cameras. The screen corners form two equal triangles and the *Möller-Trumbore* ray-triangle intersection algorithm [227] is used to calculate the intersection between the ray and the triangular facets. Thus, gazing on the screen is detected and provided as 3D fixation on the screen model. As a final step, given the known screen dimensions, the screen plane position can be refined using perspective transformation and accurate micro-scale fixation is obtained.

Figures 3.3, 3.9, 3.10 and 3.11 demonstrate the 3D gaze estimation pipeline of the proposed framework, based on the methods discussed in the previous sections.

3.12 Framework (Software) Architecture

The system is developed in *ROS* with C++ in Linux Ubuntu 14.04, to facilitate the hardware agnostic aspect of the framework. *Hardware agnostic* refers to the ability to operate the proposed framework with various hardware for each category, with only minimum integration effort. Thus, any kind of wearable eye-tracking glasses, RGB-D cameras and motion capture systems can be plugged in to the system. As such, software maintenance and adaption to new technologies are facilitated and hardware specific limitations are easier to be overcome.

Fig. 3.12 shows the ROS architecture of the system. The RGB-D data decoder receives the raw RGB-D sensor data and outputs the RGB and depth image frames. Windows PC data decoder receives the Windows PC data through UDP and streams ETG and MCS related data. 3D scene reconstruction, head pose estimation and 2D fixation classification nodes provide the necessary information for the 3D gaze estimation node,

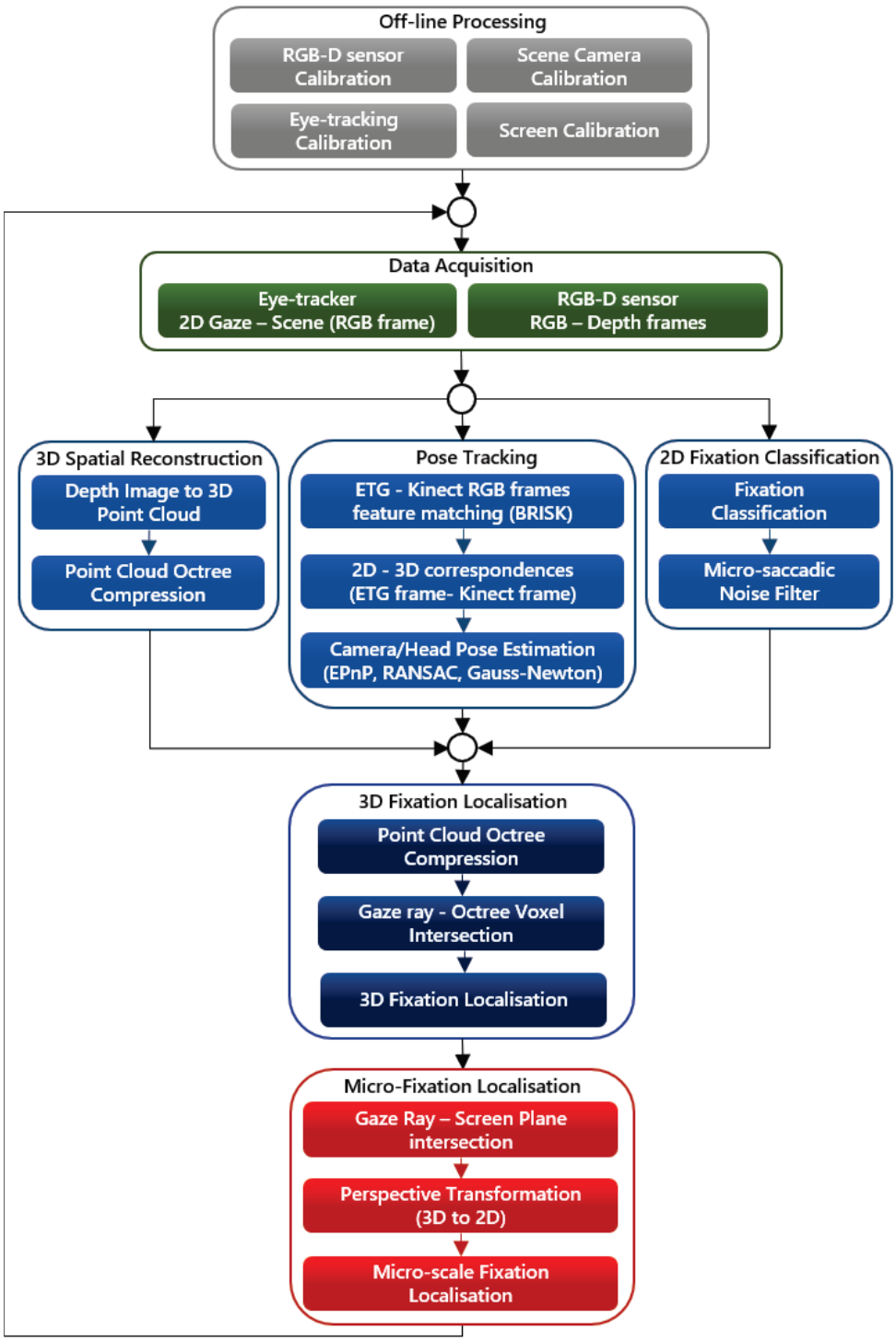


Figure 3.9: Framework flowchart using the PnP approach for head pose estimation.

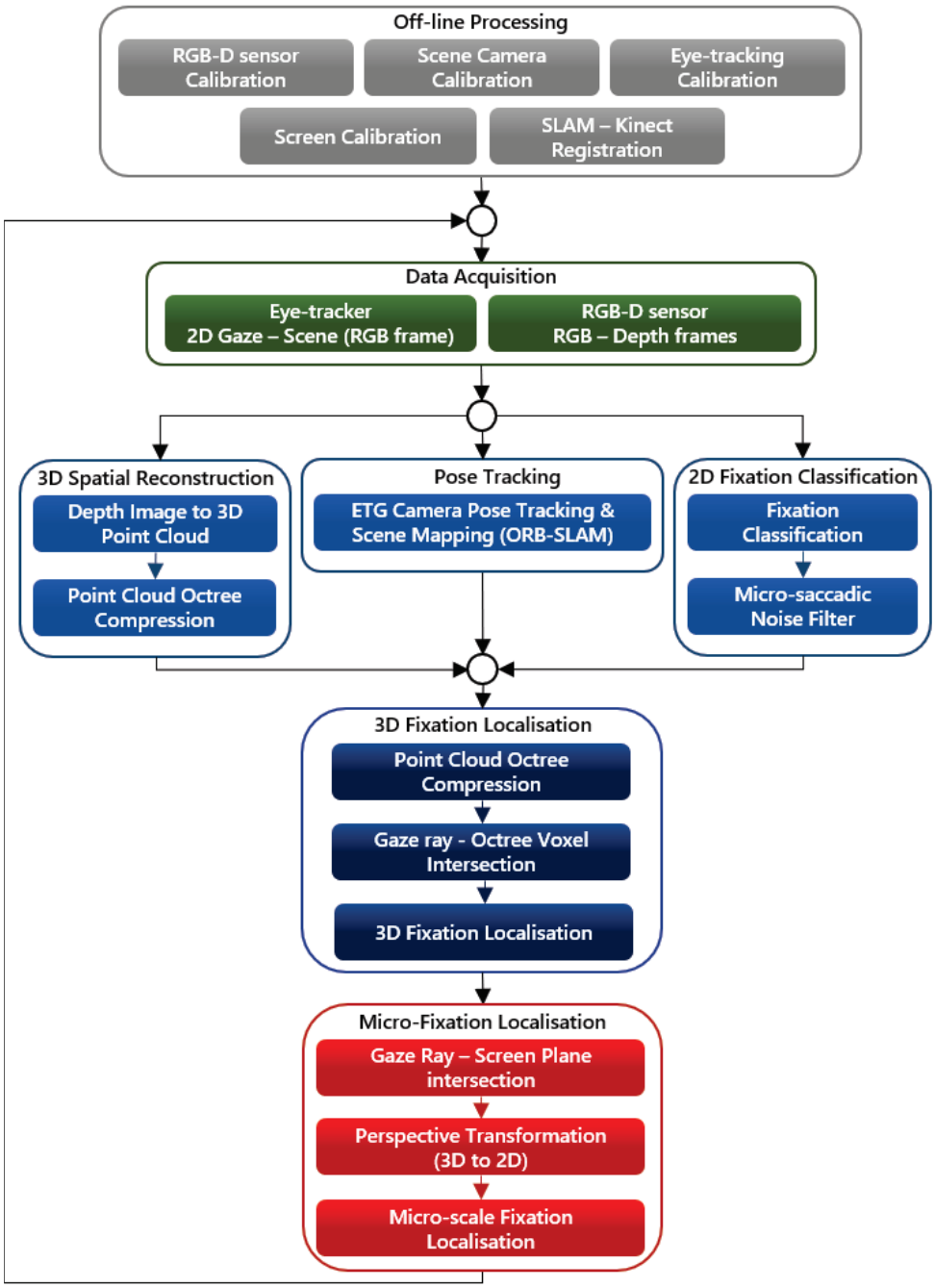


Figure 3.10: Framework flowchart using the SLAM approach for head pose estimation.

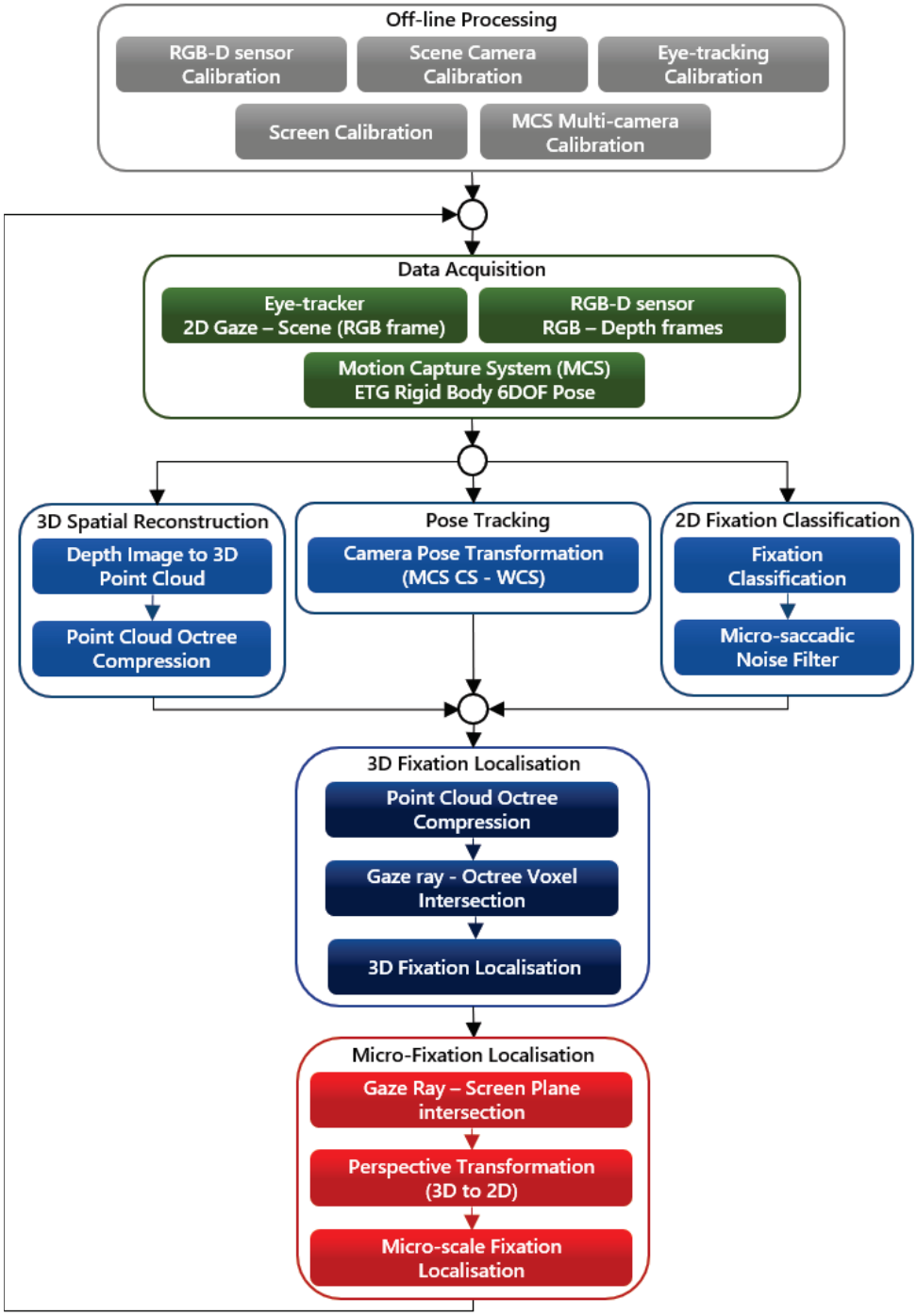


Figure 3.11: Framework flowchart using the Motion Capture System (MCS) approach for head pose estimation.

which feeds the micro gaze estimation node.

3.13 Validation Method

The accuracy of the 3D fixation localisation is assessed in conditions simulating a surgical setup, hence fixating on a surgical table and a screen, from different positions (free-view) and using a unique calibration. The validation of the 3D fixation localisation accuracy aims to define the feasibility and design requirements of the gaze-contingent framework in clinical applications.

Six subjects, aged between 24-32 years, were recruited and asked to fixate on ten predefined targets in space standing at two different positions in the simulated OR. The distance range between the subject and the targets is $97cm$–$229cm$. All 20 fixations per subject were measured with a unique calibration, to assess the performance of the framework in multi-modal applications, involving interactions both in macro- (theatre-wide) and micro-scale (screen-wise).

The task starts with the subjects fixating on 9 predefined targets to perform eye-tracking calibration. Then the subjects are asked to fixate on each of the 10 targets in order and this process is repeated from 2 different predefined positions (Fig. 3.13).

The estimated 3D fixations were compared to the actual 3D coordinates of the targets in the WCS by measuring their angular offset θ. The value θ is an angle calculated using the 3D target vector V_{CT} and 3D fixation vector V_{CF}:

$$\theta = cos^{-1}\left(\frac{V_{CT} \cdot V_{CF}}{\|V_{CT}\|\|V_{CF}\|}\right) \tag{3.5}$$

where V_{CT} and V_{CF} are the vectors defined by the camera centre of projection C_w, the target 3D point T_w and the estimated 3D fixation F_w (Fig. 3.14).

Figure 3.12: The ROS architecture of the 3D gaze framework. Black oval shapes represent nodes, red boxes sensors and grey boxes data/messages published in the ROS environment. The RGB-D data decoder receives the raw RGB-D sensor data and publishes the RGB and depth image data. The Windows PC data decoder receives the Windows PC data through UDP and streams ETG and MCS related data. 3D scene reconstruction, head pose estimation and 2D fixation classification nodes provide the necessary information for the 3D gaze estimation node, which feeds the micro gaze estimation node. All messages can be used for the implementation of gaze-contingent application nodes.

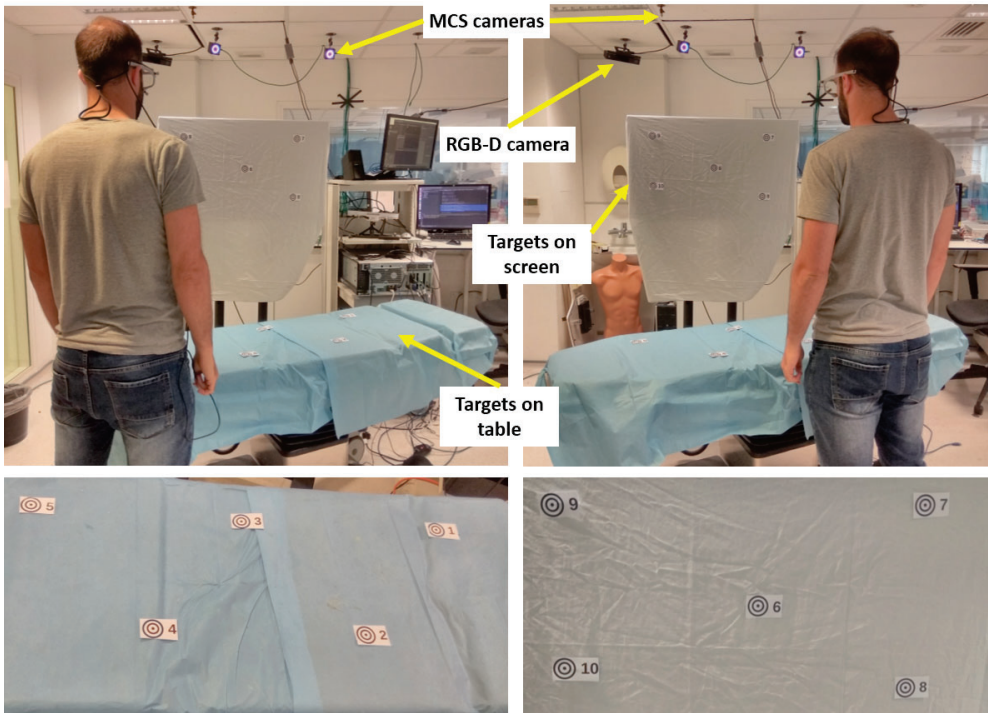


Figure 3.13: The experimental setup for the framework accuracy validation. The participants performed the tasks from 2 different positions (top). 10 targets where positioned in the setup, 5 on a surgical table and 5 on a screen (bottom).

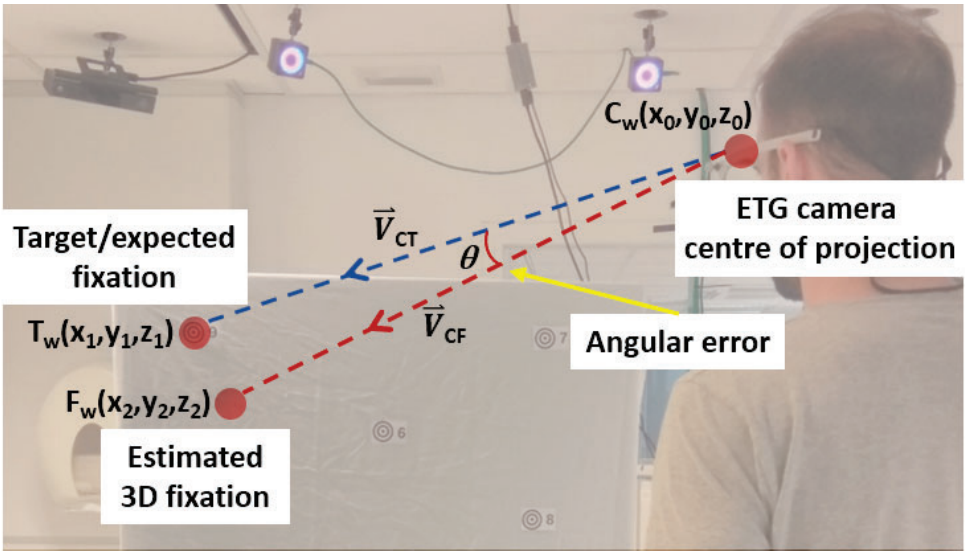


Figure 3.14: Definition of the angular error θ. The vectors V_{CT} and V_{CF} are defined by the camera centre of projection C_w, the target 3D point T_w and the estimated 3D fixation F_w. All variables are represented in the world coordinate system (WCS).

3.14 Results

The results depicted in Fig. 3.15 show the median error and distribution overall, per subject, per target and per target area (surgical table or screen). The median angular error over 120 fixations is 2.52°.

The Shapiro-Wilk Test was performed with 0.001 level. The angular error is not normally distributed ($p<.001$), therefore the median error is reported. To explore whether human factor has significant influence on the framework error, the Kruskal Wallis test was performed. The results reported that the angular error is not affected by the human factor with 0.001 level ($\chi^2(5) = 17.433$, $p = 0.004$).

3.15 Discussion and Conclusions

A novel real-time framework has been presented that allows gaze-driven interactions within a 3D environment. This is achieved by the combination of unrestricted wearable gaze-tracking, theatre 3D reconstruction and computer vision concept.

Each individual component of the framework (hardware and methodologies) introduces intrinsic error to the 3D fixation localisation pipeline, which propagates to the final system error. For example, the Kinect sensor produces an average error of 2–4mm, but depending on the distance from the target this may increase to over 4mm [206]. This error propagates to the system through its registration with the MCS CS and the 3D fixation localisation (in the WCS). Moreover, error is introduced and propagated towards the output of the system through the eye-tracker's inaccurate gaze estimation, especially over multiple distances (parallax effect [228]). This could be eliminated with the use of the framework presented here, by performing multiple use-specific calibrations over multiple distances and then accordingly switch or interpolate between the derived calibration parameters based on a resolved fixation depth. Moreover, head pose estimation is one the most significant stages of the framework and is

introducing an error, either due to the complexity of the computer vision approach used or the calibration errors in case the motion capture system is used.

The accuracy of the framework was assessed in conditions simulating a surgical setup, hence fixating on a surgical table and a screen. Most importantly, each subject performed the experiment using a unique calibration, while fixating on the targets from two different positions. The computed angular error of 2.52°signifies the feasibility of integrating the framework into clinical settings and sets the requirements to design robust gaze-contingent applications in healthcare.

The work presented in this chapter represents an introduction and experimental validation of core functionalities of a larger gaze-contingent framework. The proposed framework is geared towards a safer and more efficient surgical theatre. It is envisaged that an open-source and hardware-agnostic framework will allow large-scale deployment in several theatres. This would provide a large amount of easily anonymised data, which will help generate a large evidence base and critical mass for clinical use. Exemplar functionalities, which aim at enhancing ergonomics, safety, collaboration and training include: *gaze-guided object recognition and tracking, robotic manipulation, augmented visualisation of gaze relevant information, behavioural analysis and workflow segmentation based on perceptual information provided by the framework.*

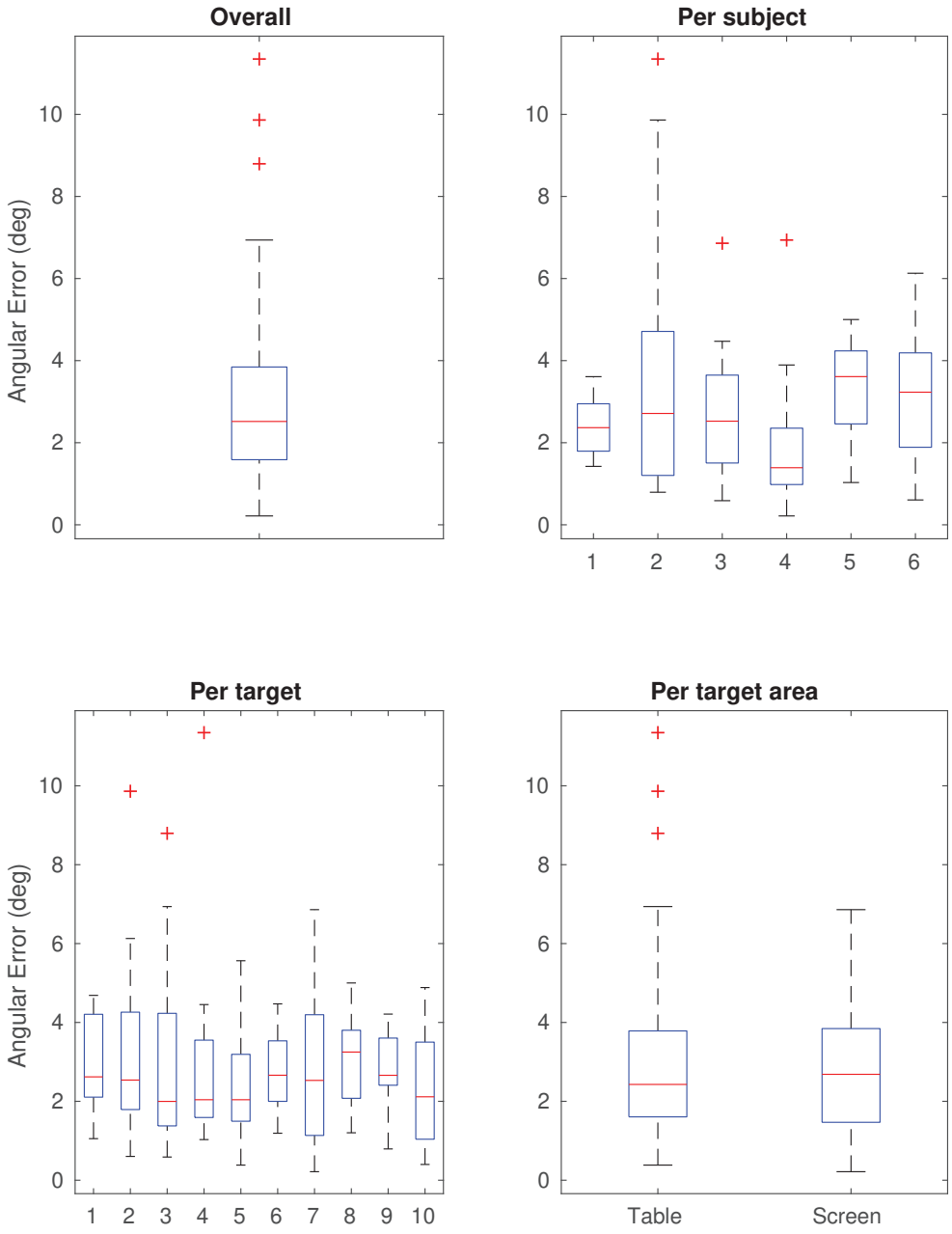


Figure 3.15: Angular error (*degrees*) of 3D fixation, from 2 different positions. Six subjects were asked to fixate on 10 targets positioned on a surgical table (5) and a screen (5). A unique calibration was used across all 20 fixations per subject.

Chapter 4

A Gaze-controlled Robotic Scrub Nurse [1]

4.1 Introduction

Technology advances within surgery have seen operating habits transform over the past number of years. Certain surgeries have seen traditional techniques replaced by robotic assisted surgery, now accepted by the surgical community as mainstream practice [229].

Thus, more research has targeted the development of further assistive robotic devices to improve operating practice. Healthcare associated human error has been reported as a leading cause of preventable patient harm and has at times resulted in avoidable patient death [230]. As such, recent advances in touchless artificial intelligence have allowed the introduction of such assistive robotic devices aimed at optimising surgical performance, operating time, operating flow and team working [231].

Eye tracking glasses worn by the surgeon can be used to measure

[1] Content from this chapter was published as:
Free-View, 3D Gaze-Guided Robotic Scrub Nurse. Kogkas A., Ezzat A., Thakkar R., Darzi A., Mylonas G. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2019. Lecture Notes in Computer Science, vol 11768, doi: 10.1007/978-3-030-32254-0_19, Springer, Cham.

parameters such as blink rate and gaze drift when fixating on a subject in order to reflect surgeon concentration [232, 233]. Other devices allow the utilisation of touch free navigation systems aimed at maximising accessibility to important surgical information required during surgery, including radiological images or patient notes [2]. This modality has been theorised to reduce the risk of infection transmission and deliver a seamless system which minimises interruptions [234].

To achieve this, systems such as Gestix has relied on predetermined user hand gestures control through 2D cameras to enable a particular action such as magnifying or changing the image [119]. HERMES VRI is a voice command-based system which was trialed within laparoscopic surgery and shown faster operating times [231]. More recently, there has been an expansion of research directed at gaze-controlled navigation due to perceived advantage in practicality during surgery and limitations of the interruptions of hand gesture control, as well as difficulty of voice recognition systems when scrubbed within the noisy operating environment [234].

Delivering surgical robotic assistance that can augment or replace human team members is an area attracting a lot of research with much promise. Such devices include the da Vinci® (Intuitive Surgical, Inc.), which is an established success story and is a robotic system controlled by the surgeon via a computer-based console. Da Vinci® has been used in 1.5 million laparoscopic surgeries to date with clinical outcomes demonstrating reduced post-operative pain and hospital stay and improved surgical accessibility and view in confined anatomical spaces [235]. Other examples include automated laparoscopic devices, such as that of a camera, and rely on the surgeon's head position to move the instrument and show up to 15% reduction in task completion [236]. Gestonurse is a magnetic based robotic scrub nurse which uses hand gesture-based selection of different surgical instruments and delivers the desired instrument. The authors reported 95% hit rate in robot gesture recognition [9]. Penelope™ has been described in the literature as the first robotic scrub nurse successfully

used in surgery. Penelope^{TM} is reported as a semi-autonomous system relying on surgeon verbal commands to pick up, predict the desired instrument and deliver. It can also detect if an instrument has not been used for a period of time and return it to the instrument tray [237].

An extension of the gaze-contingent framework (chapter 3) is presented here, that allows hands-free gaze-driven interactions with a screen and a robotic arm, which acts as a Robotic scrub nurse (RSN) assistant by transferring surgical instruments to the surgeon. The introduction of a RSN as an integral component of the operating theatre of the future, may address nursing shortages [71] and empower the team by enabling the surgeon and the Human scrub nurse (HSN) to perform a wider variety of tasks in a more efficient and safe manner. This is achieved by offering a "third hand", but most importantly one that is under the firm control of the operating surgeon.

In this chapter we present the system workflow and test the usability and acceptability of this novel eye-tracking based RSN by the operating team, during realistic surgical procedures.

4.2 System Overview

The system is developed in *ROS* with C++, to facilitate the hardware agnostic aspect of the framework. The core functionality of the real-time framework is to provide the user's 3D point of regard (PoR) in the world coordinate system (WCS), defined by multiple co-registered RGB-D sensors fixed in the theatre. It relies on a) estimating the pose of the scene camera integrated with the eye-tracking glasses (ETG) in the WCS and b) tracing the gaze ray provided by the ETG on the head frame of reference onto the 3D reconstructed space. The ETG scene camera pose is estimated with the employment of a motion capture system (MCS) and spherical markers mounted on the ETG. The provided camera pose, 2D fixation and parameters provided by an off-line calibration process, enable gaze control of a screen in space. A graphical user interface (GUI)

designed allows gaze selection of surgical instruments to be delivered by a robot arm.

4.3 Equipment

For eye-tracking, the SMI (SensoMotoric Instruments GmbH) glasses are used, with a stated accuracy of 0.5°of visual angle, a scene camera with a resolution of 1280 × 960 px and field of view (FOV) 80°× 60°. For RGB-D sensing, two Microsoft Kinect v2 cameras are used, with an RGB resolution of 1920 × 1080 px at 30Hz, time-of-flight technology, FOV of the depth sensing 70°× 60°and operation distance between 50 cm and ∼4.5 m. For head pose tracking the OptiTrack MCS is used, with four Prime 13 cameras with 240 fps and FOV 42°× 56°. The robot arm is a UR5 (Universal Robots A/S), a 6 degrees of freedom (DOF) collaborative robot with a reach radius of up to $850mm$, $\pm 0.1mm$ repeatability, ±360°joint ranges, maximum $5kg$ pay-load and weighing $18.4kg$. It has the Robotiq FT-300 Force/Torque sensor mounted on its end-effector. For the instrument selection GUI, a 42" LG screen with 1920 × 1080 px resolution is used (Fig. 4.1).

For the current implementation, a Windows 10 PC is used for acquiring and streaming the ETG and MCS data and a Linux PC with Ubuntu 14.04 is used for all other modules. The Linux PC runs on Intel Xeon Processor, NVIDIA GTX 580 1.5 GB, 16 GB RAM.

4.4 Data Acquisition

The ETG's API by SMI provides the scene video and eye related data. The MCS provides with spherical marker's positions and the 6 DOF pose of a user defined rigid body's geometry in the motion capture system coordinate frame (MCS CS). The RGB frames and the 2D gaze information by the ETG and the 6 DOF rigid body pose of the ETG are streamed timestamped through UDP to the Linux PC, where they are

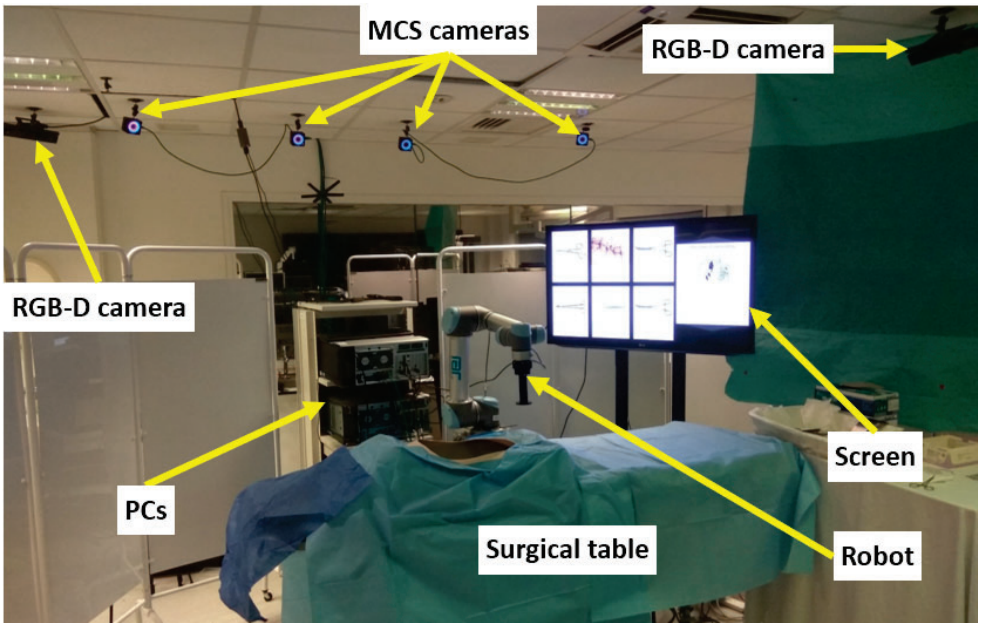


Figure 4.1: The setup of the robotic scrub nurse system.

decoded. For RGB-D camera the Kinect bridge [210] is used to acquire sensor data and convert them into ROS compatible messages.

4.5 Offline Calibration

ETG's scene/RGB camera. The intrinsic camera parameters of the eye-tracker scene camera are calibrated using a chessboard.

Eye-tracking. Eye fixations were mapped to specific points in the ETG's scene camera plane by asking the user to fixate in 9 predefined points in their FOV, keeping their head pose fixed.

Kinect sensor. The RGB and IR cameras intrinsic parameters and their rigid transformation were estimated using the calibration process provided by [210].

MCS. The spatial correlation of the four MCS cameras in the MCS CS is calibrated through the OptiTrack Motive software package. It involves moving a calibration wand consisted of spherical reflective markers of known geometry while cameras are recording the sequence. After the

cameras' calibration, the ETG rigid body is defined by six spherical reflective markers with fixed asymmetric geometry mounted on the ETG.

Multiple-Kinect setup. The WCS is defined by the synergy of two RGB-D cameras fixed with respect to the operating theatre. Their spatial correlation is defined by capturing simultaneous RGB frames with chessboard observations in their common FOV and solving the non-linear least squares problem using the Levenberg-Marquardt algorithm.

WCS – MCS CS registration. The rigid transformations between the ETG rigid body – ETG scene camera (Y) and MCS CS – WCS (X) is estimated by solving the hand-eye calibration problem $AX = YB$ [35]. This involves capturing 6 DOF poses both of the ETG rigid body in the MCS CS (A) and the ETG scene camera in the WCS (B), simultaneously. The first is provided by the MCS API. The latter is calculated employing EPnP [217], given the 2D-3D correspondences of an asymmetric checkerboard. The 2D correspondences are observed by the ETG scene camera and the 3D by the RGB-D sensor's RGB and IR camera.

Screen position in the WCS. The screen corners' 3D coordinates in the WCS are manually selected on the Kinect RGB image. The depth correspondence is estimated using the Kinect calibration and the 3D points generated.

Instrument positions in the robot coordinate system (RCS). Instruments are positioned in fixed positions on a tray, where their shape and name are printed and placed on top of it. The robot is manually moved towards each instrument and the target pose is calibrated.

4.6 Interface Design

The GUI displayed on the screen consists of two parts: instrument selection (left 2/3) and the image navigation (right 1/3).

Left: Six blocks equally split demonstrate common surgical instruments (Fig. 4.2). Micro fixation on any of the blocks initiates a traffic

light sequence (red-amber-green) followed by relevant audio feedback. Starting with red block borders, dwell time of 0.6 s into the same block turns the borders into orange and another 1 s turns the borders into green. The design is based on pilot experiments aimed to allow the user to have sufficient feedback (audio/visual) for the estimated micro fixation (red), be warned before finalising the instrument selection (amber) and confirm the action (green). The time intervals are decided in an attempt to balance avoidance of the Midas touch problem (non-intentional gaze-based selection) and disruption from the task workflow.

Right: Three slides are presented to provide information necessary for the task workflow. The user can navigate through them by fixating on the top and bottom 1/6 parts of the screen for previous and next slide respectively. Dwell time here is 1 s.

4.7 Robot Control

The selection of an instrument on the screen (server) triggers the robot (client) to handle the corresponding instrument to the user. TCP/IP is used to transmit the instrument ID to the robot client. The robot client has predefined poses for homing, instruments grasp and instruments delivery. The rounded nonlinear move mode of the robot encoder is used here. The robot moves towards instruments grasping pose (90°/s, 900°/s^2), grasps the instrument with the magnetic gripper and delivers it to the user (120°/s 400°/s^2). Then the robot stays idle until the F/T sensor senses the instrument collection by the user and 2 s extra time to ensure proper instrument collection. Finally, it returns to its homing position (191.5°/s, 900°/s^2).

4.8 Application Workflow

User head-pose (equivalent to the ETG's scene camera pose) provided by the MCS and transformed to the WCS, can be used to map 2D gaze

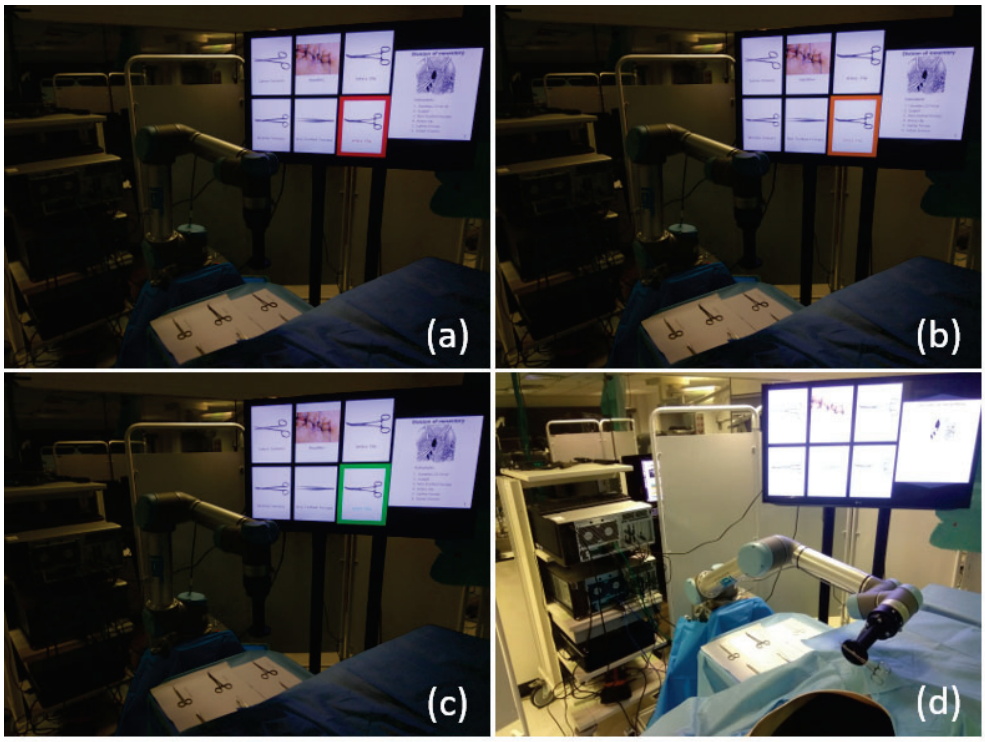


Figure 4.2: Egocentric view of the surgical instrument selection routine. (a) The surgical trainee (surgeon - ST) looks at an instrument (red), (b) the instrument is preselected (orange), (c) then selected (green) and (d) the robot delivers it to the ST.

to a unique 3D fixation in the WCS. The 3D gaze ray is used to detect fixations on the screen fixed in space (micro-fixation). The GUI consists of two parts: instrument selection (left) and image navigation (right). The image navigation part shows task workflow steps. The user can navigate through it by fixating on the right top and bottom of the screen. Micro-fixation on any of the instrument blocks initiates a traffic light sequence (red-amber-green) followed by relevant audio feedback. After a certain dwell time the robot routine is triggered. The robot moves towards a surgical instrument selected by the user, grasps it with a magnetic gripper and transfers it to the user. When the F/T sensor mounted on the robot senses the instrument is picked up, it returns to its homing pose (Fig. 4.3).

4.9 Validation Method

4.9.1 Experimental Design

Surgeons (surgical trainees – ST) were recruited to perform ex vivo resection of a pig colon and hand sewn end-to-end anastomosis. Each surgeon performed two experiments in randomised order:

- A Human scrub nurse only task (HSNt) with the assistance of a human scrub nurse (HSN).
- A Robot and human scrub nurse task (R&HSNt) with the assistance of both robotic (RSN) and human (HSN) scrub nurses.

In both experiments, a surgeon assistant aids the surgeon and a scrub nurse assistant the HSN. The instrument tray inventory consists of the 6 most frequently utilised instruments during this particular task: a suture scissors, a Mcindoe (curved) scissors, a non-toothed forceps, two artery clips and a hand suture attached to an artery clip. The main stages of the task are presented on the right part of the screen (Fig. 4.4).

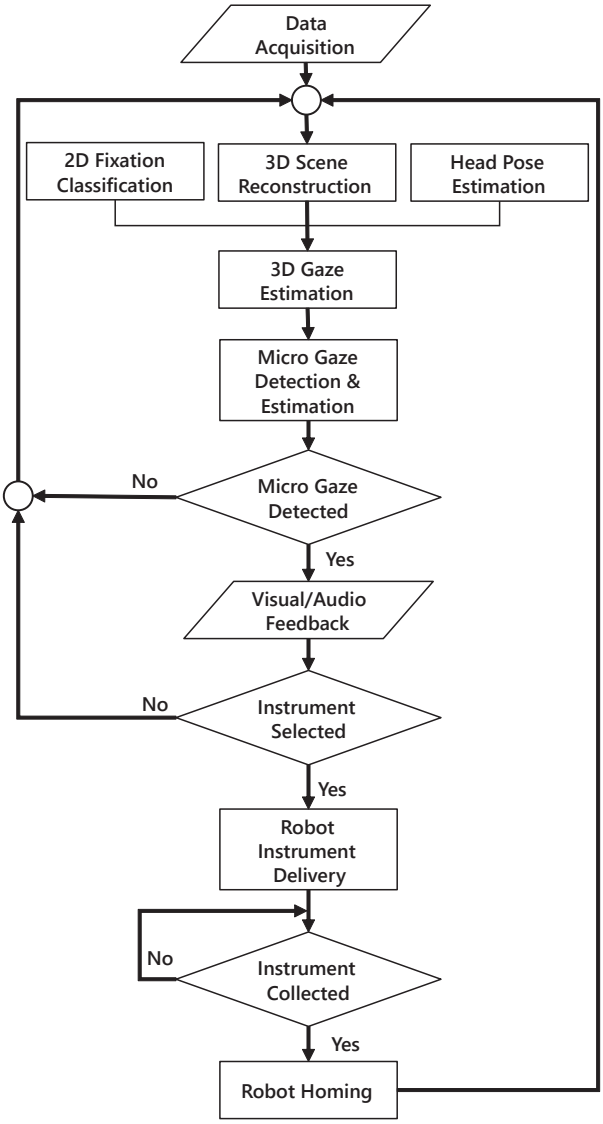


Figure 4.3: Flow chart of the Robotic scrub nurse (RSN) system. The 3D gaze ray, provided by the 3D gaze framework, is used to detect fixations on the screen (micro-fixation). Micro-fixation on any of the instrument blocks initiates a traffic light sequence (red-amber-green) followed by relevant audio feedback. After a certain dwell time the robot routine is triggered. The robot moves towards a surgical instrument selected by the user, grasps it with a magnetic gripper and transfers it to the user. When the F/T sensor mounted on the robot senses the instrument is picked up, it returns to its homing pose.

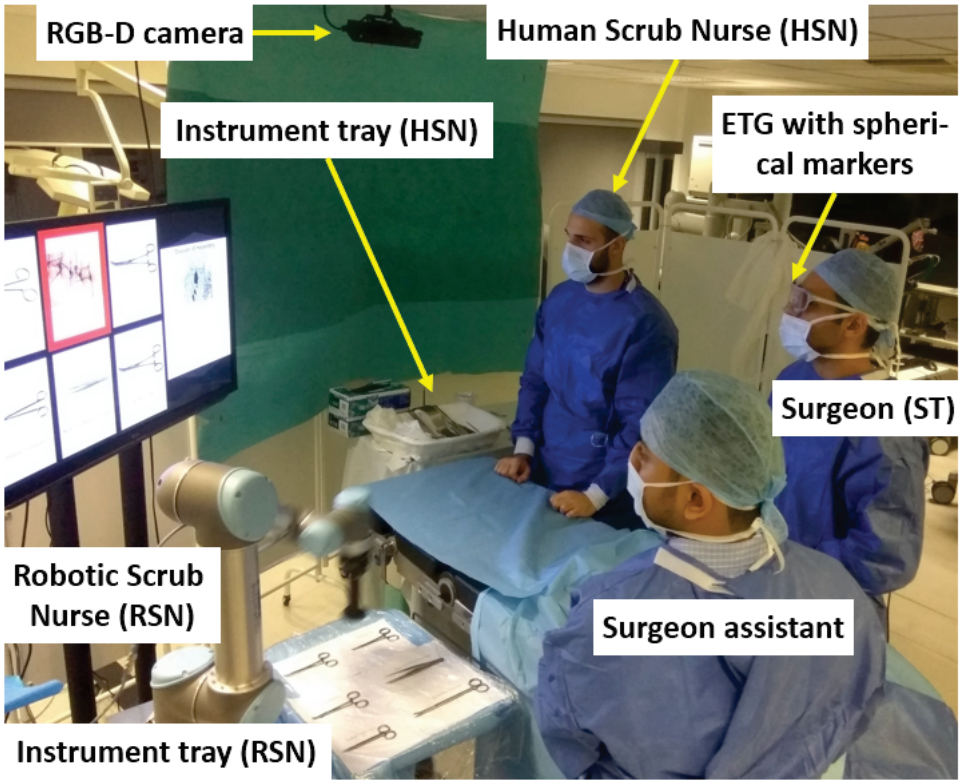


Figure 4.4: The experimental setup. The motion capture system (MCS) cameras track the spherical markers on the eye-tracking glasses (ETG) and provide its 6 DOF pose. The RGB-D cameras provide the 3D model of the operating theatre, in which the user's 3D gaze ray is estimated. The surgeon (ST) gazes on the screen to select an instrument and the robot delivers it. The surgeon assistant assists with the surgical task and returns the used instruments to the Robotic scrub nurse (RSN) tray. The Human scrub nurse (HSN) delivers instruments from a different instrument tray.

For the R&HSNt, the surgeon uses the ETG and is asked to fixate on 9 predefined points to perform eye-tracking calibration. Familiarisation with the system setup is offered for 1 minute. During the task, the surgeon looks at the screen to select an instrument and once it is delivered and collected, the surgeon assistant responds to verbal command or prior experience to return the instrument to its predefined position on the instrument tray. The surgeon uses verbal command directed towards the HSN when further instruments are required. In case the wrong instrument is delivered, the surgeon expresses the error verbally. If eye-tracking recalibration is necessary, the task continues after recalibration. During the HSNt the setup is identical. The screen and RSN are switched off and the surgeon relies entirely on the HSN to deliver instruments based on verbal commands. ETG is utilised to capture and analyse visual behaviour.

During both experiments, distractions are introduced to the HSN. The scrub nurse assistant asks the HSN to stop and perform an instrument count twice and solve a puzzle at specific task stages.

4.9.2 Participants

10 surgical trainee specialists (ST) participated (7 male and 3 female). Two had corrected vision. Surgeons were between 30-40 years with at least 6 years surgical experience. 5 trained theatre scrub nurses (HSN) were recruited. One surgical trainee, with 2 years surgical experience, acted as surgeon assistant and one medical student acted as scrub nurse assistant for all experiments.

4.9.3 Subjective Validation

Task Load

After each task, the ST and HSN were asked to complete a *NASA-TLX* (System Task Load Index defined by NASA) questionnaire. The scale assesses the mental, physical and temporal demand, own performance,

Table 4.1: NASA-TLX questions [1]

Mental Demand	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
Physical Demand	How much physical activity was required? Was the task easy or demanding, slack or strenuous?
Temporal Demand	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
Own Performance	How successful were you in performing the task? How satisfied were you with your performance?
Frustration	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?

frustration levels and effort during the task (Table 4.1). An overall task load score is calculated as described in [1].

Technology Acceptance

Technology usability and satisfaction feedback was collected immediately following the R&HSNt using the *Van Der Laan* acceptance scale [238]. The scale consists of five usefulness metrics (useful/useless, good/bad, effective/superfluous, assisting/worthless, raising alertness/sleep-inducing) and four satisfaction metrics (pleasant/unpleasant, nice/annoying, likeable/irritating, desirable/undesirable). Each item was on answered a 5-point semantic differential from -2 to +2.

4.9.4 Objective Validation

Performance

Performance was assessed in terms of overall task completion time. The task starts with the surgeon assistant's oral instruction "START" and finishes with the oral indication "FINISH".

Workflow

Workflow interruptions were measured for both tasks. During the HSNt, interruptions are defined as the events of a wrong instrument delivery by the HSN and the delay for instrument delivery by the HSN, which causes interruption of the task by the ST for $> 3s$. During the R&HSNt, the same interruptions are measured (HSN-derived) in addition to the RSN-derived events, namely incorrect instrument selection/delivery and eye-tracking recalibrations.

4.10 Results

4.10.1 Data Analysis

Previously published results have been produced performing between-subjects analysis. In this chapter, within-subjects analysis was applied were appropriate to account for possible baseline differences between-subjects. The comparisons demonstrated in the following sections were conducted using within-subjects analysis when comparing:

- Task completion time of HSNt vs R&HSNt
- Number of interruptions in HSNt vs R&HSNt
- NASA-TLX scores of ST in HSNt vs R&HSNt
- NASA-TLX scores of HSN in HSNt vs R&HSNt

Between-subjects analysis was conducted when comparing:

- NASA-TLX scores of HSNt by ST vs HSN
- NASA-TLX scores of R&HSNt by ST vs HSN
- Van der Laan's scores by ST vs HSN

For within-subjects analysis, the Shapiro-Wilk test for normality of the paired differences was performed, followed by paired-samples t-test when the test was successful and no outliers were detected. In case of non-normal distribution of the differences or the presence of outliers, the Wilcoxon signed-rank test was used.

For between-subjects analysis, the Shapiro-Wilk test for normality of the samples was performed, followed by independent-samples t-test when the test was successful. In case of non-normal distribution of any of the two samples, the Mann-Whitney U test was applied.

For all types of statistical analysis tests, a p-value <0.05 was considered significant.

4.10.2 Subjective Data

Task Load

The NASA-TLX scores overall and per category are depicted in Fig. 4.5. ST subjective feedback reported no significant difference overall (Table 4.2). ST did not report any significant change on task performance (p=0.526). ST did report significant frustration using RSN 22±10.6 vs 51.5±19.3, p=0.012. HSN feedback reported significant difference overall (39.9±19.6 vs 24.6±15.9, p=0.017). Frustration remained unchanged (p=0.833), whilst mental, physical demand and effort showed significant differences in favour of the R&HSNt. Comparison of ST vs HSN using RSN showed significant difference overall (57.5±15.8 vs 24.6±15.9, p<.001) and specifically in all sub-scales, in so demonstrating reduced HSN demands. There was a significant difference in frustration 0.7±3.86 vs 5.9±5.82, p=0.017. Comparison of ST vs HSN perceptions over the HSNt showed

no significant difference overall (p=0.161). Task performance score was significantly higher for the ST (33.5±20 vs 18±14.2, p=0.009).

Technology Acceptance

The ST group reported usefulness score of 0.5±0.73 and satisfying score of 0.43±0.74 (Fig. 4.6). ST reported that the RSN was likable 0.4±0.84, useful 0.5±1.08 and pleasant 0.8±0.79. ST feedback was neutral about RSN desirability 0.1±0.99. HSN feedback reported usefulness score of 0.76±0.92 and satisfying score of 0.78±0.79. HSN reported RSN was likable 0.6±1.26, useful 0.7±1.42 and pleasant 0.9±0.99. RSN was perceived as desirable 0.7±0.82. Upon comparison of ST vs HSN using RSN there was no statistically significant difference in *technology acceptance* domains (Table 4.3). Overall responses were positive in ST and HSN groups (usefulness score of 0.5±0.73 / satisfying score of 0.43±0.74 vs usefulness score of 0.76±0.92 and satisfying score of 0.78±0.79, respectively).

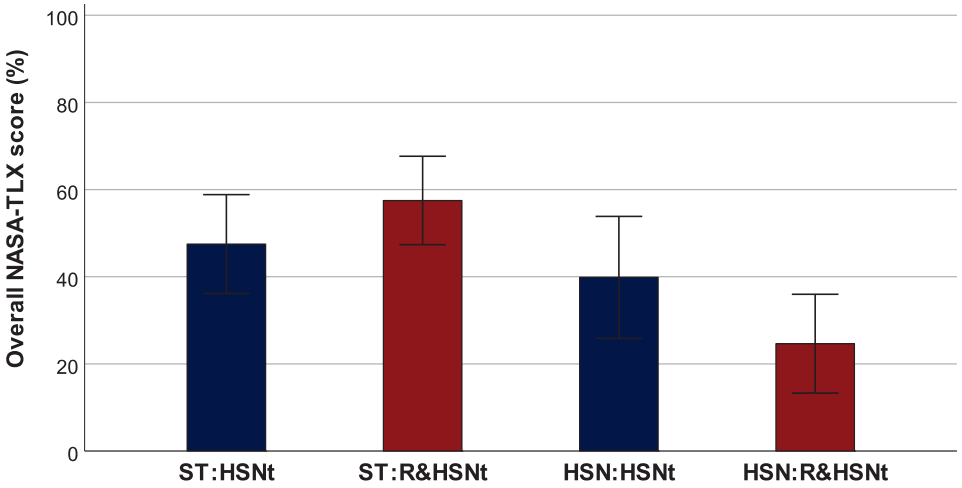
4.10.3 Objective Data

Performance

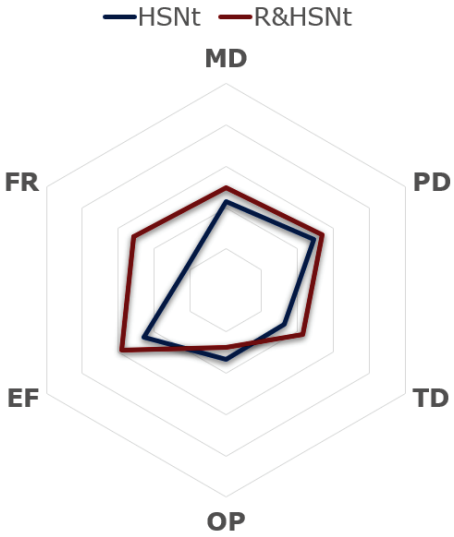
The HSNt mean duration was 22:35±6:30 minutes vs 26:04±4:50 minutes for the R&HSNt (Fig. 4.7). Comparison showed no significant difference in overall *task completion time* (p=0.074) in R&HSNt vs HSNt (Table 4.4).

Workflow

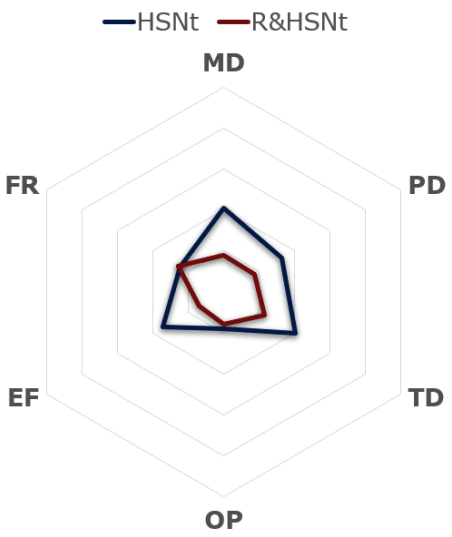
The comparative analysis of the total number of *interruptions* per task is shown in Table 4.5. No significant difference incurred (p=0.84) between R&HSNt and HSNt (2.3±0.95 vs 2.4±1.26, respectively).



(a)



(b)

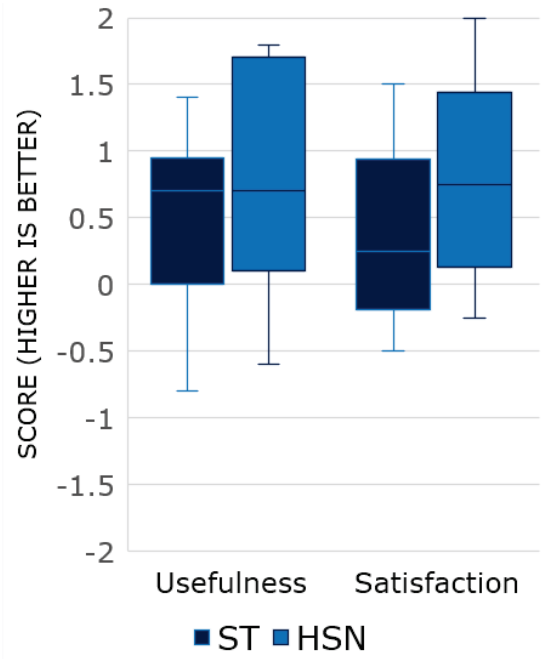


(c)

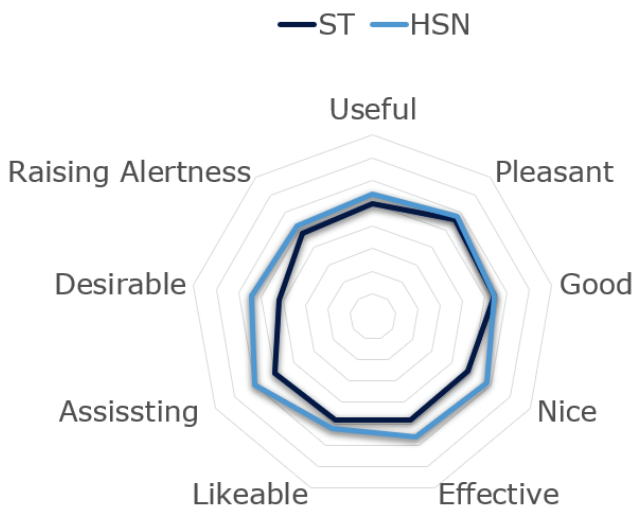
Figure 4.5: (a) Overall NASA-TLX score and analytical results (*MD, PD, TD, OP, EF, FR*) for (b) Surgeons (ST) and (c) Human scrub nurses (HSN). NASA-TLX values range between 0 and 100, with higher values indicating higher task load. *(HSNt: Human scrub nurse only task, R&HSNt: Robot and human scrub nurse task)*

Table 4.2: NASA-TLX score comparison of Surgeons (ST) and Human scrub nurses (HSN) on Human scrub nurse only task (HSNt) and Robot and human scrub nurse task (R&HSNt). *p-values* are reported.

	ST on: HSNt vs R&HSNt	HSN on: HSNt vs R&HSNt	HSNt by: ST vs HSN	R&HSNt by: ST vs HSN
MD	0.309	**0.008**	0.858	**0.004**
PD	0.812	**0.026**	0.141	**0.001**
TD	0.249	0.081	0.451	**0.02**
OP	0.526	0.657	**0.023**	**0.019**
EF	0.120	**0.009**	0.147	**<.001**
FR	**0.012**	0.833	0.73	**0.017**
Overall	0.052	**0.017**	0.161	**<.001**



(a)



(b)

Figure 4.6: (a) Overall Van der Laan's technology acceptance score by Surgeons (ST) and Human scrub nurses (HSN) and (b) analytical results. The usefulness scale derives from the average of *useful/useless, good/bad, effective/superfluous, assisting/worthless, raising alertness/sleep-inducing* metrics and satisfaction scale derives from *pleasant/unpleasant, nice/annoying, likeable/irritating, desirable/undesirable* metrics. The scale range between -2 and +2, with higher values indicating positive bias on the specific attribute.

Table 4.3: Van der Laan's technology acceptance scores comparison between Surgeons (ST) and Human scrub nurses (HSN) on Robotic scrub nurse (RSN).

	Usefulness	Satisfaction
p-value	0.491	0.320

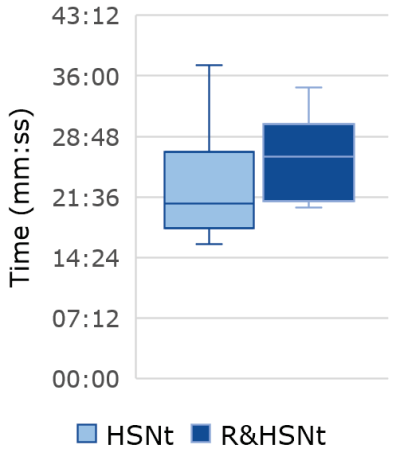
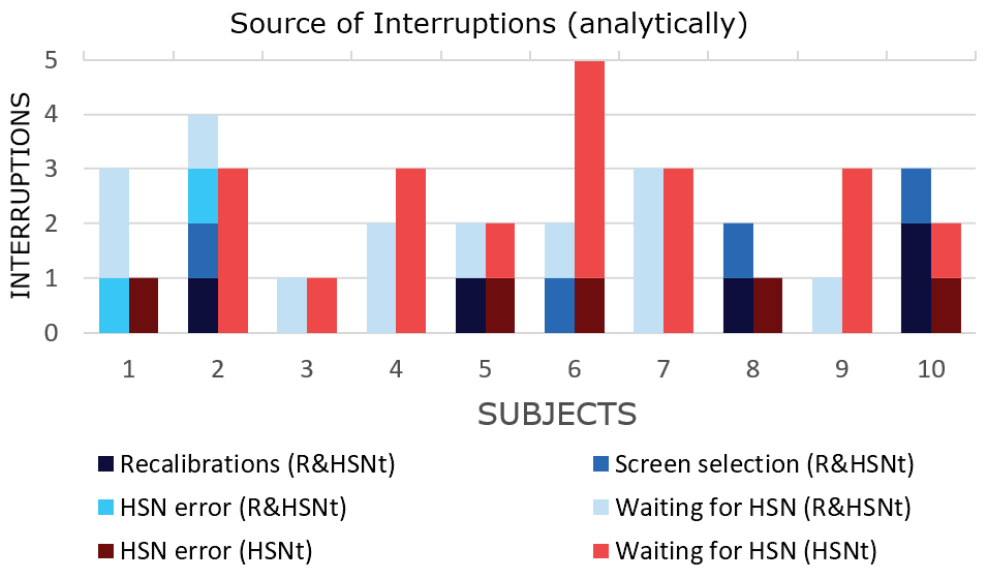


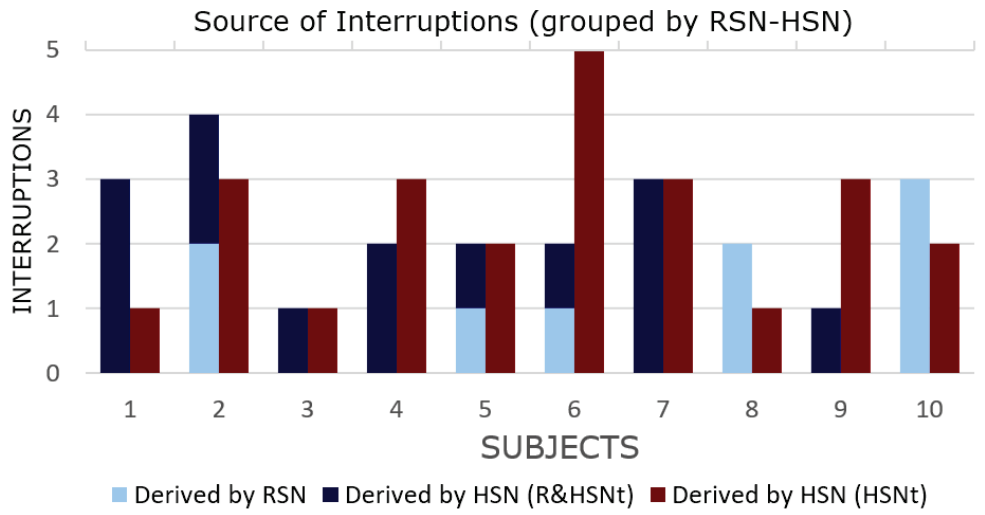
Figure 4.7: Performance comparison of the two tasks in terms of overall task completion time. The task starts with the surgeon assistant's oral instruction "START" and finishes with the oral indication "FINISH". *(HSNt: Human scrub nurse only task, R&HSNt: Robot and human scrub nurse task)*

Table 4.4: Comparison of number of task completion time ($mm : ss$) between the Human scrub nurse only task (HSNt) and Robot and human scrub nurse task (R&HSNt).

Task	Lower	Upper	Mean	SD
HSNt	16:02	37:17	22:35	6:30
R&HSNt	20:18	34:35	26:04	4:50
p-value	0.074			



(a)



(b)

Figure 4.8: (a) Source of workflow interruptions analytically and (b) grouped by Robotic scrub nurse (RSN)- and Human scrub nurse (HSN)-derived. During the HSNt, interruptions are defined as the events of a wrong instrument delivery by the HSN and the delay for instrument delivery by the HSN, which causes interruption of the task by the surgeon (ST) for $> 3s$. During the R&HSNt, the same interruptions are measured (HSN-derived) in addition to the RSN-derived events, namely incorrect instrument selection/delivery and eye-tracking recalibrations. *(HSNt: Human scrub nurse only task, R&HSNt: Robot and human scrub nurse task)*

Table 4.5: Comparison of number of interruptions between the Human scrub nurse only task (HSNt) and Robot and human scrub nurse task (R&HSNt).

Task	Lower	Upper	Mean	SD
HSNt	1	5	2.4	1.26
R&HSNt	1	4	2.3	0.95
p-value	0.84			

4.10.4 Subjective Feedback

Overall feedback was positive with all participants expressing that RSN had potential. All ST expressed looking away from surgical field can affect task flow whilst seven ST highlighted verbal commands may augment the platform. All ST highlighted that a more intuitive RSN platform that can predict the next instrument would improve usability. Three ST expressed a view that RSN would not respond as well as HSN in unpredictable events or emergency. All HSN reported positively about the RSN platform and dismissed any concerns it may replace their role entirely. All HSN agreed the RSN would allow them to perform other tasks more efficiently, especially in big operations where multiple instrument sets and assemblies are required. All HSN reported RSN would have a role in surgery.

4.11 Discussion and Conclusions

A novel robotic scrub nurse, responsive to surgeon gaze, has been proposed. This platform allows the surgeon to visually select an instrument, using an ETG device, pick it up and deliver to complete a task. We tested the RSN with 10 different surgical teams in simulating a common operative scenario with similar theatre staff representation and operative field set up. Table 4.6 summarises the key conclusions of the experimental validation.

Subjectively, RSN was received positively. NASA-TLX data demonstrated no significant difference between HSN vs RSN across perceptions relating to task performance. This affirms a perception of safety towards the platform. ST reported no significant difference across mental or temporal demands in delivering the task. Furthermore, Van der Laan technology acceptance scores were positive across ST vs HSN participants.

Objectively, R&HSNt incurred no significant difference in number of task interruptions, compared with HSNt. RSN related interruptions were attributed predominantly to recalibration where the surgeon visual gaze was not accurately represented on the instrument monitor. The RSN selected the correct instrument in 100% of tasks. In comparison,

HSNt interruptions included incorrect instrument transfers or delays in instrument delivery. HSNt interruptions and resulting errors occurred during HSN disruptions during an ongoing task (instrument count/puzzle). These findings are supported by literature into healthcare interruptions [239], with reported error rates of nearly 3.5% in drug administration when nurses were interrupted, impacting directly on patient safety and related outcomes. In tandem, patient mortality may increase due to scrub nurses shortage [71]. This has big implications in longer and more complicated surgical tasks where more disruptions exist and more personnel is required. This is partly accounted for by the person shifting cognitive load towards the "new" disruption (the puzzle for instance), in so taking longer in performing the primary task or not all [240].

We demonstrated no significant difference in overall experiment duration. Whilst, mean duration is longer in RSN group, this is in part accounted for by recalibration which will be improved in the hardware-agnostic platform modifications through the use of techniques for online ETG displacement compensation.

ST frustration was significant using RSN, although all experiments were completed. Qualitative feedback revealed frustration related to looking away from the operative field to select an instrument. ST proposed verbal commands may enhance the platform. Verbal commands alone may not be reliable due to surrounding noise [234]. In one study surgeons needed to repeat their verbal commands up to three times 30% of the time, using verbally based Penelope^{TM} platform [237].

To alleviate surgeons' frustration, the screen could be replaced with visual projection of instruments close to the operating field. Moreover, we aim to introduce an intuitive RSN, to automatically respond to surgeon instrument selection behaviours, through work flow segmentation and task phase recognition, imitating the HSN's greatest advantage of instrument anticipation [71], as was emphasised in our subjective feedback. Further improvements of the visually aided RSN include enabling real-time recognition and tracking of the surgical instruments and screen position in

space. We also aim to enable the RSN to return the instruments.

A robotic scrub nurse system, visually controlled in a mobile and unrestricted fashion was introduced. This is the first platform of its kind. Subjective feedback was positive. Task duration was similar across RSN vs HSN. Surgeon frustration was highlighted and can be improved by future sophisticated versions. Perception over performance was unchanged.

Table 4.6: Summary of the key conclusions of the experimental validation.

All experiments successfully completed
The RSN selected the correct instrument in 100% of tasks
ST overall task load showed no significant difference (HSNt vs R&HSNt)
ST frustration was significant using RSN
Technology acceptance scores were positive by ST and HSN
No significant difference in overall experiment duration (HSNt vs R&HSNt)
No significant difference in number of task interruptions (HSNt vs R&HSNt)

Chapter 5

A Gaze-controlled Robotised Flexible Endoscope [1] [2]

5.1 Introduction

Flexible endoscopy is a routinely performed medical procedure carried out by means of a flexible endoscope. The endoscope consists of a flexible tube, one or two working channels for flexible instruments to be inserted and a camera and light source at the distal steerable end (tip). The endoscopist can bend the tip left, right, up and down, by rotating with one hand two dials at the handle of the device (Fig. 5.1) and by advancing and rotating the shaft of the endoscope with the other hand. Endoscopy has traditionally been a diagnostic tool, allowing the exploration and the acquisition of tissue biopsies in the upper and lower gastrointestinal (GI)

[1] Content from this chapter was published and is reproduced with permission from: **Intuitive Gaze-Control of a Robotized Flexible Endoscope.** Oude Vrielink T.J.C., González-Bueno Puyal J., Kogkas A., Darzi A., Mylonas G. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018. ©2018 IEEE

[2] This work has been conducted in collaboration with T.J.C. Oude Vrielink and J. González-Bueno Puyal. The author of this thesis integrated the robotised flexible endoscope with the gaze-contingent framework, implemented the robot control, performed the experiments, data processing and validation. J. González-Bueno designed the motor gears and the controller. T.J.C. Oude Vrielink performed the system benchmarking in collaboration with the author.

tract. Despite its wide adoption, diagnostic endoscopy presents several challenges, such as limited dexterity, decreased spatial awareness, loop formation and overall poor ergonomics [241].

Despite these challenges, endoscopes are increasingly adopted in a more therapeutic role in lesion removal in the GI tract, with techniques such as Endoscopic Submucosal Dissection (ESD) slowly becoming more widely adopted. ESD involves an electrosurgical cutting tool introduced via the working channel of the endoscope. The aim is to dissect the submucosa, which is the tissue layer of the GI tract that supports the mucous membrane. Only limited control of the cutting tool is possible by pushing and pulling it inside the endoscope's working channel, while simultaneously steering the endoscope's tip using its control dials with the other hand. The lack of bimanual dexterity and tissue retraction -known as tissue triangulation- are the main reasons behind the technical complexity of ESD. Hybrid techniques have been investigated for ESD, offering only marginal improvement [242]. Furthermore, endoscopes with augmented functionality have been proposed. These include systems that can be manually controlled, such as the *Cobra* (USGI Medical, USA) [243] and *Endosamurai* (Olympus, Japan) [244], or robotically controlled, such as the *CYCLOPS* [245], the *STRAS* [246], the *MASTER* [247] or the *Flex* (Medrobotics, USA) [248]. These devices introduce externally controllable instruments at the tip of the endoscope, allowing bimanual dexterity and tissue triangulation. Robotic actuation is used to control the additional degrees-of-freedom (DOF) offered by the robotic attachments. However, control of the host flexible endoscope is still manual for some of those systems.

Despite the advantages of augmented endoscopes, the increased DOF require more operators. In this type of situation an operator is needed to advance and steer the endoscope, while at least another operator manipulates the instruments manually or by using tele-manipulators. This introduces obvious new challenges, such as suboptimal collaboration, communication failures, space constraints and collisions, as well as increased

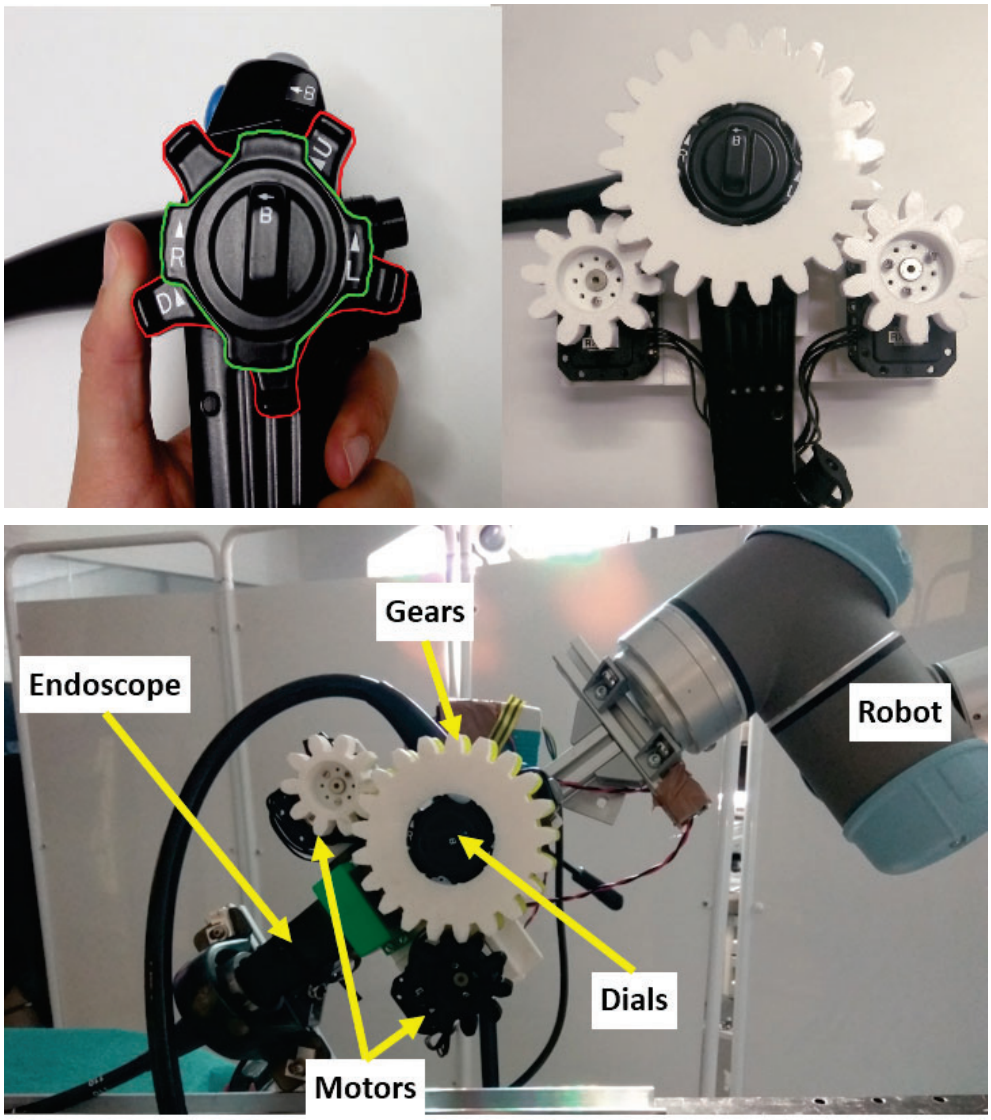


Figure 5.1: **Top left:** Flexible endoscope handle illustration. The larger diameter dial (red) controls the up/down movement, while the smaller dial (green) controls the right/left movement of the tip of the endoscope. **Top right:** The motorised system with gears placed on the dials. **Bottom:** The fully robotised system.

cost [249] [250]. Therefore, a more intuitive human-endoscope interfacing approach is required to overcome the increased complexity introduced by augmented endoscopes. We hypothesise that by motorising the flexible endoscope and by using eye-gaze control through eye-tracking, we can decrease the complexity introduced by the augmented endoscopes.

Previous studies explore the robotisation of standard flexible endoscopes. In Kume et al. work, it is controlled with one haptic device, and used for releasing the endoscopist's other hand to control the *MASTER* device [251]. Similarly, in [252] the authors use a robotised flexible endoscope to perform automatic steering of the shaft for lumen centralisation. It is important, however, to keep in mind the importance of intuitive user interfaces to increase the efficiency of the flexible endoscope steering, as underlined by [253]. A study performed by Dik et al. shows that during colonoscopy there is high correlation between the total gaze time spent in an area and its diagnostic interest [254]. Work presented in [62] [255] [21] further highlights how eye-gaze information can be used to augment and improve surgical practice. Finally, Noonan et al. control an articulated mechatronic laparoscope using 2D gaze and fixations as commands [24].

Based on this evidence, we use eye-gaze and head orientation to control a robotised flexible endoscope. We prove that this approach can improve diagnostic endoscopy through intuitive and effortless control of the endoscope. In this study, we propose a fully motorised gaze-controlled system for non-restricting, free-view flexible endoscopy. The feasibility and comparison against traditional hand control are assessed. Simulated examination of the upper GI tract (UGIt) is performed by expert and novice users.

5.2 System Overview

The system is developed in *ROS* with C++, to facilitate the hardware agnostic aspect of the framework. The work presented here allows a user to remotely control the endoscope movements without handling the device

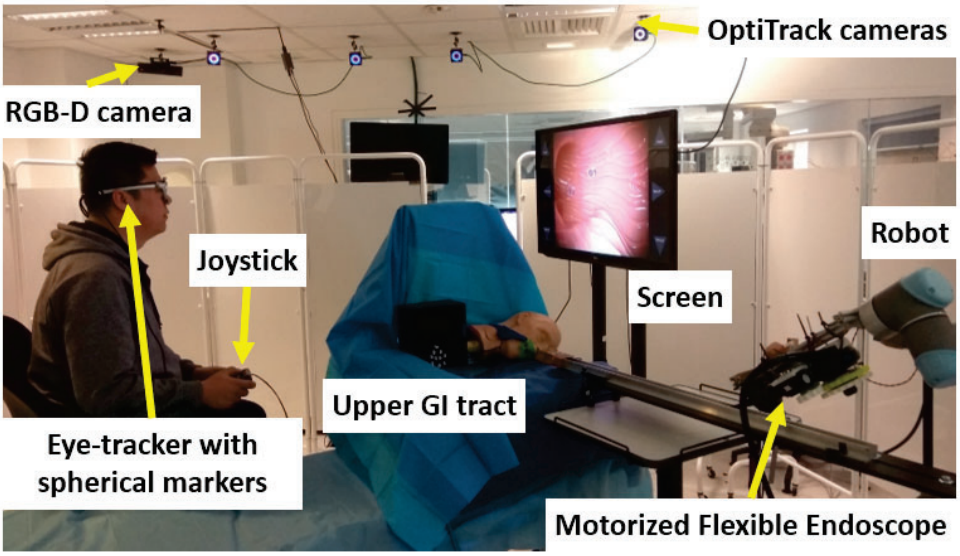


Figure 5.2: The system setup: The motorised flexible endoscope is mounted on the UR5 articulated robot arm. The Microsoft Kinect RGB-D cameras are employed for the 3D reconstruction of the operating theatre. The OptiTrack motion capture system provides the pose of the eye-tracking glasses in space. The user controls the endoscope with natural gaze, head pose and a joystick.

(Fig. 5.2). A flexible gastroscope is attached to an articulated robotic arm, mounted onto a rail and placed on top of a surgical table (Fig. 5.1). The dials used to control the distal tip steering are motorised using two 3D printed gears and two motors, controlled by a gaze-contingent closed loop velocity controller. Gaze on the screen is estimated based on a 3D gaze reconstruction framework we developed in chapter 3 with the synergy of conventional wearable eye-tracking glasses, a motion capture system and fixed in space RGB-D cameras for real-time 3D reconstruction of the environment. An articulated robotic arm controls the endoscope shaft rotation and insertion/retraction. The former is controlled with the head sideways rotations and the latter with a joystick handle, which also allows features such as system pause and automatic retroflexion. Audio feedback is provided when the user enables rotation, pauses the system or enables retroflexion.

5.3 User Interface

The functionalities of the system are summarised as following:

- Distal tip angulation is controlled by eye gaze tracking; the endoscope follows the direction of gaze on the screen.
- Axial rotation of the endoscope is controlled by head movement, tilting the head to the left or right.
- Insertion and withdrawal are controlled by head movement, towards or away from the screen. The same functionality is also available through a joystick on the control pad.
- Retroflexion[3] is controlled by pressing a button on the control pad.
- Endoscope movement is "frozen" by a button on the control pad, or when the endoscopist looks away from the screen.
- Visual and audible alert during endoscope insertion or withdrawal.
- Audible alert "left" or "right" during endoscope rotation.

5.4 Equipment

For eye-tracking, the SMI (SensoMotoric Instruments GmbH) eye-tracking glasses (ETG) are used. For RGB-D sensing, the Microsoft Kinect v2 is used and for head pose tracking the OptiTrack motion capture system (MCS) with four Prime 13 cameras. The robot arm is a UR5 (Universal Robots A/S), a 6 degrees of freedom (DOF) collaborative robot with a reach radius of up to $850mm$, $\pm 0.1mm$ repeatability, ± 360°joint ranges, maximum $5kg$ pay-load and weighing $18.4kg$. Two Dynamixel RX-24F motors were employed for the motorisation of the endoscope along with the USB2Dynamixel controller to interface them with the computer.

[3] a technique where the endoscope bends backwards

For the endoscopic task view, a 42" LG screen with 1920 × 1080 px resolution is used. The endoscope is a flexible gastroscope (Karl Storz 13801 PKS) with a 9.8mm diameter and a 1.1m long flexible shaft. The control pad consists of a 2-axis joystick controller connected to an Arduino UNO microcontroller, which streams serial data to the PC through USB interface.

For the current implementation, a Windows 10 PC is used for acquiring and streaming the ETG and MCS data and a Linux PC with Ubuntu 14.04 is used for all other modules. The Linux PC runs on Intel Xeon Processor, NVIDIA GTX 580 1.5 GB, 16 GB RAM.

5.5 Data Acquisition

The ETG's API by SMI provides the scene video and eye related data. The MCS provides with spherical marker's positions and the 6 DOF pose of a user defined rigid body's geometry in the motion capture system coordinate frame (MCS CS). The RGB frames and the 2D gaze information by the ETG and the 6 DOF rigid body pose of the ETG are streamed timestamped through UDP to the Linux PC, where they are decoded. For RGB-D camera the Kinect bridge [210] is used to acquire sensor data and convert them into ROS compatible messages.

5.6 Offline Calibration

ETG's scene/RGB camera. The intrinsic camera parameters of the eye-tracker scene camera are calibrated using a chessboard.

Eye-tracking. Eye fixations were mapped to specific points in the ETG's scene camera plane by asking the user to fixate in 9 predefined points in their FOV, keeping their head pose fixed.

Kinect sensor. The RGB and IR cameras intrinsic parameters and their rigid transformation were estimated using the calibration process provided by [210].

MCS. The spatial correlation of the four MCS cameras in the MCS CS is calibrated through the OptiTrack Motive software package. It involves moving a calibration wand consisted of spherical reflective markers of known geometry while cameras are recording the sequence. After the cameras' calibration, the ETG rigid body is defined by six spherical reflective markers with fixed asymmetric geometry mounted on the ETG.

WCS – MCS CS registration. The rigid transformations between the ETG rigid body – ETG scene camera (Y) and MCS CS – WCS (X) is estimated by solving hand-eye calibration problem $AX = YB$ [35]. This involves capturing 6 DOF poses both of the ETG rigid body in the MCS CS and the ETG scene camera in the WCS, simultaneously. The first is provided by the MCS API. The latter is calculated employing EPnP [217], given the 2D-3D correspondences of an asymmetric checkerboard. The 2D correspondences are observed by the ETG scene camera and the 3D by the RGB-D sensor's RGB and IR camera.

Screen position in the WCS. The screen corners' 3D coordinates in the WCS are manually selected on the Kinect RGB image. The depth correspondence is estimated using the Kinect calibration and the 3D points generated.

5.7 Motorisation

The system setup is shown in Fig. 5.2. An attachable actuation module was developed for the endoscope, allowing its immediate unplugging as a safety measure and permitting conversion from robotic to manual control. Gears were designed to fit the dials of the endoscope and to couple two motors. The designed mechanism is shown in Fig. 5.1.

Reilink et al. [256] carried out torque and speed measurements using a Pentax EG-2930K gastroscope and identified the following operational requirements:

- The maximum torque needed with the larger dial is approximately $0.4N \cdot m$.

Table 5.1: Gears parameters

Parameter	**Endoscope gears**	**Motor gears**
Pitch radius	$46mm$	$22mm$
Number of teeth	23	11
Pitch	$4mm$	
Dedendum	$5mm$	
Adendum	$4mm$	
Clearance	$1mm$	
Pressure angle	$20deg$	

- The required velocity is $15rpm$. Faster movements were determined as not useful and resulting in a loss of spatial orientation.

Based on these specifications, the gears were designed to be able to provide $1.0N \cdot m$ torque and $50rpm$ angular velocity, ensuring sufficient power margin. Table 5.1 lists the selected design parameters for the gears, which can be applied for other standard flexible endoscopes. For the current application, the transmission ratio between the motors and the tip bending stands as approximately 0.36, meaning that a complete turn of a motor will steer the tip approximately 130 degrees in one direction.

5.8 System Functionalities

5.8.1 Distal Tip Angulation

As endoscopy requires the use of a screen to visualise the endoscopic video, free-view gaze interaction with the screen is achieved with the real-time framework presented in chapter 3 (Fig. 5.2). The framework's core functionality is to provide the user's 3D point of regard in space.

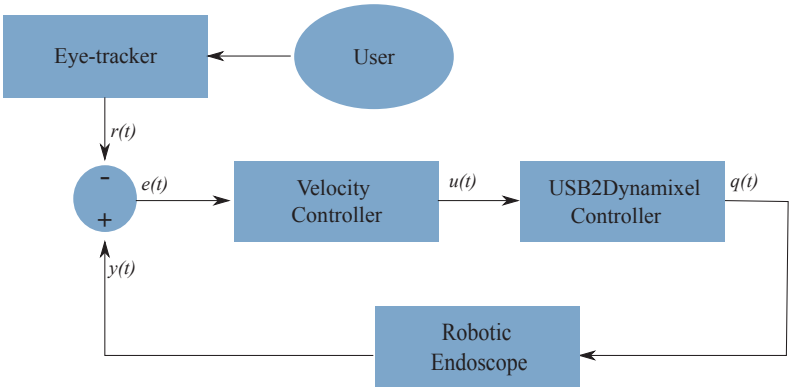


Figure 5.3: Gaze control system diagram where $r(t), y(t)$, $e(t), u(t)$ and $q(t)$ correspond to the reference signal, the feedback signal, the error signal, the desired velocity and the voltage input respectively.

It relies on estimating the pose of the eye-tracker scene camera in the world frame, and tracing the gaze ray provided on the head frame of reference, onto the 3D reconstructed space. For the work presented here, the head pose is estimated with the employment of the OptiTrack motion capture system. Spherical markers are mounted on the ETG to form an asymmetric rigid body and allow OptiTrack to provide its unique 6 DOF pose in space. The rigid transformations between the rigid body—ETG scene camera and MCS CS—WCS are calibrated. By using the hybrid macro/micro-scale model presented in chapter 3, the mode (macro or micro) and the 2D screen fixation (for micro mode) is provided as output.

A closed-loop system was implemented to control the robotic actuation, integrating the received gaze information and the motors' controller. This approach aims to use the natural gaze of the user to facilitate the task of manipulating the endoscope. Whenever the user directs his/her gaze away from the centre of the screen coordinates $y(t)$, the distance $e(t)$ between the gaze point $r(t)$ and the centre of the screen is computed. This error signal is used as an input for the controller, as depicted in Fig. 5.3.

A velocity based control was applied, where the desired velocity $u(t)$ of the motors is computed by using a PID controller. For use with gaze control an overshoot has shown to be confusing, and a fast response time

was more intuitive than a small steady-state error. The desired velocity derived from the controller is used by the USB2Dynamixel to compute the corresponding voltage input $q(t)$ for the motors and steer the camera accordingly.

Limits in terms of the velocity and torques exerted by the motors are applied by means of the USB2Dynamixel controller, in order to prevent any damage on the tissue or the endoscope.

5.8.2 Camera Rotation

The rotation of the endoscope is achieved by rotating the end-effector of the robot with a constant speed. It is initiated with the rotation of the user's head on ETG scene camera's z-axis, above a predefined angular threshold. The head orientation reference is defined at the beginning of the experiment. When the rotation threshold is exceeded, an audible alert "left" or "right" is activated. Similarly, when the head rotation reverts back to the idle angular range, an audible alert "straight" is enabled.

5.8.3 Insertion/Withdrawal

Insertion and retraction of the endoscope is implemented with the linear movement of the robot with constant velocity (20 mm/s). It is triggered by a joystick (up/down respectively) connected to an Arduino Uno which streams data to the system PC. By moving the joystick upwards, the robot starts moving inwards, a double arrow is drawn on the top of the screen (Fig. 5.4(b)) and an audible alert "in" is activated. Respectively, by moving the joystick downwards, the robot starts moving outwards, a double arrow is drawn on the bottom of the screen (Fig. 5.4(c)) and an audible alert "out" is activated.

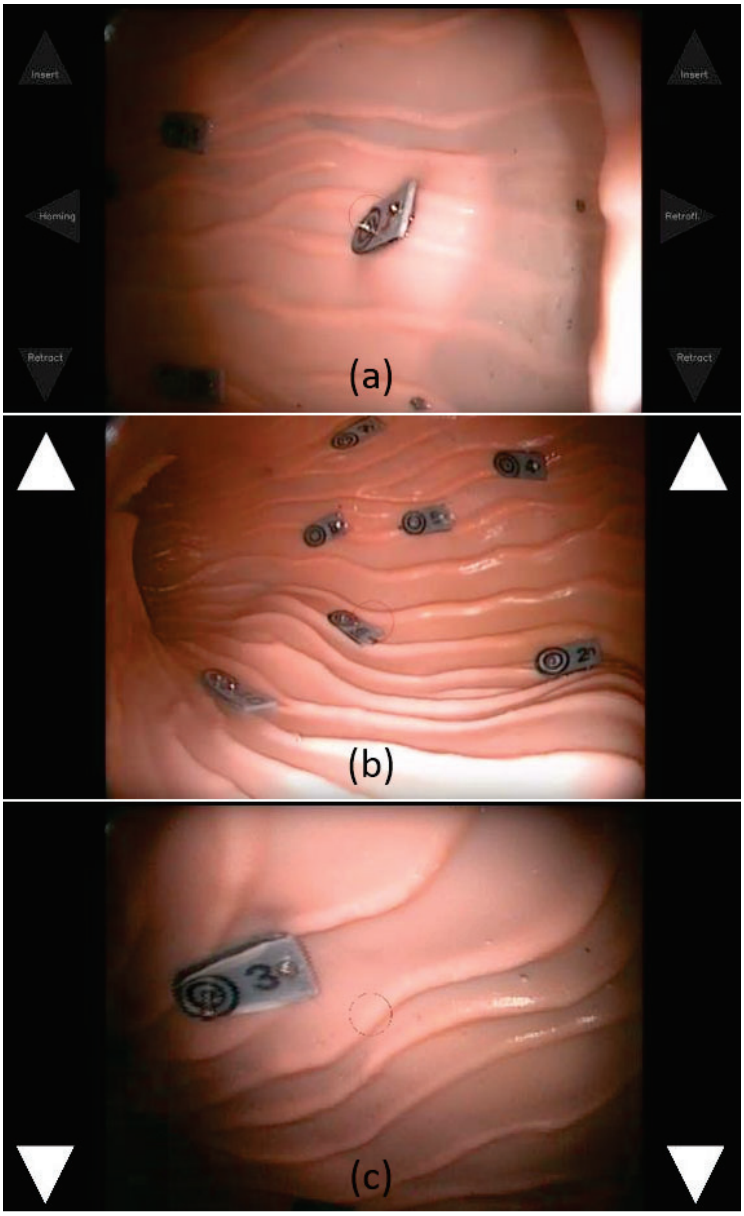


Figure 5.4: The graphical user interface (GUI) of the system. (a) The view while the system is paused. The user can see the joystick options besides the endoscope camera view. Up arrow for endoscope insertion, down arrow for endoscope retraction and right for retroflexion. (b) The view of the screen when insertion and (c) retraction are activated.

5.8.4 Retroflexion

Retroflexion of the distal tip is activated by holding the joystick to the right for $1s$. Audible feedback "retroflexion" notifies the user of this choice.

5.8.5 System Pause

A 9-point user calibration is performed for each user in order to map user-specific gaze-direction dependent ocular landmarks to unique coordinates on the his/her head frame of reference. The system is paused (motors and robot remain idle) when gaze-points correspond to positions out of the screen, or when invalid eye-movements are detected (i.e. due to blinking, bad calibration, etc.).

System paused/unpause is also enabled by pushing the joystick handle. Audible feedback "pause" or "unpause" is activated when the system is paused intentionally. This functionality serves not only safety purposes, but also facilitates seamless exploration of the endoscopic view, by not allowing the camera to move following the eye movements.

5.9 Validation Method

5.9.1 Experimental Design

Fig. 5.7 shows the experimental setup. A screen displays the endoscopic video. Users were positioned in front of the monitor and used their natural gaze tracked by the ETG, their head pose and a joystick, to manipulate the robotised flexible endoscope.

The evaluation process consists of two tasks:

- a navigation task in a spherical cavity (SPHt)
- a simulated diagnostic gastroscopy in an upper gastrointestinal tract (UGIt)

and two modalities:

- Gaze-contingent Control (GC)
- Hand Control (HC)

For this purpose, a plastic sphere was used for the SPHt and a head phantom, a silicon tube to simulate the oesophagus and a stomach phantom were used for the UGIt.

A set of ten differently numbered targets was placed on the interior of both the sphere and the stomach phantom, as can be seen in the screen display of Fig. 5.5. One of the sphere targets (number ten) was placed in challenging position, namely the user would need to reach it by retroflexion. The targets, the sphere and the phantom were maintained in a fixed position throughout the experiments to eliminate any possible variation between participants. For the SPHt, each subject was asked to locate and fit the targets in ascending order and horizontal orientation within a circle drawn at the centre of the screen (Fig. 5.5(a)). For the UGIt, each subject was instructed to intubate the oesophagus and then locate and fit the targets in ascending order within a circle drawn at the centre of the screen (Fig. 5.5(b-e)). This task allowed for the evaluation of a simulated clinical scenario, in which an endoscopist examines possible malignancies in a dexterous manner.

5.9.2 Participants

Sixteen subjects were included in the study; eight novices (non-endoscopists) and eight expert endoscopists. Participants performed the task both in a traditional manner (HC), where they controlled the endoscope with their hands, and with the proposed system (eye-gaze, head pose and joystick – GC). Which system was used first by each participant was randomised, in order to reduce the learning effect bias.

After informed consent, the subjects were taken through the experimental setup, starting either with the gaze- or hand-control setup, in randomised order, and given time to familiarise themselves with it.

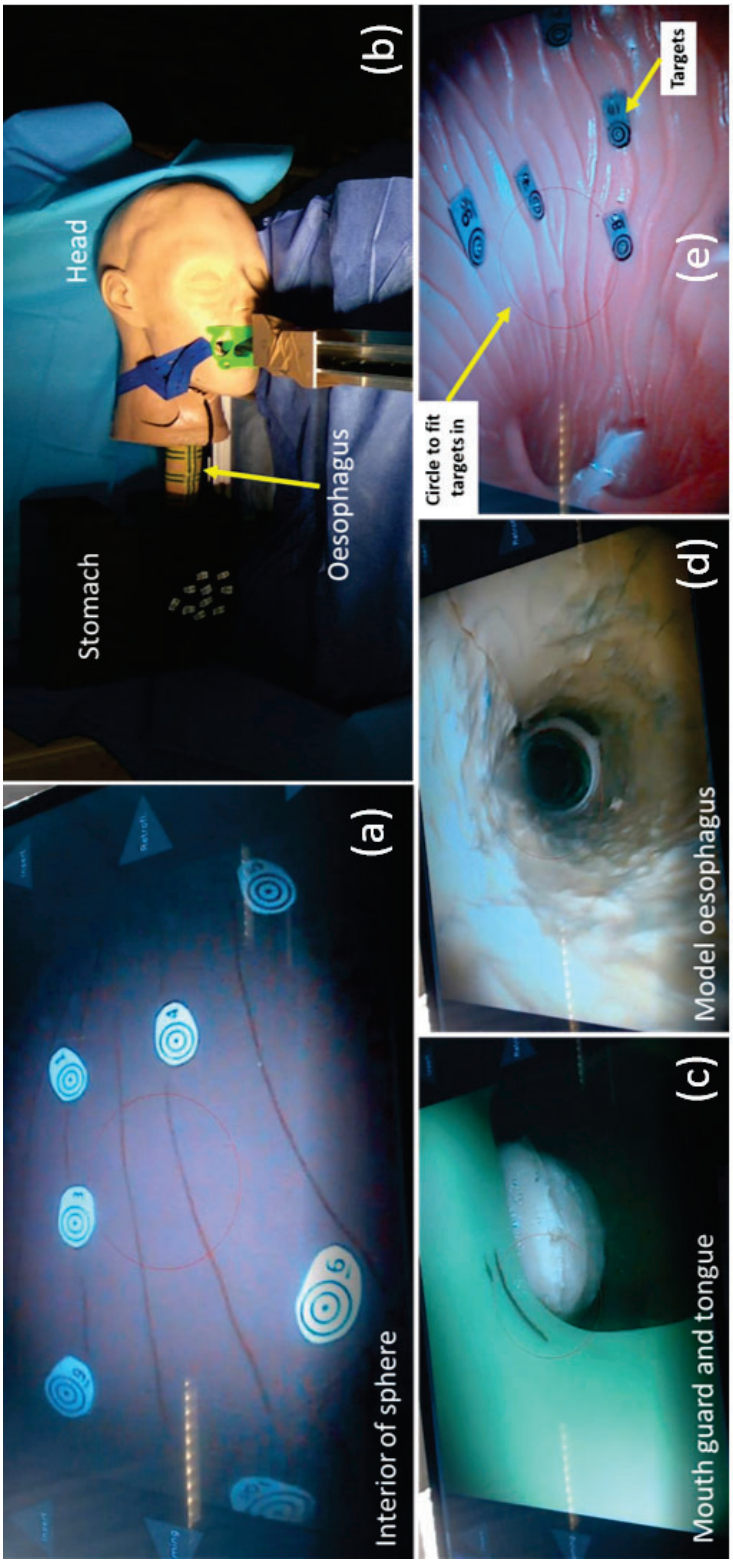


Figure 5.5: The experimental setup: (a): View of the interior of the sphere used for the user study. (b) The Upper GI tract (UGIt) silicon phantom, comprising the head, oesophagus and stomach. (c) Endoscope view of the insertion point, oesophagus (d) and the silicon stomach (e).

5.9.3 Subjective Validation

Task Load

After each task, the participants were asked to complete a *NASA-TLX* (System Task Load Index defined by NASA) questionnaire. The scale assesses the mental, physical and temporal demand, overall performance, frustration levels and effort during the task. More details on the NASA-TLX system in section 4.9.3.

Technology Acceptance

Technology usability and satisfaction feedback was collected immediately following the R&HSNt using the *Van Der Laan* acceptance scale [238]. The scale consists of five usefulness metrics (useful/useless, good/bad, effective/superfluous, assisting/worthless, raising alertness/sleep-inducing) and four satisfaction metrics (pleasant/unpleasant, nice/annoying, likeable/irritating, desirable/undesirable). Each item was on answered a 5-point semantic differential from -2 to +2.

Ergonomics Assessment

To compare the user preferences in terms of ergonomic factors, workflow and comfort, the participants completed a Likert questionnaire consisting of the following questions:

- Gaze control is more comfortable
- Gaze control is easier to learn
- Gaze control is less stressful
- Gaze control doesn't interrupt the task flow
- Gaze control doesn't cause neck discomfort
- Gaze control doesn't cause eye strain
- Gaze control doesn't cause me to become fatigued

5.9.4 Objective Validation

Benchmarking

A benchmarking study was used to measure the technical performance of the system. The first technical evaluation was the characterisation of the system relating a given input motor's position to the tip response. It is important to assess whether the transformation from input position to output tip position can be simplified to a linear relationship. In case of a non-linear transformation, the response to input will depend on the the endoscope tip's position within the workspace. For the user, this will result in areas in which the system is more responsive to the gaze control input than others. The transformation was assessed by moving the motors with the measured input θ_x and θ_y and evaluating a metric of the resulting orientation of the endoscope's tip β. The value β is an angle calculated using the camera orientation vector V_n and the base-frame vector e_z:

$$\beta = cos^{-1}\left(\frac{V_n \cdot e_z}{\|V_n\|\|e_z\|}\right) \tag{5.1}$$

where V_n and e_z are the normal vectors to the plane defined by the three markers on the tip and base, respectively (Fig. 5.6). An optical tracking system (2x Prime 13 OptiTrack Cameras, NaturalPoint, Inc.) is used to track the 3D position of the endoscope tip during these experiments (Fig. 5.7). Three passive optical markers are attached to the tip, using a lightweight nylon mount. Additionally, three markers are placed just before the flexible part of the tip, to act as a base frame.

A second technical validation was performed to characterise the controller. To make the evaluation consistent and cancel out any effects caused by voluntary and involuntary eye movements, this is performed without the gaze input from the user. Instead, a reference step input $r(t)$ is based on the position in the screen of an optical marker placed within the field of view of the endoscope, achieving visual servoing. The optical marker was placed in 12 spatially distributed positions. An adjustable

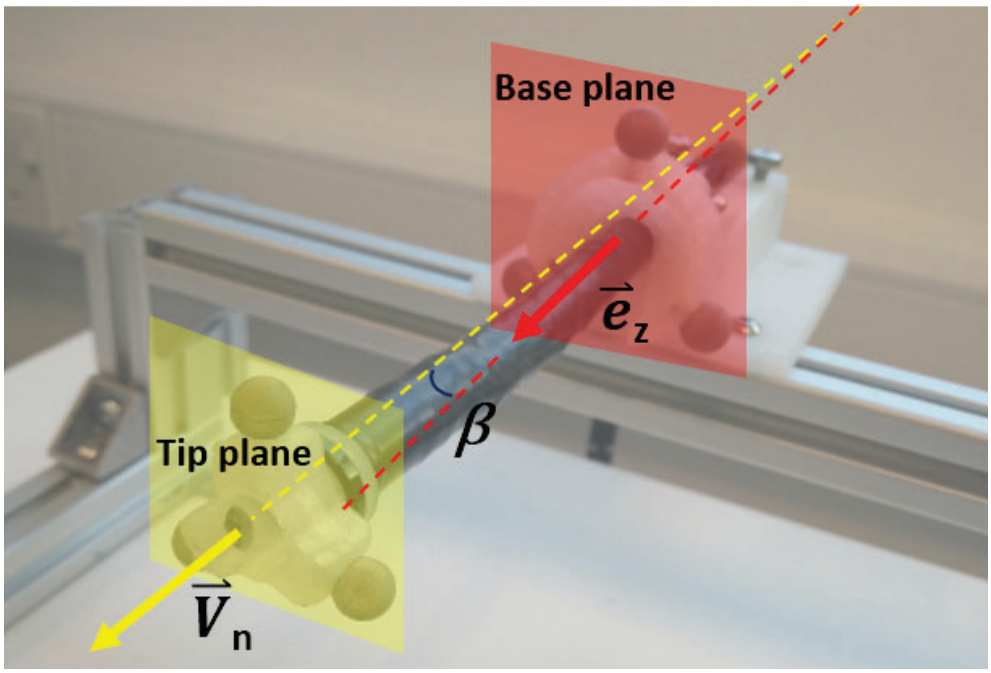


Figure 5.6: Definition of the angle β. The vectors V_n and e_z are the normal vectors to the plane defined by the three markers on the tip and base respectively.

rig to which the optical marker is attached is used to change the position throughout the field of view of the endoscope. The endoscope's light source was used to increase the intensity of the passive optical marker. In order to simulate the 2D coordinates of the gaze point, the marker was segmented from the grayscale image by using a binary threshold function combined with a circular morphological filter (OpenCV 3.2). Erosion and dilation were then applied to filter out noise. The setup, including the view from the endoscope, is shown in Fig. 5.7.

Performance

An objective evaluation was carried out for each type of control, assessing the overall completion time of each task.

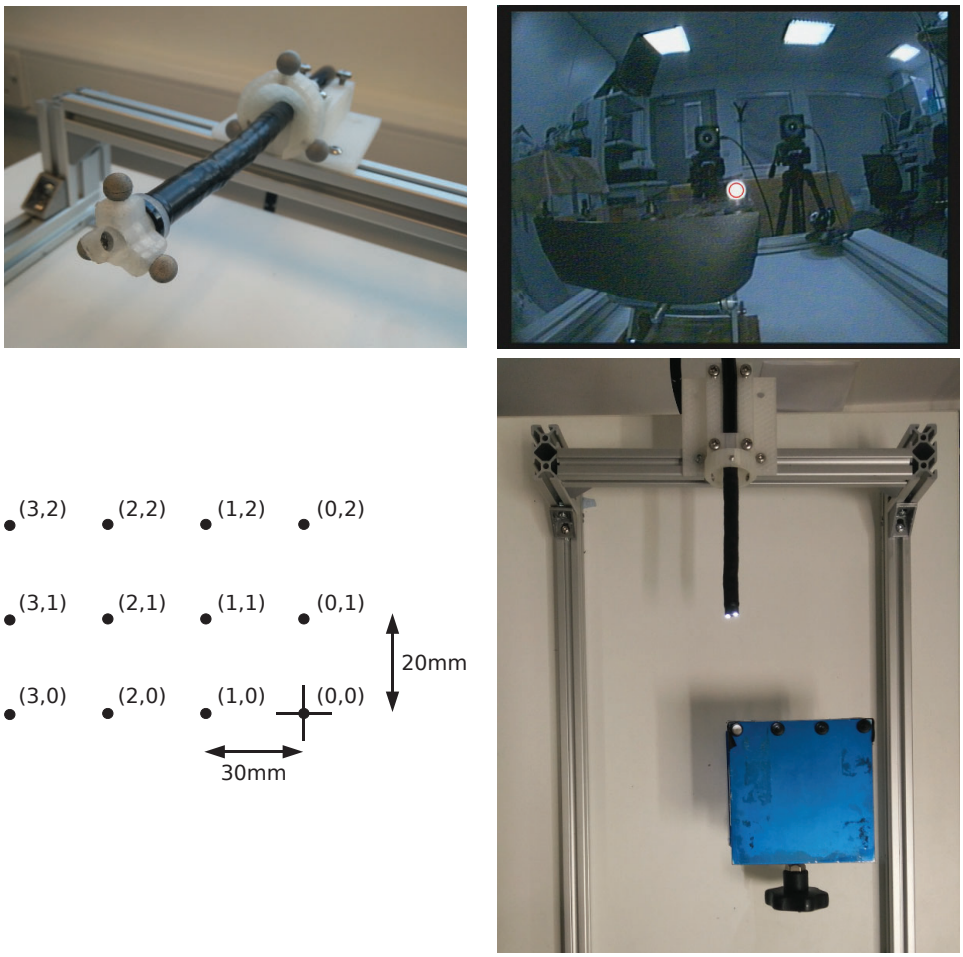


Figure 5.7: **Top Left:** Setup used for optical tracking of the endoscope tip for different motor input. Three passive optical markers are placed at the tip of the endoscope, and another three are placed at the base of the bending tip. **Top Right:** View from the endoscope during the visual servoing experiments. The passive optical marker is encircled in red. The *(x, y)* pixel position of the centre of the circle is used as input during these experiments. **Bottom Left:** Data were collected for 12 spatially distributed marker positions. The spatial distribution is based on a XY plane $75mm$ in front of the endoscope. Point *(0,0)* is in the centre of the endoscope's video at homing position, and is shown in the endoscopic image above. **Bottom Right:** The top-view of the setup. The blue adjustable platform is used to change the on-screen Y position of the marker. To change the X position, the marker is placed in different locations on the platform.

5.10 Results

5.10.1 Data Analysis

The comparisons demonstrated in the following sections were conducted using within-subjects analysis when comparing:

- Task completion time of endoscopists with HC vs GC
- Task completion time of novices with HC vs GC
- NASA-TLX scores of endoscopists with HC vs GC
- NASA-TLX scores of novices with HC vs GC

Between-subjects analysis was conducted when comparing:

- Task completion time with HC by endoscopists vs novices
- Task completion time with GC by endoscopists vs novices
- NASA-TLX scores with HC by endoscopists vs novices
- NASA-TLX scores with GC by endoscopists vs novices
- Van der Laan's scores by endoscopists vs novices
- Ergonomics assessment scores by endoscopists vs novices

For within-subjects analysis, the Shapiro-Wilk test for normality of the paired differences was performed, followed by paired-samples t-test when the test was successful and no outliers were detected. In case of non-normal distribution of the differences or the presence of outliers, the Wilcoxon signed-rank test was used.

For between-subjects analysis, the Shapiro-Wilk test for normality of the samples was performed, followed by independent-samples t-test when the test was successful. In case of non-normal distribution of any of the two samples, the Mann-Whitney U test was applied.

For all types of statistical analysis tests, a p-value <0.05 was considered significant.

5.10.2 Subjective Data

Task Load

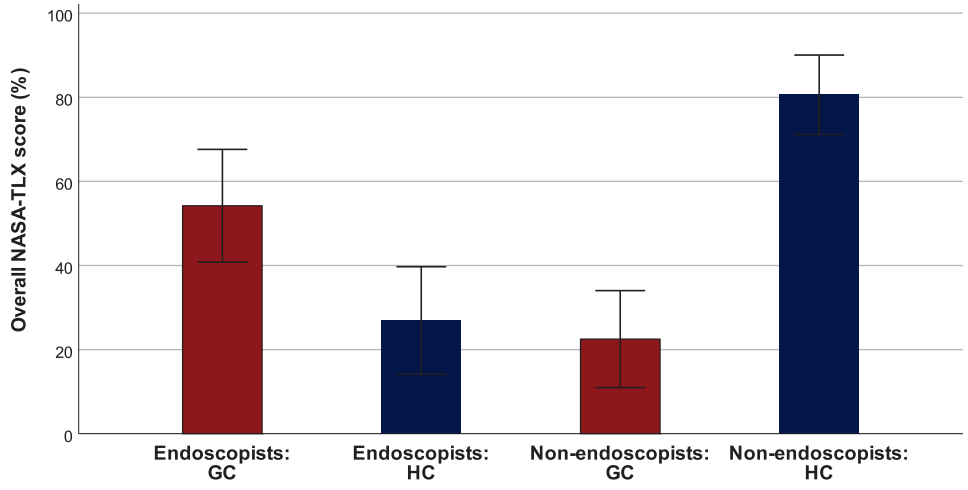
The NASA-TLX scores overall and per category are depicted in Fig. 5.8 and Table 5.2. Endoscopists reported significantly higher workload using the gaze control over hand control (54.2±16 vs 26.9±15.3, p=0.012), where perception over physical demand and frustration remained unchanged. Novices reported significantly higher workload using the conventional control (80.6±11.3 vs 22.5±13.8, p<.001). Gaze control demonstrated higher task load for endoscopists compared to novices (p=0.001), whilst hand control was reported as more demanding overall for novices compared to endoscopists (p<.001).

Technology Acceptance

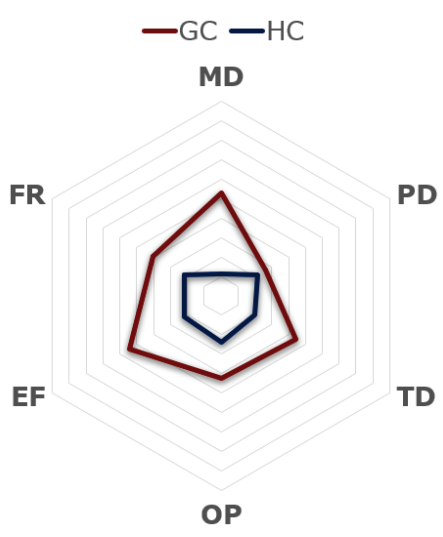
The Van der Laan scores overall and per category are depicted in Fig. 5.9 and Table 5.3. The endoscopists group reported usefulness score of 0.56±0.83 vs 1.43±0.51 by novices, p=0.065 (Table 5.4). Satisfying score shows statistically insignificant difference (p=0.222) for endoscopists vs novices (0.8±0.87 and 1.44±0.68 respectively). Novices showed greater preference in all individual metrics over endoscopists. Nevertheless, overall responses were positive in both groups across all technology acceptance metrics.

Ergonomics Assessment

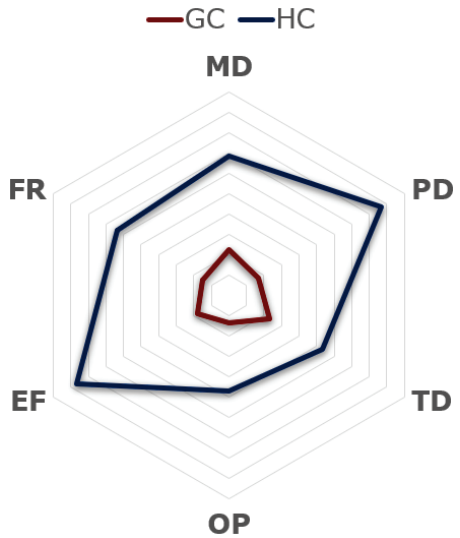
The perception over ergonomics between endoscopists and novices shows significant difference in comfort, ease of learning, stress and flow metrics (Table 5.5), with endoscopists being mostly negatively and novices mostly positively biased. Analytical results in Fig. 5.10 show endoscopists deeming gaze control less comfortable (83%), more stressful (57%), interrupting the task flow (85%), causing neck discomfort (57%), eye strain (86%) and fatigue (71%), while being neutral to the ease of learning (57%). Instead, novices perceive the gaze control as more comfortable (100%), easier to



(a)



(b)

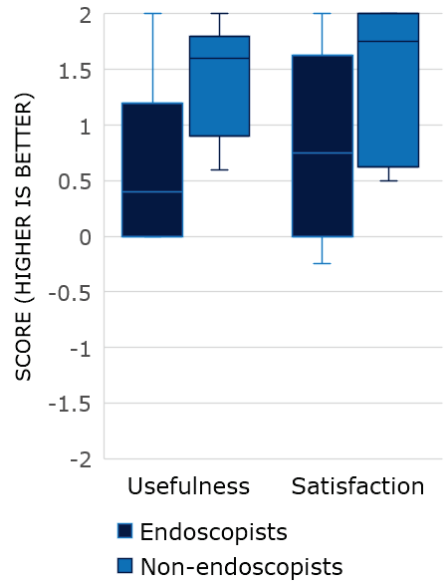


(c)

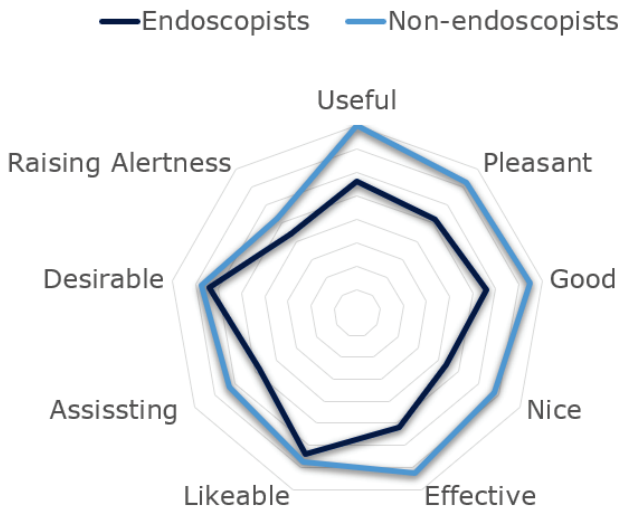
Figure 5.8: (a) Overall NASA-TLX score and analytical results (*MD, PD, TD, OP, EF, FR*) for (b) endoscopists and (c) novices. NASA-TLX values range between 0 and 100, with higher values indicating higher task load. *(HC: Hand Control, GC: Gaze-contingent Control)*

Table 5.2: Mean, standard deviation (SD) and comparison of NASA-TLX scores by endoscopists and novices on Gaze-contingent Control (GC) and Hand Control (HC).

	Endoscopists		Novices		Comparisons				
	GC (1)	HC (2)	GC (3)	HC (4)		(1) vs (2)	(3) vs (4)	(1) vs (3)	(2) vs (4)
MD	53.1[23.3]	11.3[8.3]	22.5[18.3]	68.8[26.8]	p	**0.001**	**0.002**	**0.028**	**<.001**
PD	26.3[18.7]	21.9[18.0]	16.9[16.2]	86.9[9.9]	p	0.655	**<.001**	0.234	**<.001**
TD	44.4[12.2]	20[12.0]	23.1[11.9]	53.1[26.8]	p	**0.004**	**0.003**	**0.005**	**0.008**
OP	42.5[25.8]	23.8[26.2]	13.8[7.1]	46.9[23.1]	p	**0.042**	**0.004**	**0.023**	0.115
EF	54.4[19.4]	21.9[15.7]	18.1[9.1]	86.9[9.9]	p	**0.013**	**<.001**	**<.001**	**<.001**
FR	40.6[21.0]	21.9[17.8]	15[13.4]	63.8[32.3]	p	0.076	**0.012**	**0.012**	**0.006**
Overall	54.2[16.0]	26.9[15.3]	22.5[13.8]	80.6[11.3]	p	**0.012**	**<.001**	**0.001**	**<.001**



(a)



(b)

Figure 5.9: (a) Overall Van der Laan's technology acceptance score by endoscopists and novices and (b) analytical results. The usefulness scale derives from the average of *useful/useless, good/bad, effective/superfluous, assisting/worthless, raising alertness/sleep-inducing* metrics and satisfaction scale derives from *pleasant/unpleasant, nice/annoying, likeable/irritating, desirable/undesirable* metrics. The scale range between -2 and +2, with higher values indicating positive bias.

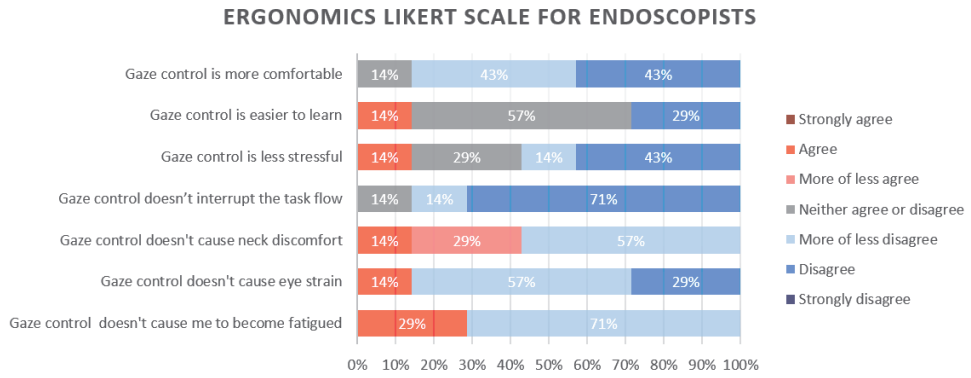
Table 5.3: Van der Laan's technology acceptance scores between endoscopists and novices. The scale range between -2 and +2, with higher values indicating positive bias on the specific attribute. Mean and standard deviation (SD) values are reported.

	Endoscopists	Novices
Useful	0.80 [1.10]	2.00 [0.00]
Good	0.80 [1.10]	1.75 [0.46]
Effective	0.60 [0.89]	1.63 [0.52]
Assisting	0.40 [1.14]	1.13 [1.36]
Raising Alert.	0.20 [1.10]	0.63 [0.92]
Pleasant	0.60 [0.89]	1.63 [0.52]
Nice	0.20 [1.30]	1.38 [0.74]
Likeable	1.20 [1.30]	1.38 [0.92]
Desirable	1.20 [0.84]	1.38 [0.74]
Usefulness	0.56 [0.83]	1.43 [0.51]
Satisfaction	0.80 [0.87]	1.44 [0.68]

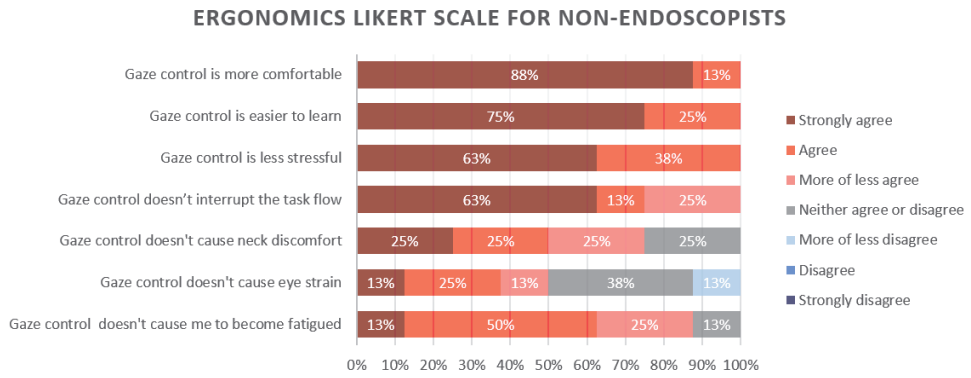
learn (100%), less stressful (100%), not interrupting the task flow (100%), not causing neck discomfort (75%), eye strain (51%) and fatigue (88%).

Table 5.4: Van der Laan's technology acceptance scores comparison between endoscopists and novices.

	Usefulness	Satisfaction
p-value	0.065	0.222



(a)



(b)

Figure 5.10: Likert scale results of ergonomics assessment for (a) endoscopists and (b) non-endoscopists.

Table 5.5: Comparison across Likert scale responses between endoscopists and novices. p-values are reported.

Comfortable	**<.001**
Easier Learning	**<.001**
Stressless	**<.001**
Uninterrupting	**<.001**
Neck Discomfort	0.337
Eye strain	0.115
Fatigue	0.368

5.10.3 Objective Data

Benchmarking

The mapping of the motor inputs to the endoscope tip position and orientation is shown in Fig. 5.11. The surface is the 2^{nd} order polynomial fitting of the data:

$$\beta(\theta_x, \theta_y) = a_0 + a_1\theta_x + a_2\theta_y + a_3\theta_x^2 + a_4\theta_x\theta_y + a_5\theta_y^2$$

The parameters a_i are found using MATLAB's (R2017a) *fit()* function and are shown in Table 5.6. The RMSE for the entire dataset, and for each quadrant are shown in Table 5.7. Higher order polynomials did not improve the RMSE fitting.

The visual servoing experiments showed that the control system does not exhibit any overshoot (Fig. 5.12). This is important as during user control overshoot might result in unexpected behaviour from the user. As the previous benchmarking showed similar behaviour for all quadrants of the endoscope's image, this experiment is only performed in the top-left quadrant (Q1, as defined in Table 5.7).

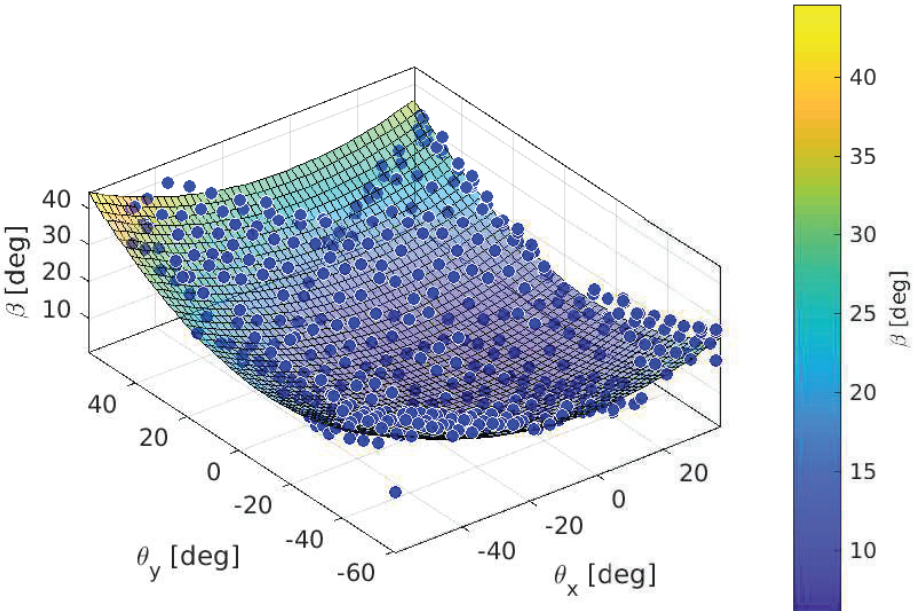


Figure 5.11: The mapping from input motor angles θ_x and θ_y to the tip angle position β. The surface is fitted by a 2^{nd} order polynomial, with parameters shown in Table 5.6.

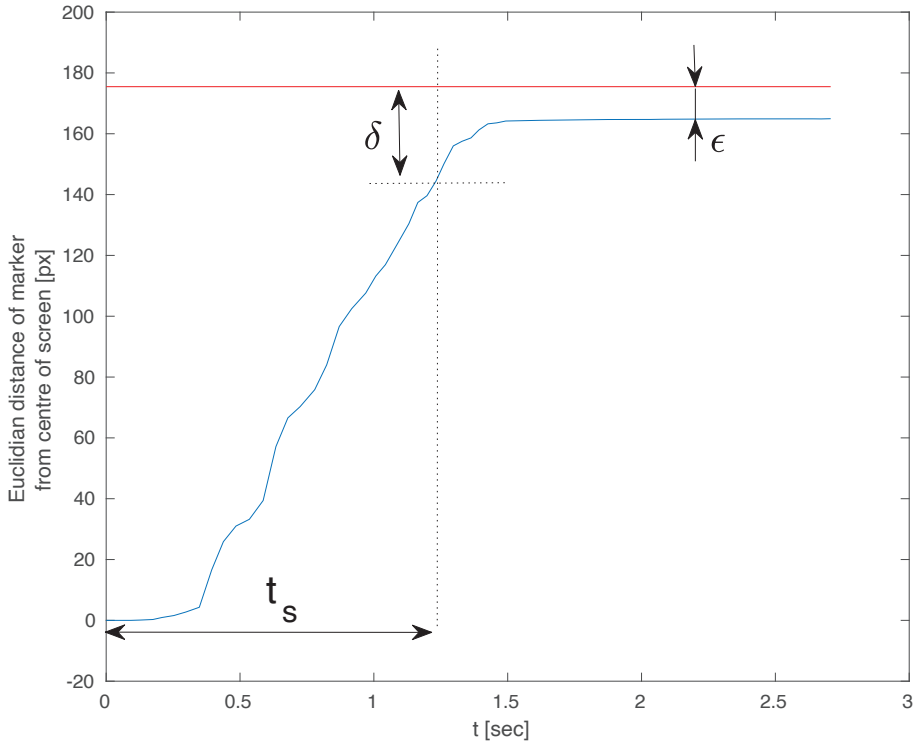


Figure 5.12: The step response of the system at point (2,0) (as defined in Fig. 5.7). The average of 20 samples is shown here.

Table 5.6: Polynomial fitting parameters

Parameter	**Value**	**95% confidence interval**
a_0	6.534	(5.953, 7.115)
a_1	0.01069	(0.0002934, 0.02109)
a_2	0.1042	(0.09567, 0.1128)
a_3	0.004279	(0.003863, 0.004695)
a_4	-9.261e-05	(-0.000359, 0.0001737)
a_5	0.004988	(0.00471, 0.005266)

For each marker position 20 repetitions were performed. Fig. 5.13 shows the settling time t_s, and the steady-state errors on the x and y position of the visual servoing of each marker (e_x and e_y, respectively). For all measurements the error on the steady-state was taken 5 seconds after the initial step input was given.

Performance

Performance of both groups (endoscopists, novices) with both modalities (GC, HC) in both tasks (SPHt, UGIt) in terms of task completion time is depicted in Fig. 5.14 and analytically in Table 5.8. Comparison showed significant difference in favour of hand control for endoscopists both during the SPHt (1:24±0:39 vs 3:18±1:14 minutes, p=0.002) and the UGIt (1:27±0:20 vs 2:10±0:35 minutes, p=0.006). Novices were significantly faster using the control both during the SPHt (3:54±1:17 vs 9:05±5:40 minutes, p=0.012) and the UGIt (1:59±0:24 vs 3:45±0:53 minutes, p<.001). While endoscopists performed significantly faster than the novices using hand control (p=0.006 in SPHt and p<.001 UGIt), gaze control was equally efficient for both groups (p=0.161 in SPHt and p=0.458 in UGIt).

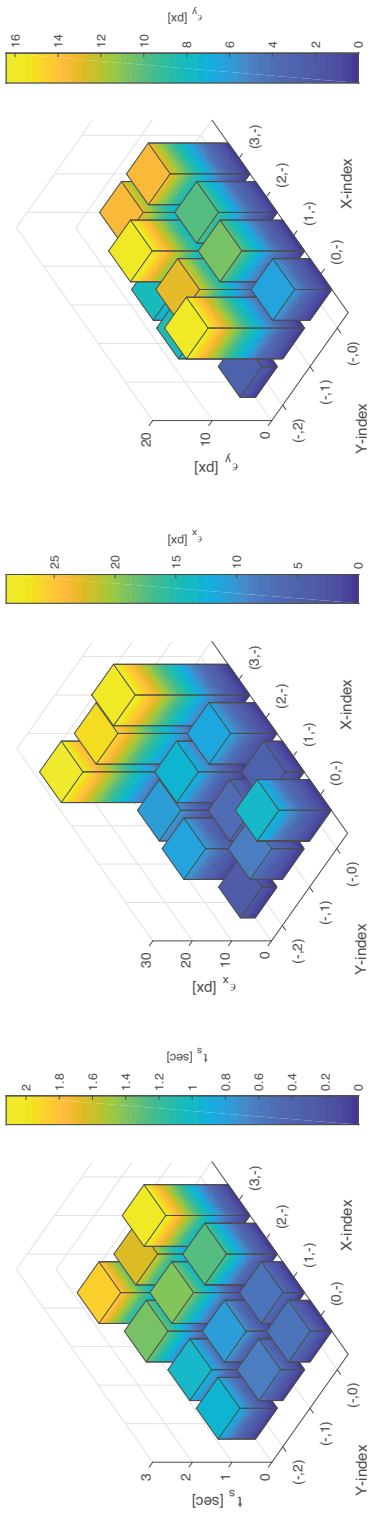


Figure 5.13: The settling time t_s, and steady-state error on the X and Y response (ϵ_x and ϵ_y, respectively). For each point 20 repetitions are recorded, resulting in the average displayed (n = 20).

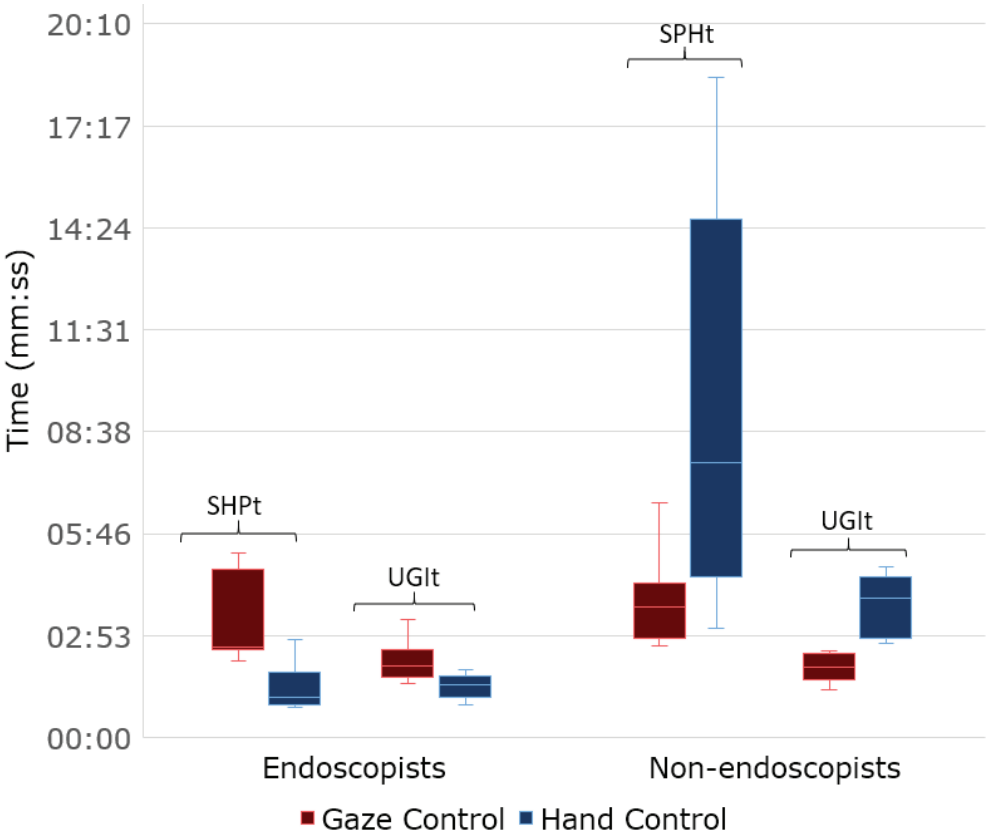


Figure 5.14: Performance comparison of the two modalities (gaze – hand control) for endoscopists and non-endoscopists on both setups (Spherical cavity task (SPHt), Upper Gastrointestinal tract task (UGIt)) in terms of overall task completion time.

Table 5.7: Root Mean Square Error on the data fitting.

Dataset	**Condition**	**RMSE [deg]**
Entire Dataset	$\forall\theta_x, \theta_y$	3.1818
Q1: top-left image quadrant	$\theta_x > 0$,$\theta_y > 0$	2.5191
Q2: bottom-left image quadrant	$\theta_x > 0$,$\theta_y < 0$	3.7296
Q3: top-right image quadrant	$\theta_x < 0$,$\theta_y > 0$	3.4971
Q4: bottom-right image quadrant	$\theta_x < 0$,$\theta_y < 0$	2.7181

5.10.4 Subjective Feedback

Overall feedback was positive with all participants expressing that the robotic platform had potential. Novices found the conventional control very challenging to learn, while they felt quite comfortable with the gaze control after a few minutes of training. Endoscopists were challenged by gaze control, mainly because of the need to adapt to a completely new technology and methodology for a task which is very familiar to them, as they perform it on a daily basis for years with the conventional way. Their feedback was focused on the translation of all information available to them during a conventional endoscopic procedure to the robotised system, such as haptic feedback.

5.11 Discussion and Conclusions

The benchmarking illustrates the relationship between the input and output and the response of the system to a simulated fixed gaze point using visual servoing. The quadratic surface fitting shows a relatively large offset parameter a_0. The large offset is most likely attributed to imperfections in the setting of the homing position. Also, the RMSE of the fitting is high. Typically, an endoscope's tip is redundantly actuated

Table 5.8: Mean, standard deviation (SD) and comparison of task completion time ($m : ss$) for the Spherical cavity task (SPHt) and Upper Gastrointestinal tract task (UGIt), performed by endoscopists and novices.

	Endoscopists		Novices		Comparisons			
	GC (1)	HC (2)	GC (3)	HC (4)		(1) vs (2)	(3) vs (4)	(1) vs (3)	(2) vs (4)
SPHt	3:18 [1:14]	1:24 [0:39]	3:54 [1:17]	9:05 [5:40]	p	**0.002**	**0.012**	0.161	**0.006**
UGIt	2:10 [0:35]	1:27 [0:20]	1:59 [0:24]	3:45 [0:53]	p	**0.006**	**<.001**	0.458	**<.001**

as it consists of multiple links and therefore degrees of freedom, whereas it is only actuated in two DOFs. As a result, the final orientation β is not fully determined by the geometry of the system: if some links are constrained in their movement, other unconstrained links will still be able to move. In a clinical setting this is important as any anatomy constraining the movement of one section will not result in the full movement to be constrained. In the experiments this translates to the large RMSE found. With no external constraints, only the internal friction will play a role in the final orientation of the system and therefore resulting in a variation of angle β. In case of constrained situations, less links actively participate in the tip position and therefore larger angle β is expected for the same motor inputs.

The visual servoing experiments show the stability of the control system in different positions. The system is optimised for the unconstrained situation, in which no overshoot is presented. In case of constraints, the transmission from motor inputs to tip output is expected to increase, and therefore likely to add an overshoot before settling to the step response of the system.

It is important to note that the benchmarking has been done for one specific endoscope. As endoscopes have a similar mechanical design, a similar motor input to tip output mapping is expected, albeit with different fitting parameters. However, for sake of usability of different endoscopes by the endoscopists, these differences are not expected to be radically different. This should be evaluated in further development of the system.

The user studies included a more realistic scenario, in which the endoscope will inadvertently be constrained by the upper gastrointestinal (GI) tract phantom and which the saccadic eye movement are included. Despite this more stochastic environment, the results show that gaze control was a feasible concept and all participants were able to navigate though the anatomy of the upper GI tract accurately and with 100% success rate. Also that gaze steering provides enhanced dexterity for

navigation and accurate target location.

Upon statistical analysis on subjective feedback, gaze control appeared more intuitive, ergonomic and implied a lower task load when compared to manual control of the endoscope among the novice users. Expert endoscopists reported higher task load with gaze control and superiority of hand control in terms of ergonomics. Nevertheless, both groups showed positive trend on the usefulness and satisfaction scales over the robotic platform.

Objective assessment of users' performance showed increased efficiency of novices with gaze control and endoscopists with hand control. However, gaze manipulation showed similar task completion time for both groups, whereas gaze outperformed conventional control for novices.

These statements support the applicability of the gaze control approach for robotic endoscopy, validating the feasibility, intuitiveness and effectiveness of the system. Diverse performance and perspectives among novice and expert users imply the significance of the effect of training in the proposed system and further studies need to be performed to assess the learning curve.

For full implementation in clinical practice, the system needs further development to fit in the clinical workflow, requiring an improved design of the hardware that has no exposed active mechanical elements. Substituting the robotic arm with a less expensive motorised unit for insertion/retraction/rotation of the endoscope would reduce the system's cost and the footprint significantly. Furthermore, incorporating haptic feedback would be a valuable feature to the endoscopist and enhance patient safety. Further evaluation of the platform is required, to compare the system to traditional manipulation also in lower gastrointestinal (GI) tract. The feasibility of tele-operation will be also investigated.

A fully robotised gaze-contingent flexible endoscope has been presented, which allows touchless control of a flexible endoscope in a free-viewing fashion. Testing with novice and endoscopist subjects in a simulated diagnostic gastroscopy (UGIt) showed that gaze controlled endoscopy is a

feasible concept. It allows ergonomic, user-friendly and intuitive control whilst maintaining the benefits of a flexible endoscope.

This chapter presents a new, more intuitive and ergonomic framework that allows easier navigation and opens the door to wider adoption of complex robotic systems with added capabilities for diagnostic endoscopy and surgery.

Chapter 6

Gaze-Guided, Assistive Robotic System for Activities of Daily Living [1] [2]

6.1 Introduction

Quadriplegia is the partial or total paralysis of all four limbs. Various illness or injury can result in this condition such as cerebral palsy, amyotrophic lateral sclerosis (ALS), muscular dystrophy, traumatic brain or spinal injury and stroke. Being unable to move around or handle objects present difficult challenges to one's daily life. For many patients, the desire to regain mobility or at least dexterity so they do not feel completely helpless, is a longing wish.

"It would almost be easier if the arms came back. You could sit in a

[1] Content from this chapter was published and is reproduced with permission from: **Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living.** Wang M-Y.*, Kogkas A.*, Darzi A., Mylonas G. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018. ©2018 IEEE

[2] This work has been conducted in collaboration with Ming-Yao Wang. The author of this thesis developed the gaze estimation module and participated to all other components, experiments, data processing and validation. Ming-Yao Wang integrated all components, performed the experiments, data processing and validation.

wheelchair, at least you could do something. When the leg comes back the only thing you learn to do is walk. But the number of things you can do with an arm..." [257].

Nowadays, there are wheelchair-mounted robotic manipulators (WMRM) available such as the JACO® [258] or iARM [259] to allow these patients to gain dexterity. The arm can be manually controlled using a joystick and pushbuttons. However, this may not be possible for patients who suffer from severe motion disabilities.

Electroencephalography (EEG) is a popular Brain Computer Interface (BCI) method that offers hands-free control. Several applications were developed, including communication [260], driving a wheelchair [261] and robotic arm control [262]. However, there are multiple challenges when using a BCI interface. The technology has long task completion time and high error rates [263]. BCI applications require high-level concentration and cognitive load which can lead to mental fatigue. A specific cognitive state may be achieved in a quiet laboratory environment but is unlikely to be produced in the real world [264]. Overall, there is no consensus on what kind of skills are required to successfully drive a BCI controlled system [265].

Eye-tracking provides a powerful alternative means of control for the disabled. Individuals with ALS or muscular dystrophy lose their muscle strength over time, eventually being unable to reach out and grasp. They also lose their ability to speak. However, they still have good control over their eyes [266]. The gaze of a person can be interpreted as the direct output from the brain. Compared to detecting brain patterns using EEG, detection of eye movement is easier, faster and has higher accuracy [264]. The current state-of-the-art gaze-based assistive devices that are commercially available are mainly screen-based systems. Screen-based systems are useful for computer related tasks such as typing, sending email, browsing the web, as the user's gaze becomes the mouse pointer. By creating specific graphical user interfaces (GUIs), control of a system can be provided to the user. *Eyedrivomatic* [267] uses arrows for users

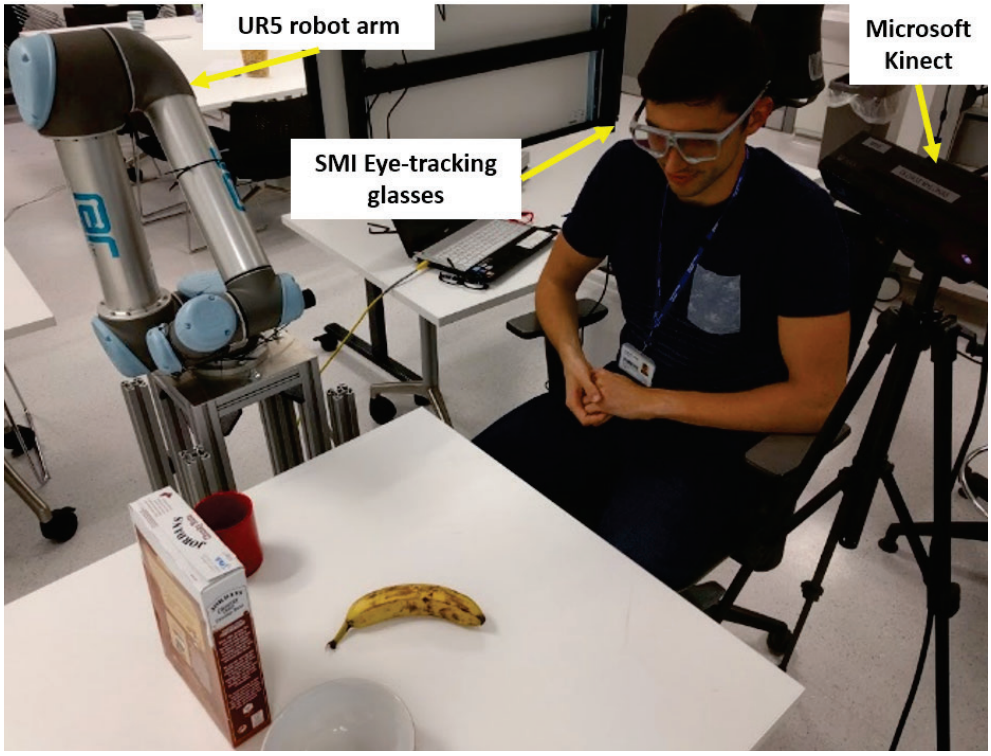


Figure 6.1: Setup of the proposed system.

to fixate and move an electrical wheelchair. A drawback of screen-based systems is that they divert the user's attention from the outside world, essentially narrowing their vision. The ideal system should grant the user control by simply looking in the real world, in other words, the ability of free-viewing gaze control. Wearable eye

Shafti et al. [158] employ wearable eye-tracking to guide assistance on reaching and grasping objects in space, by supporting the user's arm with an articulated robot. In [20] the 3D point of regard is determined using ocular vergence, followed by neural networks to improve accuracy. Using 3D gaze the user can define the contour of a target object to be grasped by the robot. However, the lack of a world frame of reference restricts the capabilities of the system to predefined and calibrated spaces. Specifically, a long calibration procedure involving 64 calibration points is required, and a head stand to prohibit head movement.

The objective of this project is to develop a system that enables patients who suffer from motor impairment to gain independence in a free-view fashion (Fig. 6.1). We achieve this by integrating free-viewing 3D fixation localisation, automatic object recognition and trajectory planning into an assistive robotic system that performs activities of daily-living (ADL). This is done with the sole use of wireless eye-tracking glasses and one RGB-D camera. The user is offered two modes of interaction with objects in space using just eye-gaze as control input and a robotic arm for manipulation. In manual mode users can control the position of the robotic arm on their head's frame-of-reference. In automatic mode a pre-defined task associated with a gaze-selected object is executed. To the authors knowledge, this is the first system of its kind, providing unconstrained freedom and flexibility in unstructured environments.

6.2 System Overview

The system consists of the following components:

- Eye-tracking glasses (ETG) from SensoMotoric Instruments (SMI) with an integrated scene camera with 1280×960 pixels resolution.
- Microsoft Kinect v2 for RGB-D sensing, with full HD 1920×1080 pixels resolution at $30Hz$ for its RGB camera and time-of-flight infrared depth sensor with $30ms$ latency.
- A 6 degrees of freedom (DOF) UR5 arm (Universal Robots A/S), for manipulation.

The setup simulates a WMRM with a wheelchair-mounted RGB-D sensor and a user wearing the ETG.

To determine the user's visual attention, the point of regard (PoR) in 3D space must be determined first. A high-level description of this task involves the following steps: (1) The RGB image information from the ETG's scene camera and the Kinect colour and depth camera images are

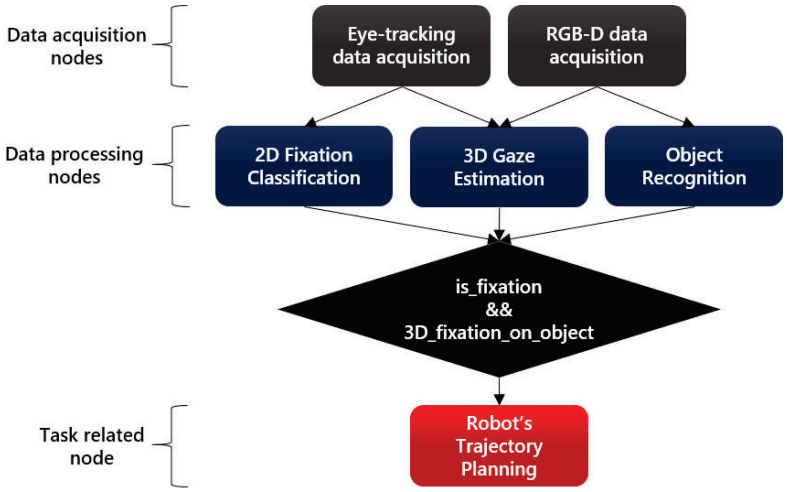


Figure 6.2: System Overview.

used to estimate the ETG pose. (2) Once this pose is retrieved, the 3D PoR can be computed as the intersection between the gaze vector and the 3D reconstructed scene. (3) Objects that are in front of the user are identified and their pose estimated. (4) Once the 3D fixation point lies on the object, the UR5 arm executes a task associated with the chosen object, depending on the mode selected. Fig. 6.2 shows the structure of the system. Object grasping is not dealt with for this project. Instead, an end-effector with a magnet attachment is used to "grip" objects.

The system is developed in *Robot Operating System (ROS)* with C++. *ROS*, being the middleware, allows effective communication to be set up between all the elements of the system. For the current implementation, a Windows 7 computer is used for acquiring and streaming the ETG data and a Linux PC with Ubuntu 14.04 is used for all other modules. The Linux PC runs on Intel Xeon Processor, NVIDIA GTX 1050 2GB, 16 GB RAM. This section discusses the methodology behind the core modules of the system, namely the coordinate frames registration, 2D fixation classification, head pose estimation, 3D gaze estimation, object recognition, trajectory planning and operation modes.

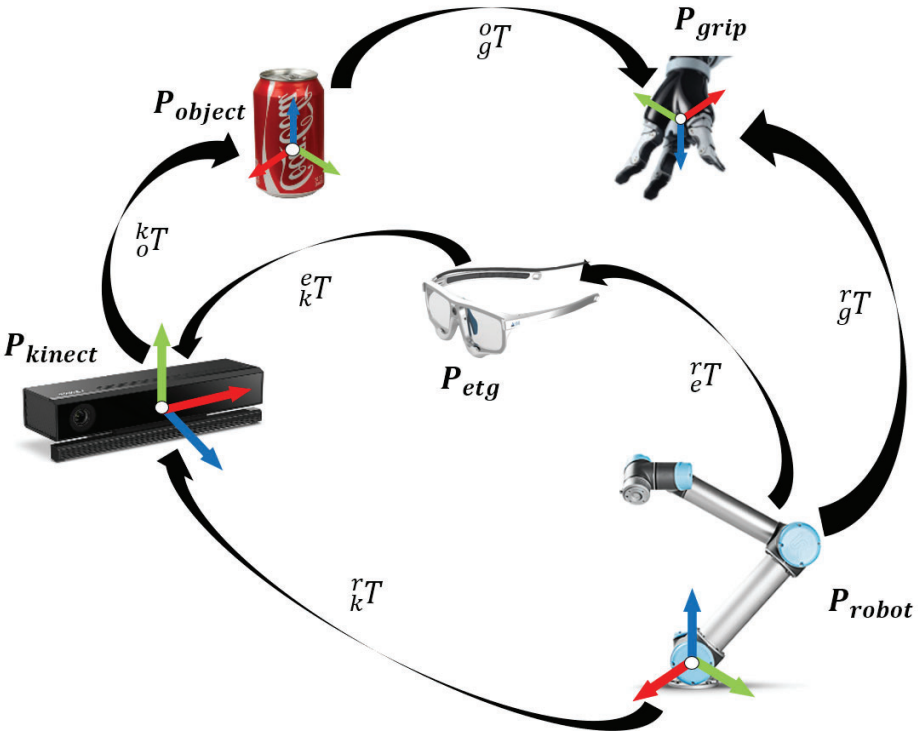


Figure 6.3: The transformations among the coordinate systems.

6.3 Coordinate Frames Registration

In the proposed system, we use the robot's coordinate system as the world frame of reference. To align multiple local frames to the global one, calibration between the robot and the RGB-D camera is necessary. The method we chose involves manually positioning the robot's end effector on the corners of a printed checkerboard, which is visible by the RGB-D camera at the same time. By performing this we estimate the rigid transformation between the robot and the RGB-D camera, as both are assumed rigidly mounted on the frame of a wheelchair. The transformations shown in Fig. 6.3 are described by the following equations:

$$ {}^{r}_{g}T = {}^{r}_{k}T * {}^{k}_{o}T * {}^{o}_{g}T \tag{6.1} $$

$$ {}^{r}_{k}T = {}^{r}_{e}T * {}^{e}_{k}T \tag{6.2} $$

6.4 2D Fixation Classification

The ETG provide the 2D PoR on the user's head frame-of-reference. As a safety precaution, the activation routine of the robotic manipulation task is based on the 2D fixation dwell time. First, the velocity of eye movement is estimated [226] and a threshold of $36deg/s$ is set to filter out fast saccadic movement. Moreover, we only consider fixations over a dwell time threshold of $2s$.

6.5 Head Pose and 3D Gaze Estimation

The 3D gaze estimation component is based on the novel framework proposed in [62] and relies on the combination of advanced computer vision techniques, RGB-D cameras and ETG. With reference to Fig. 6.4 the process consists of two tasks: user's head pose estimation and 2D to 3D gaze mapping. The user's head pose is equivalent to the ETG's RGB/scene camera pose in space. For the camera pose estimation, BRISK features [36] are detected and matched in both the ETG's frame and the RGB camera frame of the RGB-D sensor. The RGB-D extrinsic camera calibration [210] provides the depth values of the matched RGB features and consequently the 2D-3D correspondences for the ETG's features (2D points on ETG's RGB/scene camera – respective 3D coordinates in the Kinect's coordinate system). Then, EPnP with RANSAC and Gauss-Newton Optimisation [37] provide the ETG's scene camera pose in space. For the last step, we use ray casting to backproject the gaze ray from the compressed model of the 3D reconstructed environment (to improve performance) on the estimated camera pose origin, allowing real-time and free-viewing 3D fixation localisation. Fig. 6.4 outlines the 3D gaze estimation framework.

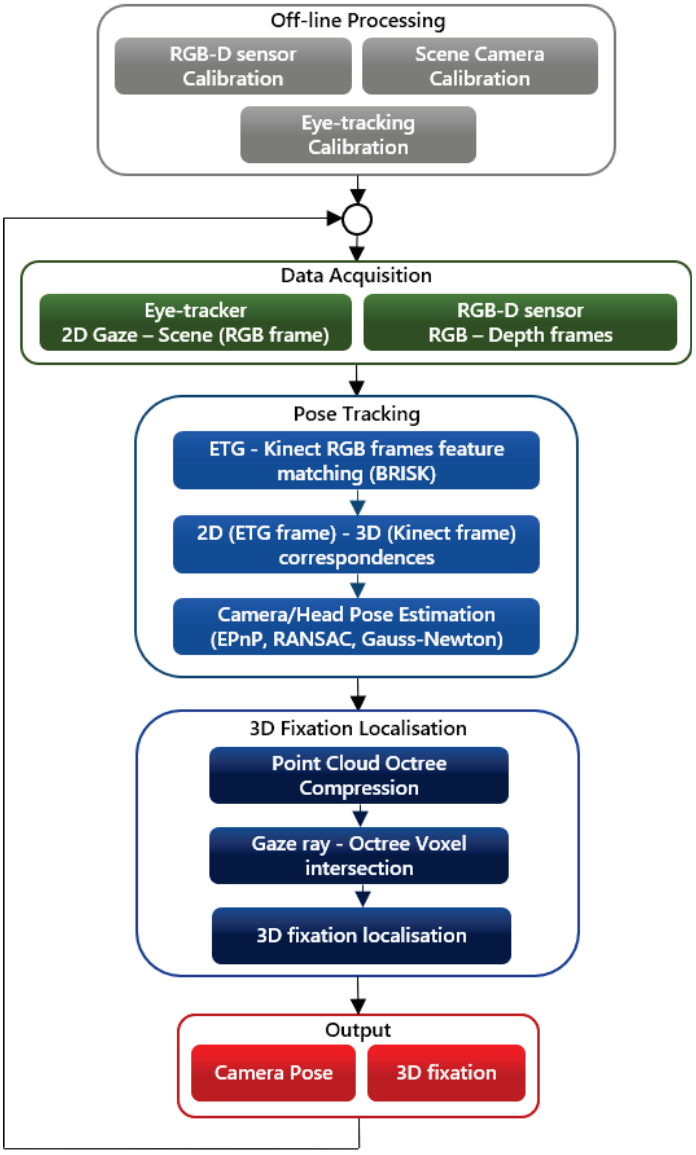


Figure 6.4: 3D gaze estimation module.

6.6 Object Recognition and Selection

For object detection and pose estimation, *LINEMOD* [268] was used with *Object Recognition Kitchen (ORK)* [269] as a backend. *LINEMOD* is a real-time template matching method and *ORK* is a framework which offers various techniques for object recognition. This includes setting up a local database to store a 3D mesh file of each object and generating the templates of the stored objects.

The 3D fixation corresponds to a point from the point cloud of the Kinect scene. To identify whether this point is on any of the recognised objects, a set of neighbouring points around the fixations is compared with *ORK*'s point cloud. A *k-dimensional (k-d)* tree algorithm was deployed to search for nearest neighbours with a radius of $1cm$. In case the 3D fixation is detected within the point cloud, the next step is to identify the specific object being fixated. As *ORK* provides the centroid for each object, the Euclidean distances between the fixation point and the centroids for all objects were calculated. The object with the shortest distance would be the fixated object and its pose then becomes the input for the trajectory planning module. Obstacle detection has yet to be implemented at this stage.

6.7 Trajectory Planning

To control the UR5 arm, the *Moveit!* framework [270] was selected. *Moveit!* is an open-source software for robotic manipulation, motion planning and control and is fully integrated with *ROS*. Therefore, it allows easy communication with our Kinect perception and gaze-control module.

From the object recognition module, the pose of the selected object's centroid is received. From the object's centroid, the contact point for the magnetic gripper is calculated on the object's surface, based on its known dimensions. Depending on which objects are selected, different

manipulation poses are established for the task intended (pre-grip poses), followed by object pick up. All movements in the manipulation module can be divided into two types: *motion planning* and *cartesian path planning*. Motion planning is based on planning a collision-free path from the current state to a designated pose, while cartesian path planning relies on computation of waypoints. The former was used to generate a trajectory from the robot's home pose to the object's pre-grip pose, while the latter was deployed once the arm reached the pre-grip position and in the *manual mode* (6.8.2). Safe zones, such as where the user is and the table, have been set up to prevent path planning from taking place within this space.

6.8 Operation Modes

The system offers the user two modes of interaction with the objects.

6.8.1 Automatic Mode

The automatic mode executes a pre-defined task associated with a selected object. The user triggers the task by fixating on a recognised object. According to [271], meal preparation and drink retrieval were considered top desired tasks for disabled patients. It was decided that the automatic mode should incorporate these functions.

6.8.2 Manual mode

The manual mode provides the user with positional control of the end-effector in the X, Y, Z axes with respect to the ETG frame. The transformation between the ETG and the world frame, which is aligned to the robot frame, is initially calculated (6.2). This allows the end-effector's position to be determined in the ETG frame. The 2D gaze coordinates from the ETG are translated to a movement in one of the three directional axes. A dead zone of 300×300 pixels was created in the centre of the

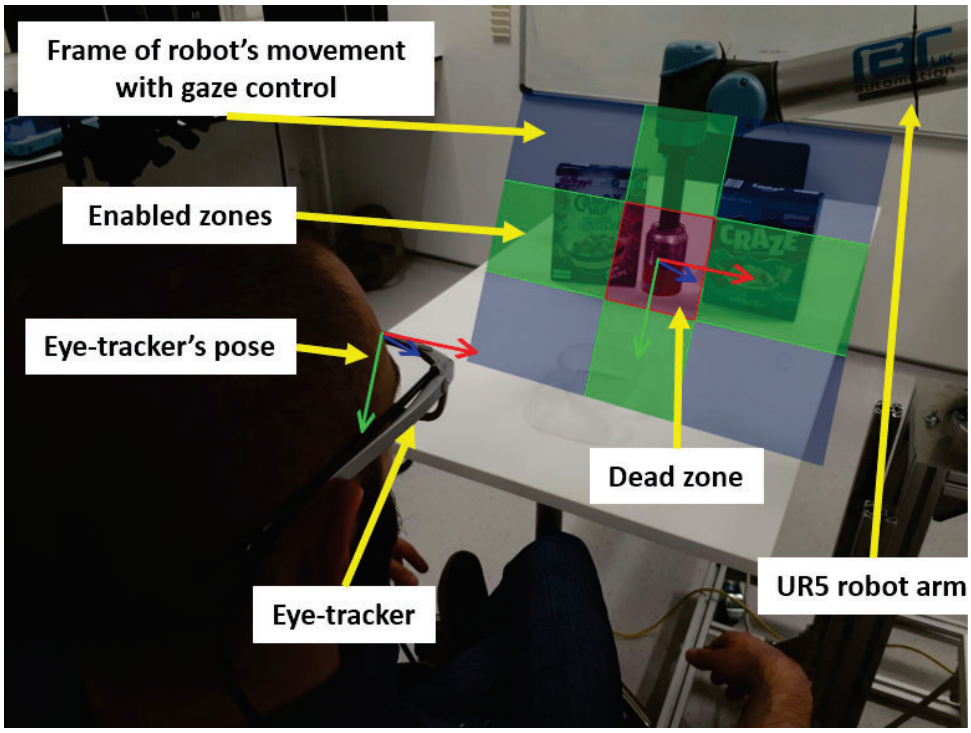


Figure 6.5: Control plane corresponding to the user's view in manual mode.

ETG RGB image. The robot will not move if the 2D PoR is within this zone. If the user's PoR is to the left of this zone, the robot moves to the left by a small pre-defined offset of $2cm$; this also applies to right, up and down. In and out depth movement is performed by closing one or the other eye. This discrete motion of the manipulator was chosen over continuous action, as it was found that the user can perform the task safer and more intuitively. Orientation control is not included at this stage as this might increase complexity for the user. Once the new pose has been determined in the ETG frame, this gets transformed into a coordinate in the robot frame and the robot moves in a step manner. Fig. 6.5 shows a visualisation of the control plane projected in front of the user, in the same orientation as the ETG pose (scene camera). Synthesised voice feedback acknowledging the directional commands is provided for assistance, as the user's centre of gaze may not always be on

the end-effector and also it was found that feedback helps with the overal confidence of the user during task execution. The small steps allow the user to perform fine positioning of the end-effector, ideal for situations where the pose of an object is inaccurately determined due to point cloud distortions or other artifacts.

6.9 Application Workflow

The workflow of the system starts with an off-line pipeline required by the object recognition module and the Kinect-to-Robot registration. First, the 3D mesh models of the objects are loaded to *ORK*. Then, the RGB-D camera is registered to the UR5 robot (world coordinate system). Finally, the user wears the ETG and performs a standard eye-tracking calibration procedure to align the ETG's scene camera frame with captured gaze vectors while fixating on three different and spread out in space points. Finally, the user is ready to fixate on the trained objects to trigger the automatic or the manual mode.

6.10 Validation Method

6.10.1 3D Gaze Estimation Evaluation

The accuracy and computation time of the 3D fixation localisation were examined. A subject was recruited and asked to fixate on 10 predefined targets from 6 different positions.

The accuracy of the 3D fixation localisation is computed by measuring the Euclidean distance of the estimated 3D fixations (output of the 3D gaze framework) and the predefined targets as observed by the RGB-D camera (on the Kinect coordinate system).

Moreover, the computation time of the 3D fixation localisation was computed as the interval between the moments the subject's PoR was classified as a fixation and the 3D fixation was estimated by the framework.

6.10.2 Trajectory Planning Performance

The success rate of the trajectory planning was examined. On the Kinect cloud, 3D points which belong to the objects of the experimental setup were manually selected and the rate of successful trajectory planning was estimated. The time between the moment a point was selected and the moment the robot started the object-specific task was also measured. Two objects were used for this experiment, a mug and a cereal box, which require different griping orientation by the robot. Each object was placed in 3 different positions on a table, within the robot's maximum reach. The process was repeated 10 times for each object.

6.10.3 Overall Evaluation of the System

An experimental study was performed to assess the usability of the overall system. Two experiments were carried out to validate each operation mode. The study measured the system's performance objectively as well as the users' subjective experience. The experiments were carried out in a well-lit room and objects were placed on a nonreflective table. Five healthy subjects, aged between 21–26 years participated in the study. Two subjects had normal vision while the rest had corrected vision. Prior to the experiment, each subject was briefed on the purpose of the study, the technology involved and the expected tasks outlined below. A three-point calibration was performed at the beginning of each experimental session to ensure that the ETG were correctly tracking the subject's pupils and subsequently providing the accurate gaze direction.

Automatic Mode

The experimental setup involved placing a coffee mug, a cereal box, a bowl, a banana and a plastic container on a table. Fig. 6.6 shows the setup of the experiment. All objects were placed between 100–120cm away from the Kinect sensor but within the UR5's working space (85cm reach). Three tasks were implemented for the study:

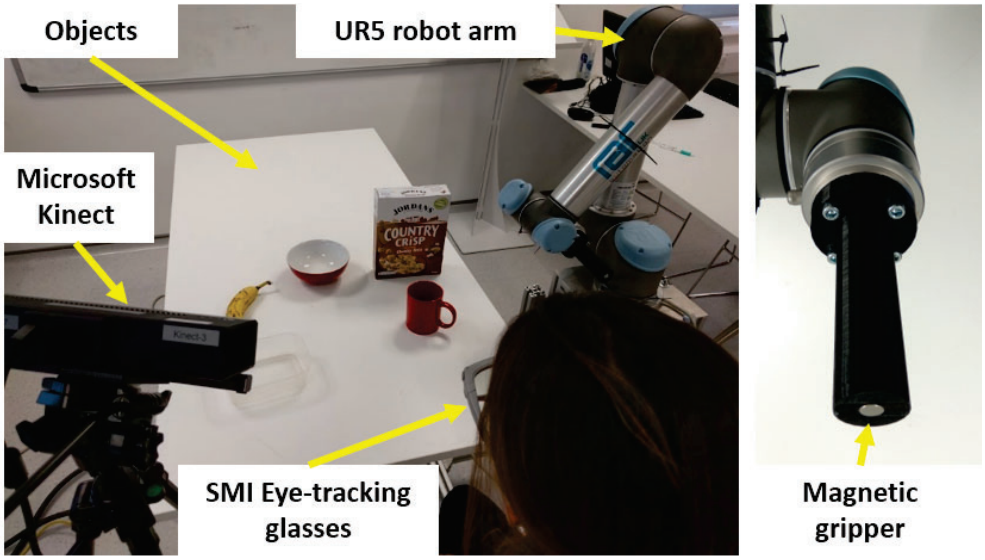


Figure 6.6: Experimental setup simulating a WMRM, assuming an external mount on the left side of the wheelchair for the Kinect sensor.

- By fixating on the mug, the robot would reach inside the mug and bring it towards the user.
- By fixating on the cereal box, the robot would pick it up, locate the bowl and pour cereals into it. The robot then places the box beside the bowl.
- By fixating on the bowl, the banana and the plastic container should not prompt any robotic action (the latter two are not loaded to *ORK* and are considered distractors).

An instructor then requests the subject to fixate on one of the objects on the table to prompt the above tasks. The order of fixation was given randomly by the instructor. Once the set of fixations on five different objects has been completed, the positions of the objects were randomised for the next set. Each subject was asked to perform three sets of trials.

Manual Mode

The object of choice for this experiment is an aluminium soft drink can. The reason being, reflective objects do not get accurately detected by the RGB-D sensor due to multipath interference, therefore the estimated pose is incorrect. We made use of this occurrence and requested the subjects to fixate on the can. The system would output the incorrect pose of the object and the robot would move towards the pre-grip pose, somewhere close to the can. The subjects were then instructed to steer the robot with their gaze to pick up the can and place it in a plastic container with dimensions $12{\times}15{\times}5cm$ positioned $30cm$ away from the can. The subjects were instructed to activate each direction once with the instructed eyes gestures prior to the experiment, but no training runs were provided. This experiment was performed twice for every subject.

Control Modalities Evaluation

Individual elements were evaluated simultaneously during the study along with the overall success rate of the system. Measurements for automatic and manual mode are as follows:

- *Automatic Mode*

  Successful selection of the object – Five different objects were used in the experiment to assess the performance of the object recognition and 3D gaze estimation elements of the system. It was considered a success when the system planned a path to the predefined pose of the selected object.

  Activation time – The elapsed time from when the user begins fixating on the object to when the robot starts moving. This outcome signifies real-time usability.

  Task completion success rate – When the robot successfully performs the intended task that corresponds to the object selected, without colliding with other objects or faulting out.

- *Manual Mode*

  Task completion time – The time elapsed from the user gaining control of the robot to when the can touched the bottom of the container.

  Task completion success rate – Successful or not successful.

Selection of object and activation time were not measured in manual mode as this was validated during automatic mode. After the experiment, the subjects were asked to fill out a questionnaire regarding their experience using the assistive system. A 5-point *Likert scale* ranging from 1 – strongly disagree to 5 – strongly agree, was provided to rate their opinion.

6.11 Results

6.11.1 3D Gaze Estimation Evaluation

The 3D gaze estimation is evaluated in terms of accuracy and computational time. For this, 10 markers were placed on objects positioned at different depths. The distance between the RGB-D camera and the objects is 100–130cm, which forms a realistic workspace for the specific application, considering the UR5's maximum reach of 85cm. The average error is 2.31±1.03cm and the computation time was measured at 0.69±0.09s (Fig. 6.7(a-b)). The computation time comprises of the camera pose estimation and the 3D fixation localisation parts.

6.11.2 Trajectory Planning Performance

The activation time of the robot's path planning was measured. As shown in Fig. 6.7(c), the interval is 2.3±2.26s for the cereal and 1.23±1.81s for the mug. Moreover, 100% success rate was achieved by the trajectory planning modules, while 91.67% was the rate for the successful grasping of the targeted objects.

6.11.3 Overall Evaluation of the System

Automatic Mode

Table 6.1 shows the success rate of the system modules along with the overall success rate for the automatic mode. The high success rate of the gaze-guided object recognition demonstrates that the system is capable of recognising the objects on the table and the 3D gaze estimation is accurate enough to trigger the intended robotic task. The path planning can also be considered reliable, failing only one time out of the 30 attempted plans. The overall system success rate dropped below 90%, despite the previous modules having over 96% success rate. This is due to the non-deterministic nature of the sampling-based motion planner. Although

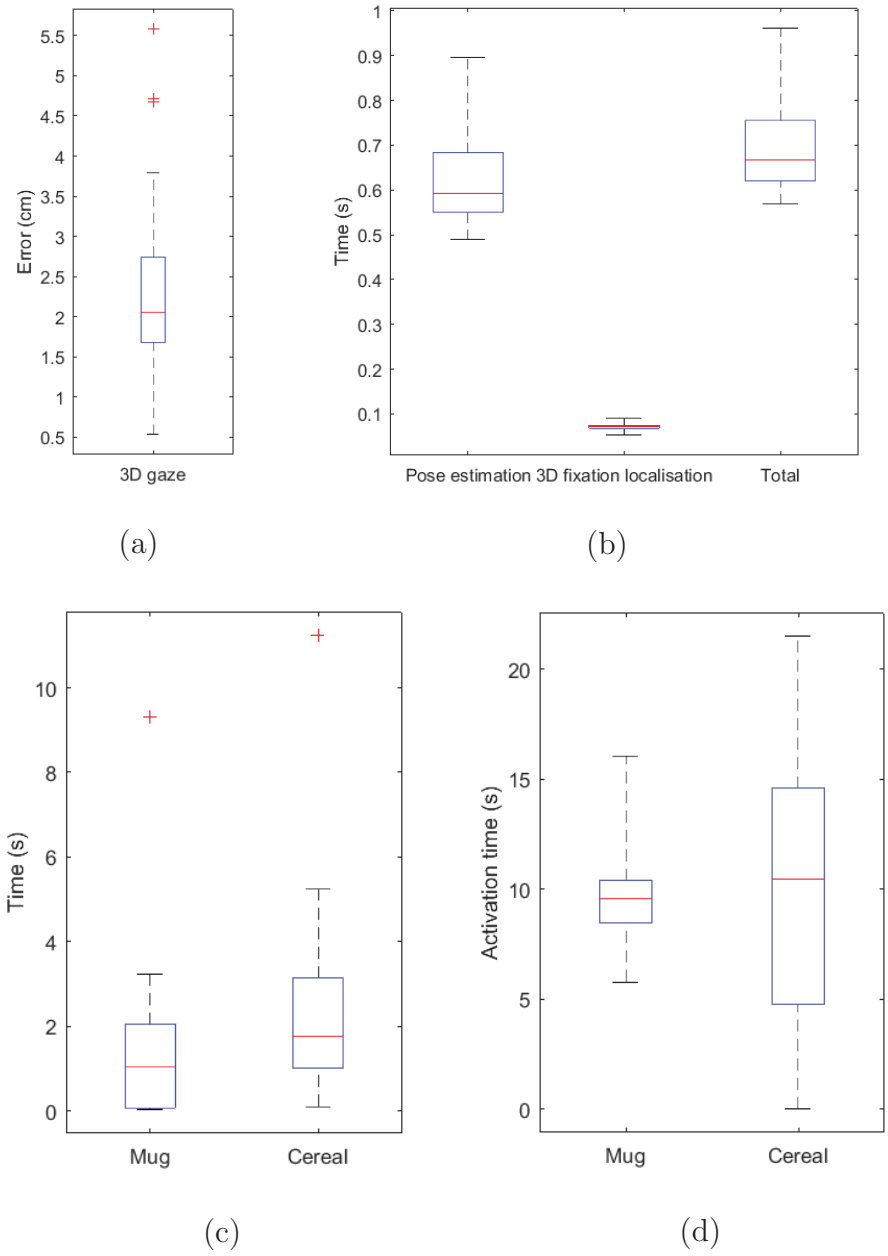


Figure 6.7: (a) 3D gaze error and (b) time of pose estimation and 3D fixation localisation. (c) Path planning time of the robot, targeting the cereal and the mug. (d) Activation times for mug and cereal, from user beginning fixating to robot moving. This timing includes the 2s dwell time threshold and the 1s of *ROS* sleep.

the generated path was valid, without obstacle detection implemented it was possible that it collided with an object as it travelled through its trajectory. This, however, was considered a fail during the experiment.

Fig. 6.7(d) shows the activation times for each object. The resulting average activation time was $9.92 \pm 4.78s$. Removing the fixation requirement of $2s$ and the *ROS* node sleep rate of $1s$, the average time to determine the user's 3D fixation point and to plan a valid path is $6.92s$. Although activation time is an important aspect for a Human-Robot Interaction system, studies showed that patients did not feel the time to complete the task was significant, but rather they are content with being able to perform the task independently [271].

Manual Mode

All subjects were able to complete the task of picking up the can and placing it in the plastic container, demonstrating a success rate of 100%. Each subject showed the ability to grasp the control within the first run and improved the execution speed on the second run, as seen in Fig. 6.8(a). This study showed that the system was intuitive enough as no training was provided beforehand.

User Experience

All subjects' feedback is shown in Fig. 6.8(b). Questions regarding the negative aspects of the system generally received a low score, indicating the users were not frustrated or fatigued while operating the system. The

Table 6.1: Automatic Mode Success Rates

Gaze Guided Object Recognition	98.67%
Path Planning	96.67%
Overall System	86.67%

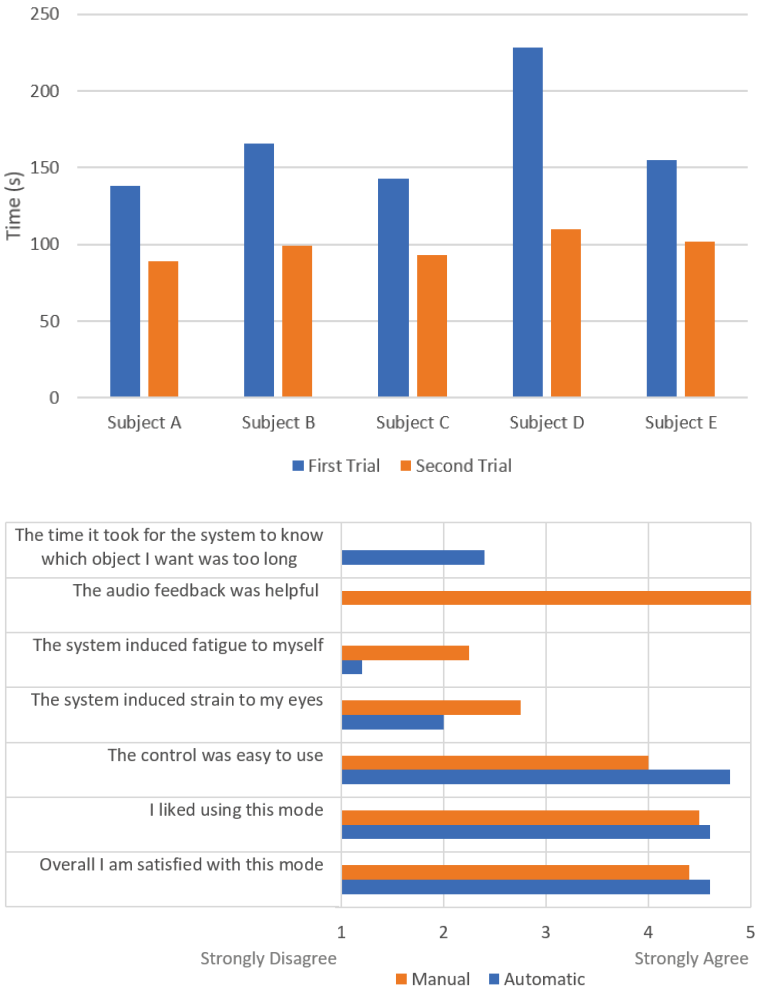


Figure 6.8: Top: Completion time for each subject for pick and place task. Bottom: Subjects' feedback for both manual and automatic modes.

time for the system to know which object was targeted trended towards a neutral score. This is related to the activation time and how some users experienced a longer wait in some occasions. The cause could arise from the inability to detect their eyes, the inability to compute the ETG pose or the random nature of motion planning. The question related to the system inducing strain to the user's eyes for the manual mode had a neutral score of 2.75. This was expected as the person is fully controlling the robot compared to the other mode, which is relying on activation just by the fixation. The positive aspects of the system received high scores, with the overall satisfaction score being 4.6 / 5.

6.12 Discussion and Conclusions

6.12.1 System Limitations

Being at an early stage of development, the system has some limitations, which can affect its success rate and practical usability. As mentioned previously, the current implementation does not yet include obstacle detection, therefore the valid paths that trajectory planning produces have the possibility of objects collisions. Overcoming this limitation is feasible by using Octomap [221] to convert the RGB-D data into occupied space. *Moveit!* will then be able to plan around the occupied region and generate collision-free trajectories.

In order for the system to be usable in everyday life, there is the evident need of a grasper. Integration to a commercial WMRM solves this issue and the product also contains pre-defined ADL tasks. However, prior to integration, the system needs to be able to switch between the different modes during runtime. This allows the patient to correct for any errors the system makes in pose estimation while granting them total control of the manipulation. Potential methods for switching between modes can range from closing one's eyes for a certain duration, draw a pattern with gaze gesture or even using additional hardware, such as

Augmented Reality (AR) glasses, just to name a few possibilities. Finally, the last mile, i.e. allowing the robotic manipulator to approach the user's lips and complete the task, is not handled with the current version of the system, but this can be solved with an additional camera for face tracking.

6.12.2 Conclusion

In this chapter, we presented a proof-of-concept for a gaze guided assistive robotic system used in a real environment. The system relies on wireless eye-tracking glasses and an RGB-D camera to achieve free viewing 3D gaze estimation in real-time, object recognition and trajectory planning. A robotic arm is used to execute activities of daily living, such as meal preparation and drink retrieval. Automatic and manual operation modes were implemented to provide useful interaction between the user and desired objects. The results show that the system is accurate, intuitive and easy to use even without training. For its practical deployment and extensive evaluation with actual patients, collision avoidance will have to be implemented and the RGB-D camera and a lightweight robotic arm have to be integrated with a wheelchair.

As the system is designed for home use, 3D models of household items can be added to the object recognition database. We can utilise the RGB-D sensor to scan the object and create a 3D mesh of it. This will allow the patient to scan objects of their choice, creating a personalised database.

Additional hardware, such as AR glasses, will enhance the user experience and allow further independence to the user, bringing the system closer to its actual integration in the everyday life of patients with severe motion disabilities. Future work involves actual patients.

Chapter 7

Conclusions and Future Research Directions

Patient safety and quality of care remain the focus of the smart operating room of the future. Some of the most influential factors with a detrimental effect on these two areas are related to suboptimal communication among the staff, poor flow of information, staff workload and fatigue, ergonomics and the sterility in the operating theatre. The integration of new technologies into the surgical workflow adds significant complexity. Nevertheless, while technological developments constantly transform the operating room layout and the interaction between surgical staff and machinery, a vast array of opportunities arise for the design of systems and approaches, that can enhance patient safety and improve workflow and efficiency. In the age of information we live in, the surgical domain endeavours to follow the industrial paradigm shift towards data-driven processes for improved products and services. To this end, perceptually enabled data is the foundation stone that will lead to cognition-guided surgery.

The proposed framework allows touchless interaction with medical technologies in the OR. As such, the surgical team can alleviate communication failures, poor information flow and ergonomic design, which would lead to patient harm and staff dissatisfaction. Most importantly, the data

provided by the employment of the framework into the surgical workflow, can provide invaluable perceptually enabled information.

7.1 Achievements and Contributions of the Thesis

Key aspects of the motivation behind this work are presented in **chapter 2**. It is an attempt to highlight potential applications of the proposed gaze-contingent framework, in the context of touchless interactions in the sterile environment. Moreover, the utility of the perceptually enabled data which derive from these applications is signified, and how they can lead to Surgery 4.0 through Surgical Data Science. Therefore, an overview of predominant safety risks in the OR is provided, followed by a review of touchless interaction modalities with emphasis in gaze-contingent systems and applications. Finally, future operating theatre concepts are discussed as reported in the literature.

In **chapter 3** the free-viewing 3D gaze framework is introduced. A core aspect of the proposed framework is its capability to estimate the 3D PoR of the user. Wearable eye-tracking glasses (ETG) can be used to provide 2D gaze information and a scene video on the head frame-of-reference of a user. After a short calibration routine, gaze vectors can be mapped to unique 2D gaze points on a virtual plane attached to the scene camera of the ETG. This plane is also fixed to and rotates with the user's head. Consequently, there is no direct quantitative correlation between 2D fixations and 3D positions of objects in space. To overcome this limitation, localisation of 3D fixations is achieved through the combined use of conventional wearable eye-tracking, fixed in space RGB-D cameras for 3D reconstruction of the environment and (occasionally) a motion capture system (MCS) for the head pose estimation. The framework relies on the ability to provide an accurate estimate of one's head pose (equivalent to the ETG's scene camera pose) on a world coordinate system

(WCS) fixed with respect to the operating theatre. The pose is then used to map the 2D gaze information reported by the eye-tracker to a unique 3D fixation in the world frame-of-reference. Then, the 3D fixation can be translated into screen 2D fixation information when the user gazes on a screen in space, allowing simultaneous macro- (theatre-wise) and micro-scale (patient/screen-wise) fixation localisation.

The introduction of a robotic scrub nurse as an integral component of the Smart-OR, may address nursing shortages and empower the team by enabling the surgeon and the human scrub nurse to perform a wider variety of tasks in a more efficient and safe manner. This is achieved by offering a "third hand", but most importantly one that is under the firm control of the operating surgeon. Extending the capabilities of the framework introduced in chapter 3, a robotic scrub nurse system, visually controlled in a mobile and unrestricted fashion, was presented in **chapter 4**. This platform allows hands-free gaze-driven interactions with a screen and a robotic arm, which acts as a robotic nurse assistant by transferring surgical instruments to the surgeon. The surgeon uses natural gaze via wearable eye-tracking glasses to select surgical instruments on a screen, in turn initiating the robot to deliver the desired instrument. The platform was tested with ten different surgical teams in simulating a common operative scenario with similar theatre staff representation and operative field set up. The system workflow, usability and acceptability were evaluated by the operating team, during an ex vivo task on pig colon. Quantitative and qualitative feedback was positive. In comparison between human- and robot-assisted tasks, no significant difference was found in overall task load of the surgeon, task workflow (interruptions) or operative time.

Flexible endoscopy is a routinely performed medical technique, which has been traditionally a diagnostic tool, allowing the exploration and the acquisition of tissue biopsies in the upper and lower gastrointestinal (GI) tract. Despite its wide adoption, diagnostic endoscopy presents several challenges, such as limited dexterity, decreased spatial awareness,

loop formation and overall poor ergonomics. Therefore, building on the framework presented in chapter 3, a robotised flexible endoscope controlled by the user's natural gaze and head orientation is proposed in **chapter 5**. It is a fully motorised gaze-contingent system for non-restricting, free-view flexible endoscopy and is based on a robotised system, which allows hands-free control of the endoscopic view in an intuitive fashion, using the natural gaze of the user to steer the endoscope tip. Eight experienced endoscopists and Eight novice users were recruited to assess feasibility and feasibility and comparison against traditional hand control, in a simulated diagnostic task of the upper gastrointestinal (GI) tract. Whilst endoscopy was proven a feasible concept, expert endoscopist performed better with the traditional hand control. However, novice users showed significantly higher performance and preference on the proposed system. At the same time, both groups showed comparable efficiency with gaze control, providing initial indications about the learning curve of both modalities.

The potential applications of the multi-modal framework proposed in chapter 3, is not naturally restrained in the context of the OR. In line with the trend of surgery adopting paradigms by other industries, such as aviation or petroleum, to improve surgical outcome, technologies designed for surgical use may have applications in other domains in healthcare or industry in general. **Chapter 6** attempts to augment the capacity of applications where the work of this thesis can be valuable, by proposing an assistive system for people with motor disabilities. It is an assistive robotic system with an intuitive free-view gaze interface, which allows the user to interact with objects using fixations. The gaze-guided object recognition routine starts with the user's gaze direction, provided by wearable eye-tracking glasses. Subsequently, it is mapped in real-time in a world frame, defined by the presence of an RGB-D camera which is rigidly mounted on the patient's wheelchair. The 3D gaze information allows free head movement and is combined with object recognition and trajectory planning. Two operational modes have been implemented to cater for

different eventualities. The automatic mode performs a pre-defined task associated with a gaze-selected object, while the manual mode allows gaze control of the robot's end-effector position on the user's frame of reference. User studies reported effortless operation in automatic mode. A manual pick and place task achieved a success rate of 100% on the users' first attempt.

7.2 Future Research Directions

Software suite for visualisation, surgical workflow and operator analysis

The novel framework can provide objective insight into the theatre attendants' visual behaviour and their interactions with a fully-registered surgical environment in real-time and 3D. This is expected to reveal a vast array of perceptually-enabled information. Information from several other modalities can be acquired and recorded, including patient physiological data. Multi-modal recordings and playback of this information could provide a training tool, as well as a rich database for further clinical and behavioural investigations, surgical workflow and operator analysis. It is envisaged that a standardised open-source framework deployed in several theatres could provide a large amount of data, which will be made available in an open-access basis between interested researchers around the globe. Development of a software suite concurrently with the research framework will allow off-line visualisation and analysis of acquired and disseminated data.

Workflow segmentation and safety alarms

One attractive implementation that could deal with safety issues in the operating theatre is modelling and recognition of surgical workflow. Detecting deviation from normal workflow patterns could be an indication of physical or cognitive fatigue and error. With the proposed framework,

real-time safety alarms could reduce relevant errors. Moreover, the surgeon could be notified at specific predefined workflow stages with relevant contextual information (e.g. tips, automatic presentation of patient data, reminders on patient's specificities), which could be critical during complex and long surgical procedures.

Intelligent robotic scrub nurse

The rich perceptual data provided by the framework could enhance the modelling of the workflow. Further development of the visually aided robotic scrub nurse can be investigated, by enabling it to automatically respond to surgeon instrument selection behaviours, through task phase recognition. This way, the system could imitate the human scrub nurse's greatest advantage of instrument anticipation, as was emphasised in our subjective feedback.

Free-viewing collaborative eye-tracking

Revealing the visual attention of two or more attendants in the operating theatre is hypothesised to enhance collaboration by improving speed, accuracy, and reliability during collaborative tasks, as shown in laparoscopic [49] and robotic [21] surgical settings. Free-viewing collaborative eye-tracking involves sharing visual projections of each collaborator's 3D fixation within the operating room, and 2D fixation overlays on the laparoscope monitor (macro- and micro-scale), or other screens whenever applicable. This is expected to provide an additional interaction channel between the surgeon and the supporting personnel, enable more efficient handling of surgical instruments, tackle verbal communication issues within the surgical team, as well as facilitate training. The collaborative framework will be further enhanced to include functionalities such as the use of augmented reality glasses or gaze-contingent projection for displaying contextually relevant information and augmented reality visualisation at the fixation location.

Gaze-contingent flexible endoscopy

The evaluation of the fully robotised gaze-contingent flexible endoscope supported the applicability of the gaze control approach for robotic endoscopy. Feasibility, intuitiveness and effectiveness of the system were validated. Diverse performance and perspectives among novice and expert users imply the significance of the effect of training on the proposed system and further studies need to be performed to assess the learning curve. However, for full implementation in clinical practice, the system needs further development to fit in the clinical workflow, requiring an improved design of the hardware that has no exposed active mechanical elements. Substituting the robotic arm with a less expensive motorised unit for insertion/retraction/rotation of the endoscope would reduce the system's cost and the footprint significantly. Furthermore, incorporating haptic feedback would be a valuable feature to the endoscopist and enhance patient safety. Further evaluation of the platform is required to compare the system to traditional manipulation in the lower gastrointestinal (GI) tract. The feasibility of tele-operation will be also investigated.

Gaze-guided assistive robotic system

The gaze-guided assistive robotic system for daily-living activities is an introduction to the use of our framework for enhancing disabled people in every day life. An extension of the current system could involve additional modes of operation, such as semi-automatic, where the users can pick an object and a specific spot where the robot will place it by using only their gaze. Further features can also be integrated, such as enriching the inventory of object specific gripping routines through attachment of relevant robotic hands developed or augmented visualisation of the different modalities the system offers.

Gaze-guided surgical light

The surgical lighting system is one of the medical technologies with notable ergonomic shortcomings; every 7.5 minutes a luminaire action takes place [272]. In [62] we demonstrated the core functionalities of the proposed framework by co-registering an articulated robot to guide a laser diode, which was mounted on the robot's end effector, to highlight the user's point of gaze in space. The same principle can be followed to guide the surgical lamp towards the surgeon's fixation point. A robotic design for a surgical light can be implemented and safety considerations integrated in the surgical workflow.

Body tracking and multi-sensor fusion

The essential equipment for the framework implementation consists of wearable eye-tracking glasses, RGB-D cameras and the workstation. The data provided by the framework as processed information, can be used in conjunction with the plethora of raw data deriving by the sensors:

- 3D gaze framework
  - 3D gaze in world coordinates
  - 6 DOF Head pose
  - 3D point cloud of the environment
- Eye-tracker
  - 2D gaze (PoR)
  - Scene RGB image
  - Further eye related data (eye images, pupil diameter)
- RGB-D camera
  - RGB image
  - Depth Image

For example, the Kinect has been well-established for estimating human skeletal poses. Eye-tracking can then be used for monitoring mental workload especially through the dilation of the pupil [273]. Larger pupil size has been shown to correlate to heavier mental workload [273]. The continuous and dynamic monitoring of the system could then be used to further the understanding of the onset of fatigue. Other sensory inputs (EMG, EEG, ECG, endoscopic video, etc.) could further enrich the ensemble of perceptually enabled data, which could allow AI methods to reveal a new horizon of semantic information, such as error detection or even prediction (Fig. 7.1).

7.3 Conclusion

Overall, the work presented here draws inspiration from the increasing utilisation of data from diverse sources and is fundamentally driven by the need to keep the surgeons and their physical interactions with the environment tightly integrated into the decision-making process. The ultimate goal is to develop functionalities, methodologies, open-source software and a low cost generic hardware framework that can be adapted to any operating theatre with minor modifications and effort. Exemplar functionalities of this multi-sensor framework, which aim to enhance safety and improve surgical outcome, include: *gaze-guided object recognition and tracking, robotic manipulation, augmented visualisation of gaze relevant information, behavioural analysis and workflow segmentation based on perceptual information provided by the framework*. The proposed framework is expected to lead to a safer and more efficient surgical environment and provide improvements in healthcare delivery and outcome. By deploying the framework in several theatres could provide a large amount of anonymised data, which will help generate a large evidence base and critical mass to facilitate the establishment of Surgery 4.0.

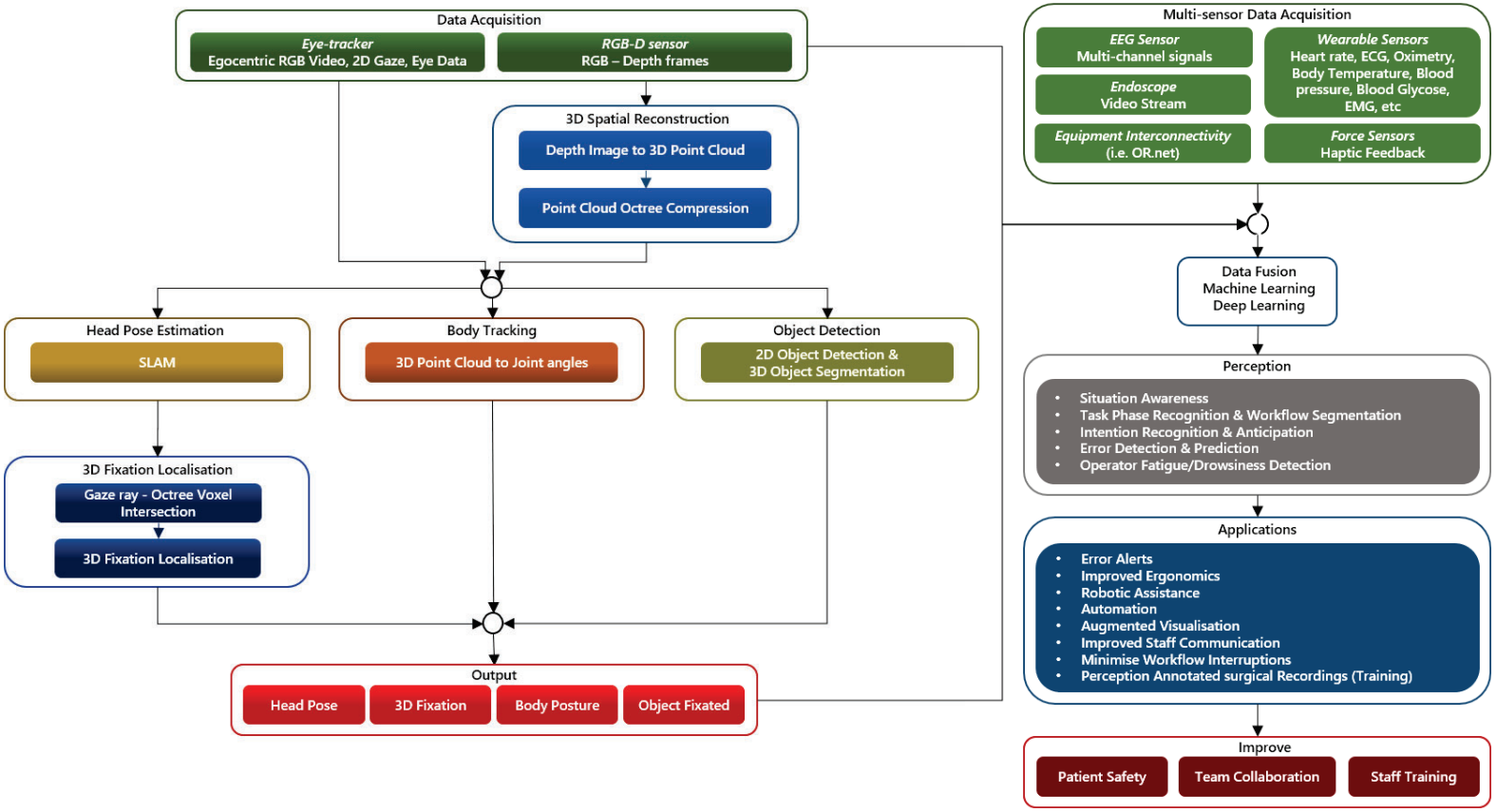


Figure 7.1: Multi-sensor data fused with the perceptually enabled data provided by the framework proposed in this thesis, can be used for a vast array of applications towards the improvement of patient safety, team collaboration and staff training.

Appendix A

Case Study: A Gaze-guided Robotic Laser [1]

This section describes the use of the framework to guide a robotic laser in the operating theatre. It is presented in [62] and is inspired by the concept of collaborative eye-tracking, which is demonstrated in robotic [21] and conventional laparoscopic [49] surgical settings, by sharing the visual attention of multiple collaborators on a screen. Results have shown to significantly improve verbal and non-verbal communication, task understanding, cooperation, task efficiency and outcome.

It should be noted that the laser-holding robotic arm serves no clinical use case as presented here. On this occasion the robot is used to demonstrate its integration and achieved accuracy within the framework using a SLAM approach for head pose estimation. Additionally, more economic ways are available for displaying one or more laser points in the theatre. Moreover, comparing to the framework methods presented in chapter 3, this work relies on static spatial 3D reconstruction and the ray-triangle intersection algorithm [227] to calculate the intersection between the gaze

[1] Content from this chapter was published as:
Gaze-contingent perceptually enabled interactions in the operating theatre. Kogkas A., Darzi A., Mylonas G. International Journal of Computer Assisted Radiology and Surgery, 12, 1131–1140 (2017), doi: 10.1007/s11548-017-1580-y.

ray and a triangle in 3D, thus providing the 3D fixation coordinates.

A.1 Methodology

A co-registered robot arm is used to point a laser pointer at the position of the resolved 3D fixation.

During initialisation, the eye-tracking glasses and the Kinect sensor are calibrated. The local coordinate systems (robot, 3D map extracted by the eye-tracker monocular RGB scene camera) are registered to the Kinect's world coordinate system and the laser pointer is aligned to the robot's end-effector. A 3D model of the operating theatre is then extracted by the Kinect sensor and the pose of the eye-tracker's scene camera is estimated within it using the SLAM technique (section 3.8.2). Subsequently, the 2D fixations provided by the eye-tracking glasses are mapped to 3D world coordinates and provided to the robot. Finally, the appropriate robot pose is estimated in order to highlight the 3D fixation with the laser attached to its end-effector.

A.1.1 Equipment

The robot arm is a UR5 by Universal Robots. It is a collaborative robot providing 6 degrees of freedom, $\pm 360^{\circ}$joint ranges, a reach radius of up to $850mm$ and $\pm 0.1mm$ repeatability. It weighs $18.4kg$ and is capable of maximum $5kg$ pay-load.

To highlight the 3D fixation in the theatre, a green laser diode is attached on the robot's end-effector using a 3D printed mount. As the laser beam is not exactly coincident with its z-axis, any alignment errors are corrected using a calibration step.

A.1.2 Calibration

The accuracy of the calibration process is of paramount importance. Four types of calibrations are performed:

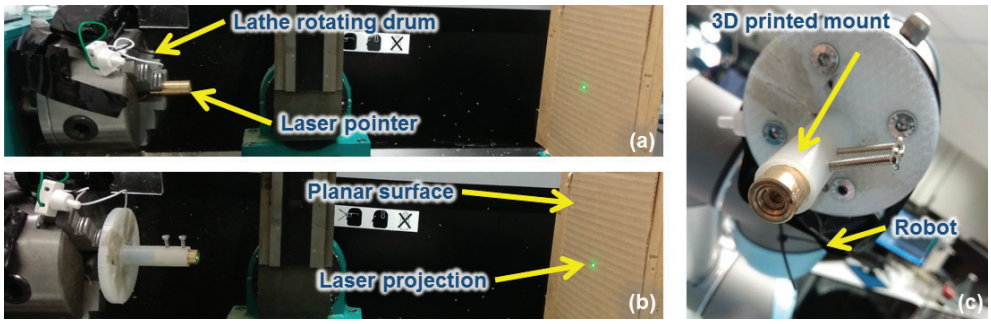


Figure A.1: The laser module's intrinsic calibration process

- Camera calibration for the eye-tracker's RGB scene camera
- User-specific eye-tracking calibration of the eye-tracking glasses
- RGB-depth calibration for the Microsoft Kinect sensor
- Laser module to robot's end-effector calibration

The first three calibration routines are described in section 3.5. The *laser module* requires intrinsic calibration, as it produces an offset angle of $\sim 0.8°$, which is significant for projections over large distances. A mechanical offset calibration is used to align the laser module's vector with the end-effector's z-axis. The laser module is calibrated using a 3D printed component and screws. The pointer is first mounted on a lathe's drum (Fig. A.1(a)). By rotating the lathe and observing the laser projection on a planar surface, the projection centre and the diode angular offset direction are determined. Then, the pointer is mounted on a 3D printed base (mounted on the lathe) making sure the offset direction vector intercepts the line connecting the 2 screws (Fig. A.1(b)). The screws are adjusted while the lathe rotates, until the laser projects accurately to the projection centre on the planar surface. Finally, it is mounted on the robot's end-effector (Fig. A.1(c)).

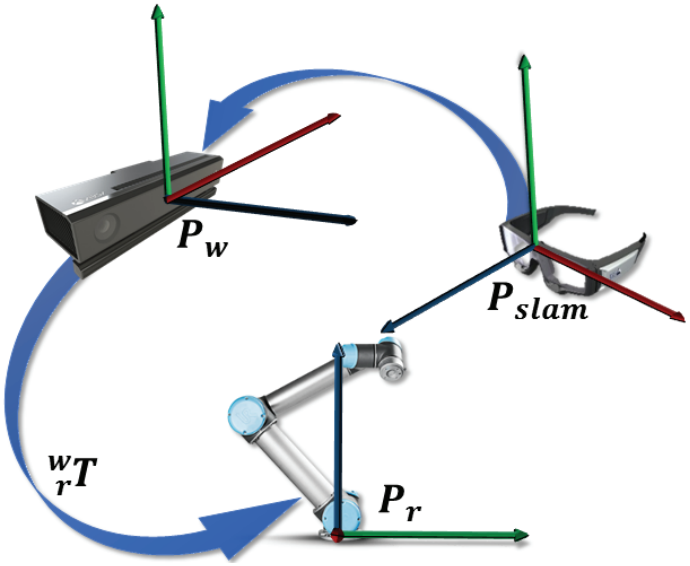


Figure A.2: The transformations among the coordinate systems

A.1.3 Registration

In the proposed system, we use the Kinect's coordinate system as the word frame of reference. To align multiple local coordinate systems to the global one, two main registrations are performed; a *SLAM-to-Kinect* (section 3.5) and a *Kinect-to-Robot registration* (Fig. A.2). Accurate *Kinect-to-Robot registration* is necessary since minor inaccuracies can lead to significant deviation from the desired waypoints. The coordinate system of the robot is defined with respect to its base. The manipulation of a 6-axis robot involves calculation of 3D coordinates and rotation vectors, defining its pose. Therefore, Kinect-to-Robot registration is performed off-line using a chessboard pattern on the robot's end-effector and the hand-eye calibration methodology presented in [274].

For every gaze-guided task, the TCP position is estimated in the world coordinate system. The robot receives the target pose in the robot coordinate system, calculated by

$$P_r =^{w}_{r} T * P_w \tag{A.1}$$

Where:
P_r is the TCP position in the robot coordinate system
P_w is the TCP position in the world coordinate system
${}^{w}_{r}T$ is the transformation matrix from the world to the robot coordinate system obtained using the hand-eye calibration method [274].

A.1.4 Robotic Laser Task

To highlight the 3D fixation in the operating room using a robotic laser, the estimated 3D fixation should be converted to a corresponding robot pose. To this end, the Kinect-to-Robot registration is not sufficient. We need to define one of the multiple poses with which the z-axis of the robot's end effector intersects the 3D fixation (Fig. A.3). A sphere with a predefined radius is defined and its centre is placed on a point along the z-axis of the robot. The intersection of the ray —defined by the coordinates of this centre point and the 3D fixation— with the sphere will be the translation of the robot's end effector. The rotation is defined by the z-axis of the robot's end-effector, which should be aligned with the line defined by the sphere intersection and the 3D fixation. The x- and y-axes are set arbitrarily. Finally, the pose is transformed to the robot's coordinate system and transmitted to it.

A.2 Experiments and Results

For the experimental evaluation of our framework, 20 targets are placed in the operating theatre (Fig. A.4). The distance range between the subject and the targets is $92cm$–$212cm$ and between the robot and the targets $42cm$–$193cm$. The task involves fixating on the targets for more than $4s$ and this process is repeated 3 times. The accuracy and the real-time performance of the system are evaluated over 60 fixations. The accuracy of the system can be affected by multiple factors: the eye-tracker intrinsic error, the head pose estimation (ORB-SLAM) error, the robot calibration

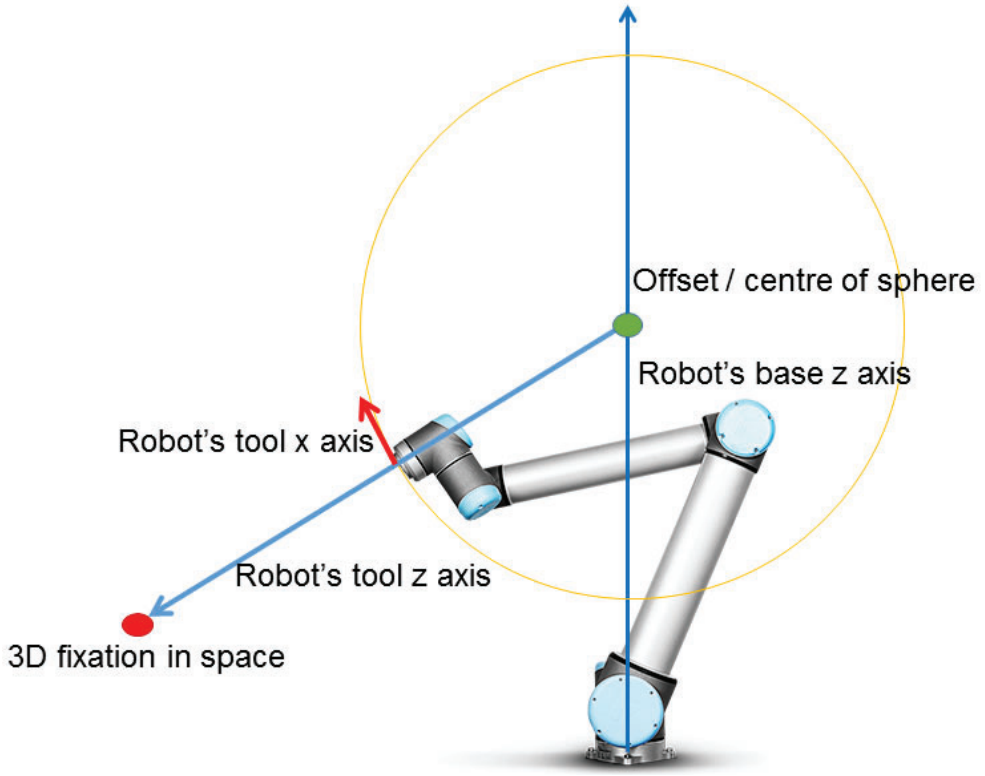


Figure A.3: Estimation of robot's pose to highlight the 3D fixation (sphere approach)

error and the Kinect sensor. In this validation we measure:

- The *eye-tracker error*, which is a 2D distance in pixels on the eye-tracker's scene-camera frame, expressed as a % of its resolution (720p) and based on comparing the actual and the expected 2D fixations.
- The *framework error* comparing the actual and the expected 3D fixations (compounded by the eye-tracker's error).
- The *robotic laser error* derived by the Kinect-to-robot calibration and the laser module's intrinsic offset, by manually repositioning the robot to accurately highlight the 3D targets.
- The *overall system error*, comparing 3D target coordinates with the

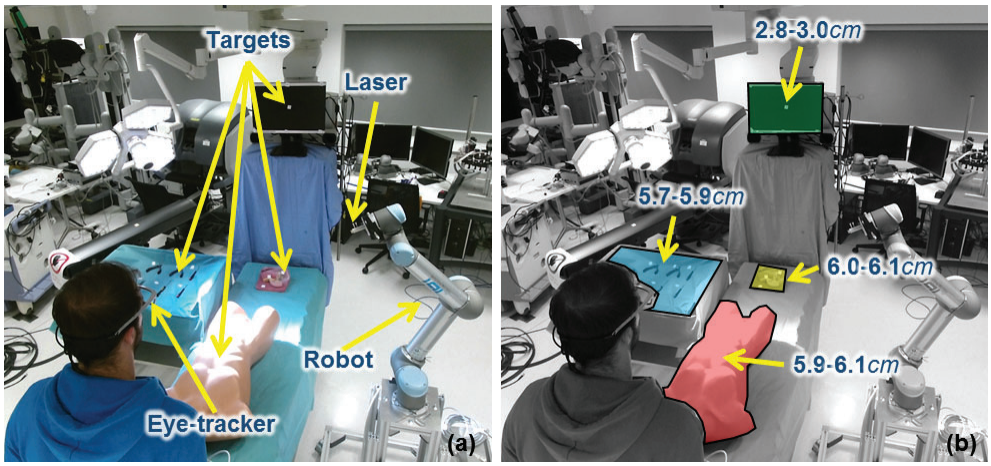


Figure A.4: (a) The experimental setup (view from the Kinect sensor): As the subject fixates on predefined targets, the pose of the eye-tracker scene camera is estimated. When a fixation is detected, the 2D gaze is mapped to 3D coordinates and the robotic laser highlights the fixated spot. (b) The error range within the main fixated areas of interest.

3D coordinates of the laser projection. This also depends on the geometry of the surface where the laser is projected.

The results summarised in Fig. A.5 show the system accuracy over all measured fixations. The overall system error is $3.98cm$.

It is of paramount importance to identify the contribution of each constituent component of the implementation (hardware and methodologies) to the overall system error (Fig. A.6). The Kinect sensor produces an average error of 2–$4mm$, but depending on the distance from the target this may increase to over $4mm$ [206]. This error propagates to multiple stages of the system. Eye-tracking may introduce variable error due to the parallax effect [228] occurring over large fixation distances. The Kinect-to-robot registration method used [274] exhibits an error of $0.75cm$ for calibration using ~ 25 poses. Last but not least, the head pose estimation is one the most significant stages of the framework and is speculated to introduce an error.

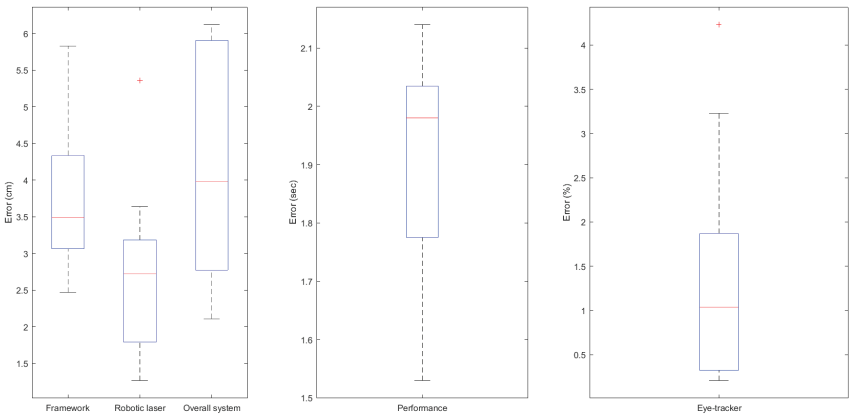


Figure A.5: Error analysis.

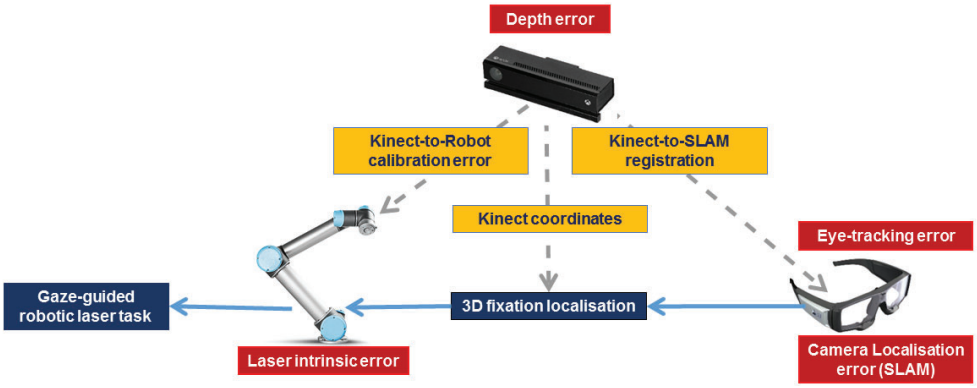


Figure A.6: Sources of error and their interaction. The Kinect sensor introduces a depth inaccuracy, which propagates to the system through its calibration with the robot, its registration with the SLAM local map and the 3D fixation localisation (in Kinect coordinates). Moreover, error is introduced and propagated towards the output of the system through the eye-tracker's inaccurate gaze estimation, the inaccuracy of the ORB-SLAM algorithm, which localises the camera within the 3D space, and the error produced by the offset of the laser pointer (reduced after its calibration)

Appendix B

Intellectual Property Re-use Permissions

B.1 IEEE Permissions

All images and textual content from IEEE publications re-used in this thesis have been referenced according to the IEEE guidelines listed on the IEEE website.

For the re-use of substantial material from the author's own publications the following has been obeyed:
The IEEE does not require individuals working on a thesis to obtain a formal reuse license. If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: ©[year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication].

For the re-use of figures of an IEEE copyrighted paper in a thesis the following has been obeyed:
The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant: In the case of illustrations or tabular material, we require that the copyright line ©[Year of original publication] IEEE

appear prominently with each reprinted figure and/or table.

B.2 Springer Permissions

The following licenses were obtained from Springer for re-use:

- "Free-View, 3D Gaze-Guided Robotic Scrub Nurse"
  License number: 4765871285831
- "Touchless interaction with software in interventional radiology and surgery: a systematic literature review"
  License number: 4764050908155
- "Laparoscopic visual field: Voice vs foot pedal interfaces for control of the AESOP robot"
  License number: 4764051293600
- "Gaze-contingent 3d control for focused energy ablation in robotic assisted surgery"
  License number: 4764080295889
- "Auditory display as feedback for a novel eye-tracking system for sterile operating room interaction"
  License number: 4764080668941
- "Surgical data science for next-generation interventions"
  License number: 4764080988062
- "FACTS-a computer vision system for 3D recovery and semantic mapping of human factors"
  License number: 4764081237032
- "RT-GENE: Real-time eye gaze estimation in natural environments"
  License number: 4764251167910

- "Segmentation and Guidance of Multiple Rigid Objects for Intra-operative Endoscopic Vision"
  License number: 4767181432818

B.3 Elsevier Permissions

The following licenses were obtained from Elsevier for re-use:

- "Hand gesture guided robot-assisted surgery based on a direct augmented reality interface"
  License number: 4764061221418

- "Context-based hand gesture recognition for the operating room"
  License number: 4764060389693

B.4 Permissions from Other Publishers

- Publisher: Georg Thieme Verlag KG
  "The Voice-Controlled Robotic Assist Scope Holder AESOP for the Endoscopic Approach to the Sella"
  License number: 4767700340002

- Publisher: IOP Publishing
  "Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain–machine interfaces"
  License number: 1017164-1

References

[1] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index)," *Advances in Psychology*, 1988.

[2] A. Mewes, B. Hensen, F. Wacker, C. Hansen, F. Wacker, and C. Hansen, "Touchless interaction with software in interventional radiology and surgery: a systematic literature review," *Int J CARS*, vol. 12, pp. 1–15, 2017.

[3] C.-A. O. A. O. Nathan, V. Chakradeo, K. Malhotra, H. D'Agostino, and R. Patwardhan, "The voice-controlled robotic assist scope holder AESOP for the endoscopic approach to the sella," *Skull Base*, vol. 16, no. 3, pp. 123–132, 2006.

[4] C. Doignon, F. Nageotte, and M. de Mathelin, "Segmentation and Guidance of Multiple Rigid Objects for Intra-operative Endoscopic Vision," in *Dynamical Vision* (R. Vidal, A. Heyden, and Y. Ma, eds.), (Berlin, Heidelberg), pp. 314–327, Springer Berlin Heidelberg, 2007.

[5] M. E. Allaf, S. V. Jackman, P. G. Schulam, J. A. Cadeddu, B. R. Lee, R. G. Moore, and L. R. Kavoussi, "Laparoscopic visual field Voice vs foot pedal interfaces for control of the AESOP robot," *Surgical Endoscopy*, vol. 12, no. 12, pp. 1415–1418, 1998.

[6] E. Carpintero, C. Pérez, R. Morales, N. García, A. Candela, and J. M. Azorín, "Development of a robotic scrub nurse for the operating theatre," *2010 3rd IEEE RAS and EMBS International*

Conference on Biomedical Robotics and Biomechatronics, BioRob 2010, pp. 504–509, 2010.

[7] M. G. Jacob and J. P. Wachs, "Context-based hand gesture recognition for the operating room," *Pattern Recognition Letters*, vol. 36, pp. 196–203, 2014.

[8] J. Wachs, H. Stern, Y. Edan, and M. Gillam, "Real-Time Hand Gesture Interface for Browsing Medical Images," *Proc. Int. J Intel. Comp. Med.*, vol. 2, no. 1, pp. 15–25, 2007.

[9] M. G. Jacob, Y.-T. Li, and J. P. Wachs, "Gestonurse: A multimodal robotic scrub nurse," *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, vol. 1, pp. 153–154, 2012.

[10] J. P. Wachs, K. Vujjeni, E. T. Matson, and S. Adams, ""A window on tissue"-Using facial orientation to control endoscopic views of tissue depth," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, no. November, pp. 935–938, IEEE, 2010.

[11] R. Wen, W.-L. Tay, B. P. Nguyen, C.-B. Chng, and C.-K. Chui, "Hand gesture guided robot-assisted surgery based on a direct augmented reality interface," *Computer methods and programs in biomedicine*, vol. 116, no. 2, pp. 68–80, 2014.

[12] J. Hettig, A. A. Mewes, O. Riabikin, M. Skalej, B. Preim, and C. Hansen, "Exploration of 3D medical image data for interventional radiology using myoelectric gesture control," in *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*, pp. 177–185, Eurographics Association, 2015.

[13] G. M. Rosa and M. L. Elizondo, "Use of a gesture user interface as a touchless image navigation system in dental surgery: Case series report," *Imaging science in dentistry*, vol. 44, no. 2, pp. 155–160, 2014.

[14] E. Whitmire, L. Trutoiu, R. Cavin, D. Perek, B. Scally, J. Phillips, and S. Patel, "EyeContact: Scleral coil eye tracking for virtual reality," in *International Symposium on Wearable Computers*, 2016.

[15] "CHRONOS VISION GmbH." https://www.chronos-vision.de/.

[16] A. Bulling, D. Roggen, and G. Tröster, "What's in the eyes for context-awareness?," *IEEE Pervasive Computing*, vol. 10, no. 2, pp. 48–57, 2011.

[17] "Tobii AB." https://www.tobiipro.com/.

[18] "The EyeTribe." https://theeyetribe.com/.

[19] "SMI - SensoMotoric Instruments." http://www.smivision.com/.

[20] S. Li, X. Zhang, and J. D. Webb, "3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2824–2835, 2017.

[21] K.-W. W. Kwok, L.-W. W. Sun, G. P. Mylonas, D. R. C. James, F. Orihuela-Espina, and G.-Z. Z. Yang, "Collaborative gaze channelling for improved cooperation during robotic assisted surgery," *Annals of Biomedical Engineering*, vol. 40, no. 10, pp. 2156–2167, 2012.

[22] D. Stoyanov, G. P. Mylonas, and G.-Z. Yang, "Gaze-contingent 3d control for focused energy ablation in robotic assisted surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11, pp. 347–355, Springer, 2008.

[23] K. Fujii, G. Gras, A. Salerno, and G. Z. Yang, "Gaze gesture based human robot interaction for laparoscopic surgery," *Medical Image Analysis*, vol. 44, pp. 196–214, 2018.

[24] D. P. Noonan, G. P. Mylonas, J. Shang, C. J. Payne, A. Darzi, and G.-Z. Z. Yang, "Gaze contingent control for an articulated mechatronic laparoscope," in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, pp. 759–764, IEEE, IEEE, 2010.

[25] C. Staub, S. Can, B. Jensen, A. Knoll, and S. Kohlbecher, "Human-computer interfaces for interaction with surgical tools in robotic surgery," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pp. 81–86, IEEE, 2012.

[26] B. Hatscher, M. Luz, L. E. Nacke, N. Elkmann, V. Müller, and C. Hansen, "GazeTap: towards hands-free interaction in the operating room," *Proceedings of the 19th ACM international conference on multimodal interaction*, no. February 2018, pp. 243–251, 2017.

[27] D. Black, M. Unger, N. Fischer, R. Kikinis, H. Hahn, T. Neumuth, and B. Glaser, "Auditory display as feedback for a novel eye-tracking system for sterile operating room interaction," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 1, pp. 37–45, 2018.

[28] Karl Storz Endoscopy, "KARL STORZ OR1™ NEO." https://www.karlstorz.com/gb/en/karl-storz-or1-neo-a-truly-integrated-operating-theatre.htm.

[29] L. Maier-hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin, "Surgical data science for next- generation interventions," vol. 1, no. September, 2017.

[30] W. Abbott and A. Faisal, "Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain–machine interfaces," *Journal of Neural Engineering*, vol. 9, no. 4, p. 046016, 2012.

[31] C. McMurrough, C. Conly, V. Athitsos, and F. Makedon, "3D point of gaze estimation using head-mounted RGB-D cameras," p. 283, Association for Computing Machinery (ACM), 10 2012.

[32] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 4511–4520, 2015.

[33] L. Paletta, K. Santner, G. Fritz, A. Hofmann, G. Lodron, G. Thallinger, and H. Mayer, "FACTS-a computer vision system for 3D recovery and semantic mapping of human factors," in *International Conference on Computer Vision Systems*, no. July, pp. 62–72, Springer, 2013.

[34] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11214 LNCS, pp. 339–357, 2018.

[35] M. Shah, "Solving the Robot-World/Hand-Eye Calibration Problem Using the Kronecker Product," *Journal of Mechanisms and Robotics*, vol. 5, p. 31007, 6 2013.

[36] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2548–2555, IEEE, 2011.

[37] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[38] G. Berci, E. H. Phillips, and F. Fujita, "The operating room of the future: what, when and why?," *Surgical endoscopy*, vol. 18, no. 1, pp. 1–5, 2003.

[39] F. J. Seagull, G. R. Moses, and A. E. Park, "Pillars of a Smart, Safe Operating Room," pp. 1–13, 2008.

[40] R. Bharathan, R. Aggarwal, and A. Darzi, "Operating room of the future," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 27, no. 3, pp. 311–322, 2013.

[41] Karl Storz Endoscopy, "Karl Storz OR1™." https://www.karlstorz.com/at/en/karl-storz-or1.htm.

[42] M. Rockstroh, S. Franke, M. Hofer, A. Will, M. Kasparick, B. Andersen, and T. Neumuth, "OR.NET: multi-perspective qualitative evaluation of an integrated operating room based on IEEE 11073 SDC," *International Journal of Computer Assisted Radiology and Surgery*, 2017.

[43] C.-h. Fan, "Driver Fatigue Detection Based," pp. 7–12, 2004.

[44] H. Jarodzka, K. Scheiter, P. Gerjets, T. van Gog, and M. Dorr, "How to Convey Perceptual Skills by Displaying Experts' Gaze Data," *Cogsci*, pp. 2920–2925, 2009.

[45] S. P. Marshall, "Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity," 2000.

[46] K. Kilingaru, J. W. Tweedale, S. Thatcher, and L. C. Jain, "Monitoring pilot 'Situation Awareness'," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 3, pp. 457–466, 2013.

[47] R. Murray-smith, *Eye Gaze Tracking for Human Computer Interaction*. PhD thesis, 2010.

[48] D. P. McMullen, G. Hotson, K. D. Katyal, B. a. Wester, M. S. Fifer, T. G. McGee, A. Harris, M. S. Johannes, R. J. Vogelstein, A. D. Ravitz, W. S. Anderson, N. V. Thakor, and N. E. Crone, "Demonstration of a semi-autonomous hybrid brain-machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 784–796, 2014.

[49] A. S. A. Chetwood, K.-W. Kwok, L.-W. Sun, G. P. Mylonas, J. Clark, A. Darzi, and G.-Z. Yang, "Collaborative eye tracking - a potential training tool in laparoscopic surgery," *Surgical Endoscopy*, vol. 26, no. 7, pp. 2003–2009, 2012.

[50] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair," *Surgical Endoscopy*, vol. 29, no. 2, pp. 405–413, 2015.

[51] S. Eivazi, R. Bednarik, M. Tukiainen, M. Von Und Zu Fraunberg, V. Leinonen, and J. E. Jääskeläinen, "Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings," *Eye Tracking Research and Applications Symposium (ETRA)*, vol. 1, no. 212, pp. 377–380, 2012.

[52] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye metrics as an objective assessment of surgical skill.," *Annals of surgery*, vol. 252, no. 1, pp. 177–182, 2010.

[53] M. R. Wilson, S. J. Vine, E. Bright, R. S. W. Masters, D. Defriend, and J. S. McGrath, "Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: a randomized,

controlled study," *Surgical Endoscopy*, vol. 25, no. 12, pp. 3731–3739, 2011.

[54] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Eye tracking for skills assessment and training: a systematic review," *Journal of Surgical Research*, vol. 191, no. 1, pp. 169–178, 2014.

[55] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 878–885, 2005.

[56] F. DRAMAS, S. J. THORPE, and C. JOUFFRAIS, "Artificial Vision for the Blind: a Bio-Inspired Algorithm for Objects and Obstacles Detection," *International Journal of Image and Graphics*, vol. 10, no. 04, pp. 531–544, 2010.

[57] B. F. G. Katz, F. Dramas, G. Parseihian, O. Gutierrez, S. Kammoun, A. Brilhault, L. Brunet, M. Gallay, B. Oriola, M. Auvray, P. Truillet, M. Denis, S. Thorpe, and C. Jouffrais, "NAVIG: Guidance system for the visually impaired using virtual augmented reality," *Technology and Disability*, vol. 24, no. 2, pp. 163–178, 2012.

[58] A. Kolarow, M. Brauckmann, M. Eisenbach, K. Schenk, E. Einhorn, K. Debes, and H. M. Gross, "Vision-based hyper-real-time object tracker for robotic applications," *IEEE International Conference on Intelligent Robots and Systems*, pp. 2108–2115, 2012.

[59] M. Litzenberger, C. Posch, D. Bauer, a. Belbachir, P. Schon, B. Kohn, and H. Garn, "Embedded Vision System for Real-Time Object Tracking using an Asynchronous Transient Vision Sensor," *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, pp. 173–178, 2006.

[60] B. Benfold and I. Reid, "Stable Multi-Target Tracking in Real-Time Surveillance Video," *Cvpr*, pp. 3457–3464, 2011.

[61] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1050–1058, 2013.

[62] A. A. Kogkas, A. Darzi, and G. P. Mylonas, "Gaze-contingent perceptually enabled interactions in the operating theatre," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2017.

[63] C. K. Christian, M. L. Gustafson, E. M. Roth, T. B. Sheridan, T. K. Gandhi, K. Dwyer, M. J. Zinner, and M. M. Dierks, "A prospective study of patient safety in the operating room," *Surgery*, vol. 139, no. 2, pp. 159–173, 2006.

[64] H. Alfredsdottir and K. Bjornsdottir, "Nursing and patient safety in the operating room.," *Journal of advanced nursing*, vol. 61, no. 1, pp. 29–37, 2008.

[65] R. L. Helmreich and J. M. Davies, "Human factors in the operating room: interpersonal determinants of safety, efficiency and morale," *Baillière's Clinical Anaesthesiology*, vol. 10, no. 2, pp. 277–295, 1996.

[66] World Alliance for Patient Safety WHO, *WHO Guidelines for Safe Surgery*. 2008.

[67] A. L. Halverson, J. T. Casey, J. Andersson, K. Anderson, C. Park, A. W. Rademaker, and D. Moorman, "Communication failure in the operating room," *Surgery*, vol. 149, no. 3, pp. 305–310, 2011.

[68] D. W. Rattner and A. Park, "Advanced Devices for the Operating Room of the Future," *Surgical Innovation*, vol. 10, no. 2, pp. 85–89, 2003.

[69] K. Cleary, A. Kinsella, and S. K. Mun, "OR 2020 workshop report: Operating room of the future," *International Congress Series*, vol. 1281, pp. 832–838, 2005.

[70] H. G. Kenngott, M. Apitz, M. Wagner, A. A. Preukschas, and S. Speidel, "Paradigm shift : cognitive surgery," vol. 2, no. 3, pp. 139–143, 2017.

[71] C. A. Velasquez, R. Mazhar, A. Chaikhouni, T. Zhou, and J. P. Wachs, "Taxonomy of communications in the operating room," in *Advances in Intelligent Systems and Computing*, 2018.

[72] L. L., R. R., E. S., R. G., and D. I., "Team communications in the operating room: Talk patterns, sites of tension, and implications for novices," *Academic Medicine*, vol. 77, no. 3, pp. 232–237, 2002.

[73] J. Moss and Y. Xiao, "Improving Operating Room Coordination: Communication Pattern Assessment," *Journal of Nursing Administration*, vol. 34, no. 2, pp. 93–100, 2004.

[74] L. L. Leape, "Error in medicine.," *JAMA : the journal of the American Medical Association*, vol. 272, no. 23, pp. 1851–1857, 1994.

[75] M. S. Bogner, *Human error in medicine*. 2018.

[76] J. T. Reason and K. Mycielska, *Absent-minded?: The psychology of mental lapses and everyday errors*. Prentice Hall, 1982.

[77] J. Reason, *Human error*. Cambridge university press, 1990.

[78] D. T. Risser, M. M. Rice, M. L. Salisbury, R. Simon, G. D. Jay, and S. D. Berns, "The potential for improved teamwork to reduce medical errors in the emergency department," *Annals of Emergency Medicine*, 1999.

[79] M. Leonard, S. Graham, and D. Bonacum, "The human factor: The critical importance of effective teamwork and communication in providing safe care," 2004.

[80] J. Firth-Cozens and D. Mowbray, "Leadership and the quality of care," *Quality and Safety in Health Care*, 2010.

[81] J. Firth-Cozens, "Multidisciplinary teamwork: The good, bad, and everything in between," 2001.

[82] L. Lingard, S. Espin, S. Whyte, G. Regehr, G. R. Baker, R. Reznick, J. Bohnen, B. Orser, D. Doran, and E. Grober, "Communication failures in the operating room: an observational classification of recurrent types and effects.," *Quality & Safety In Health Care*, vol. 13, no. 5, pp. 330–4, 2004.

[83] J. Carthey, M. R. De Leval, D. J. Wright, V. T. Farewell, and J. T. Reason, "Behavioural markers of surgical excellence," *Safety Science*, vol. 41, no. 5, pp. 409–425, 2003.

[84] J. Clayton, A. N. Isaacs, and I. Ellender, "Perioperative nurses' experiences of communication in a multicultural operating theatre: A qualitative study," *International Journal of Nursing Studies*, vol. 54, pp. 7–15, 2016.

[85] H. AFKARI, *INTERACTION IN THE MICRO-NEUROSURGERY OPERATING ROOM: The potentials for gaze-based interaction with the surgical microscope*. PhD thesis, University of Eastern Finland, 2018.

[86] M. A. Makary, J. B. Sexton, J. A. Freischlag, C. G. Holzmueller, E. A. Millman, L. Rowen, and P. J. Pronovost, "Operating Room Teamwork among Physicians and Nurses: Teamwork in the Eye of the Beholder," *Journal of the American College of Surgeons*, vol. 202, no. 5, pp. 746–752, 2006.

[87] N. M. Saufl, "Universal protocol for preventing wrong site, wrong procedure, wrong person surgery," 2004.

[88] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "An evaluation of outcome from intensive care in major medical centers," *Annals of Internal Medicine*, 1986.

[89] J. Defontes and S. Surbida, "Preoperative Safety Briefing Project.," *The Permanente journal*, 2004.

[90] R. L. Helmreich, H. C. Foushee, R. Benson, and W. Russini, "Cockpit resource management: Exploring the attitude-performance linkage," *Aviation Space and Environmental Medicine*, 1986.

[91] B. Z. Posner and W. A. Randolph, "Perceived situational moderators of the relationship between role ambiguity, job satisfaction, and effectiveness," *Journal of Social Psychology*, 1979.

[92] L. H. Aiken, S. P. Clarke, D. M. Sloane, J. A. Sochalski, R. Busse, H. Clarke, P. Giovannetti, J. Hunt, A. M. Rafferty, and J. Shamian, "Nurses' reports on hospital care in five countries," *Health Affairs*, 2001.

[93] M. Manojlovich and B. DeCicco, "Healthy Work Environments, Nurse-Physician Communication, and Patients' Outcomes," *American Journal of Critical Care*, 2007.

[94] E. Matthys, R. Remmen, and P. Van Bogaert, "An overview of systematic reviews on the collaboration between physicians and nurses and the impact on patient outcomes: what can we learn in primary care?,"

[95] D. Stalpers, B. J. de Brouwer, M. J. Kaljouw, and M. J. Schuurmans, "Associations between characteristics of the nurse work environment and five nurse-sensitive patient outcomes in hospitals: A systematic review of literature," 2015.

[96] C. Schmalenberg, M. Kramer, C. R. King, M. Krugman, C. Lund, D. Poduska, and D. Rapp, "Excellence through evidence: Securing collegial/collaborative nurse-physician relationships, part 1," *Journal of Nursing Administration*, 2005.

[97] J. S. Martin, W. Ummenhofer, T. Manser, and R. Spirig, "Interprofessional collaboration among nurses and physicians: Making a difference in patient outcome," 2010.

[98] A. Xyrichis and K. Lowton, "What fosters or prevents interprofessional teamworking in primary and community care? A literature review," *International Journal of Nursing Studies*, 2008.

[99] I. Supper, O. Catala, M. Lustman, C. Chemla, Y. Bourgueil, and L. Letrilliart, "Interprofessional collaboration in primary health care: A review of facilitators and barriers perceived by involved actors," *Journal of Public Health (United Kingdom)*, 2015.

[100] C. Davies, "Getting health professionals to work together," *BMJ*, 2000.

[101] G. Fitzpatrick and G. Ellingsen, *A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas*, vol. 22. 2013.

[102] F. Dexter, A. Macario, R. D. Traub, M. Hopwood, and D. A. Lubarsky, "An operating room scheduling strategy to maximize the use of operating room block time: Computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time," *Anesthesia and Analgesia*, 1999.

[103] P. Scupelli, Y. Xiao, and S. Fussell, "Supporting coordination in surgical suites: physical aspects of common information spaces," in *Proceedings of the*, 2010.

[104] N. Stylopoulos and D. Rattner, "Robotics and ergonomics," 2003.

[105] G. Palmer, J. H. Abernathy, G. Swinton, D. Allison, J. Greenstein, S. Shappell, K. Juang, and S. T. Reeves, "Realizing Improved Patient Care through Human-centered Operating Room Design," *Anesthesiology*, 2013.

[106] S. P. Rodrigues, A. M. Wever, J. Dankelman, and F. W. Jansen, "Risk factors in patient safety: Minimally invasive surgery versus conventional surgery," *Surgical Endoscopy*, 2012.

[107] T. Catanzarite, J. Tan-Kim, E. L. Whitcomb, and S. Menefee, "Ergonomics in Surgery: A Review," 2018.

[108] U. Matern and S. Koneczny, "Safety, hazards and ergonomics in the operating room," in *Surgical Endoscopy and Other Interventional Techniques*, 2007.

[109] C. C. H. Stucky, K. D. Cromwell, R. K. Voss, Y. J. Chiang, K. Woodman, J. E. Lee, and J. N. Cormier, "Surgeon symptoms, strain, and selections: Systematic review and meta-analysis of surgical ergonomics," 2018.

[110] F. J. Seagull, "Disparities between industrial and surgical ergonomics," in *Work*, vol. 41, pp. 4669–4672, 2012.

[111] S. Janki, E. E. Mulder, J. N. IJzermans, and T. C. Tran, "Ergonomics in the operating room," *Surgical Endoscopy*, 2017.

[112] P. Katz, "Ritual in the Operating Room," *Ethnology*, 2007.

[113] R. Johnson, K. O'Hara, A. Sellen, C. Cousins, and A. Criminisi, "Exploring the potential for touchless interaction in image-guided interventional radiology," *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, no. May, pp. 3323–3332, 2011.

[114] D. Black, M. Unger, N. Fischer, R. Kikinis, H. Hahn, T. Neumuth, and B. Glaser, "Auditory display as feedback for a novel eye-tracking

system for sterile operating room interaction," *International Journal of Computer Assisted Radiology and Surgery*, 2018.

[115] M. Schultz, J. Gill, S. Zubairi, R. Huber, and F. Gordin, "Bacterial Contamination of Computer Keyboards in a Teaching Hospital," *Infection Control & Hospital Epidemiology*, vol. 24, no. 04, pp. 302–303, 2003.

[116] H. M. Mentis, K. O'Hara, A. Sellen, and R. Trivedi, "Interaction Proxemics and Image Use in Neurosurgery," 2012.

[117] C. Graetzel, T. Fong, S. Grange, C. Baur, C. Grätzel, T. Fong, S. Grange, and C. Baur, "A non-contact mouse for surgeon-computer interaction," *Technology and Health Care*, vol. 12, no. 3, pp. 245–257, 2004.

[118] K. O'Hara, N. Dastur, T. Carrell, G. Gonzalez, A. Sellen, G. Penney, A. Varnavas, H. Mentis, A. Criminisi, R. Corish, and M. Rouncefield, "Touchless interaction in surgery," *Communications of the ACM*, vol. 57, no. 1, pp. 70–77, 2014.

[119] J. P. Wachs, H. I. Stern, Y. Edan, M. Gillam, J. Handler, C. Feied, and M. Smith, "A Gesture-based Tool for Sterile Browsing of Radiology Images," *Journal of the American Medical Informatics Association*, vol. 15, no. 3, pp. 321–323, 2008.

[120] K. O'hara, R. Harper, H. Mentis, A. Sellen, and A. Taylor, "On the naturalness of touchless," *ACM Transactions on Computer-Human Interaction*, vol. 20, no. 1, pp. 1–25, 2013.

[121] M. A. Van Veelen, C. J. Snijders, E. Van Leeuwen, R. H. M. Goossens, and G. Kazemier, "Improvement of foot pedals used during surgery based on new ergonomic guidelines," *Surgical Endoscopy and Other Interventional Techniques*, 2003.

[122] C. Vara-Thorbeck, V. F. Muñoz, R. Toscano, J. Gomez, J. Fernández, M. Felices, and A. Garcia-Cerezo, "A new robotic endoscope manipulator: A preliminary trial to evaluate the performance of a voice-operated industrial robot and a human assistant in several simulated and real endoscopic operations," *Surgical Endoscopy*, vol. 15, no. 9, pp. 924–927, 2001.

[123] A. Perrakis and W. Hohenberger, "Integrated operation systems and voice recognition in minimally invasive surgery : comparison of two systems," pp. 575–579, 2013.

[124] K. McLaughlin, "Microsoft Partners See Kinect Going Beyond Games." https://www.crn.com/news/applications-os/225700575/microsoft-partners-see-kinect-going-beyond-games.htm.

[125] J.-F. F. Collumeau, E. Nespoulous, H. Laurent, and B. Magnain, "Simulation interface for gesture-based remote control of a surgical lighting arm," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4670–4675, IEEE, 2013.

[126] R. H. Gong, O. Guler, M. Kurkluoglu, J. Lovejoy, and Z. Yaniv, "Interactive initialization of 2D/3D rigid registration," *Medical physics*, vol. 40, no. 12, p. 121911, 2013.

[127] S. K. Herniczek, A. Lasso, T. Ungi, and G. Fichtinger, "Feasibility of a touch-free user interface for ultrasound snapshot-guided nephrostomy," in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9036, p. 90362F, International Society for Optics and Photonics, 2014.

[128] A. Nishikawa, T. Hosoi, K. Koara, D. Negoro, A. Hikita, S. Asano, H. Kakutani, F. Miyazaki, M. Sekimoto, and M. Yasui, "FAce MOUSe: A novel human-machine interface for controlling the

position of a laparoscope," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 825–841, 2003.

[129] B. M. Kraft, C. Jäger, K. Kraft, B. J. Leibl, and R. Bittner, "The AESOP robot system in laparoscopic surgery: Increased risk or advantage for surgeon and patient?," *Surgical Endoscopy And Other Interventional Techniques*, vol. 18, no. 8, pp. 1216–1223, 2004.

[130] R. J. K. Jacob, "What you look at is what you get: eye movement-based interaction techniques," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 11–18, ACM, 1990.

[131] P. Majaranta and A. Bulling, "Eye Tracking and Eye-Based Human–Computer Interaction," in *Advances in Physiological Computing* (S. H. Fairclough and K. Gilleade, eds.), pp. 39–65, London: Springer London, 2014.

[132] B. Velichkovsky, A. Sprenger, and P. Unema, "Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem"," in *Human-Computer Interaction INTERACT'97*, pp. 509–516, Springer, 1997.

[133] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 71–78, ACM, 2000.

[134] P. Blignaut, "Fixation identification: The optimum threshold for a dispersion algorithm," *Attention, Perception, & Psychophysics*, vol. 71, no. 4, pp. 881–895, 2009.

[135] R. Bednarik, H. Vrzakova, and M. Hradis, "What do you want to do next: a novel approach for intent prediction in gaze-based interaction," in *Proceedings of the symposium on eye tracking research and applications*, pp. 83–90, ACM, 2012.

[136] R. Bixler and S. D'Mello, "Toward fully automated person-independent detection of mind wandering," in *International Conference on User Modeling, Adaptation, and Personalization*, pp. 37–48, Springer, 2014.

[137] S. Li and X. Zhang, "Implicit Intention Communication in Human – Robot Interaction Through Visual Behavior Studies," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 437–448, 2017.

[138] Y. Cao, S. Miura, Y. Kobayashi, K. Kawamura, S. Sugano, and M. G. Fujie, "Pupil variation applied to the eye tracking control of an endoscopic manipulator," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 531–538, 2016.

[139] K. Harezlak and P. Kasprowski, "Application of eye tracking in medicine: A survey, research issues and challenges," *Computerized Medical Imaging and Graphics*, vol. 65, pp. 176–190, 2018.

[140] E. B. Huey, *The psychology and pedagogy of reading*. The Macmillan Company, 1908.

[141] D. A. Robinson, "Movement Using a Scieral Search in a Magnetic Field," *IEEE Transactions on Bio-Medical Electronics*, vol. 10, no. 4, pp. 137–145, 1963.

[142] G. Z. Yang, L. Dempere-Marco, X. P. Hu, and A. Rowe, "Visual search: Psychophysical models and practical applications," *Image and Vision Computing*, vol. 20, no. 4, pp. 291–305, 2002.

[143] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg, "Detecting eye contact using wearable eye-tracking glasses," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, p. 699, 2012.

[144] S. H. Fairclough and K. Gilleade, *Advances in physiological computing*. Springer, 2014.

[145] G. P. Mylonas, A. Darzi, and G. Z. Yang, "Gaze-contingent control for minimally invasive robotic surgery.," *Computer Aided Surgery*, vol. 11, pp. 256–66, 1 2006.

[146] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior research methods, instruments, and computers*, vol. 34, no. 4, pp. 455–470, 2002.

[147] M. Carter, J. Newn, E. Velloso, and F. Vetere, "Remote gaze and gesture tracking on the microsoft kinect: Investigating the role of feedback," in *OzCHI 2015: Being Human - Conference Proceedings*, pp. 167–176, 2015.

[148] A. L. Yarbus, "Eye movements and vision," *Neuropsychologia*, 1967.

[149] A. Bojko, *Eye tracking the user experience*. 2013.

[150] M. Y. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas, "Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living," in *IEEE International Conference on Intelligent Robots and Systems*, 2018.

[151] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using hypernetworks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3060–3065, IEEE, 2018.

[152] H. Wang and B. E. Shi, "Gaze awareness improves collaboration efficiency in a collaborative assembly task," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, p. 88, ACM, 2019.

[153] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2009.

[154] R. J. K. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," in *The mind's eye*, pp. 573–605, Elsevier, 2003.

[155] C. Ware and H. H. Mikaelian, "An evaluation of an eye tracker as a device for computer input2," *Acm sigchi bulletin*, vol. 18, no. 4, pp. 183–188, 1987.

[156] T. E. Hutchinson, K. P. White, W. N. Martin, K. C. Reichert, and L. A. Frey, "Human-computer interaction using eye-gaze input," *IEEE Transactions on systems, man, and cybernetics*, vol. 19, no. 6, pp. 1527–1534, 1989.

[157] T. L. Chen, M. Ciocarlie, S. Cousins, P. Grice, K. Hawkins, K. Hsiao, C. Kemp, C. H. King, D. Lazewatsky, A. E. Leeper, H. Nguyen, A. Paepcke, C. Pantofaru, W. Smart, and L. Takayama, "Robots for humanity: Using assistive robotics to empower people with disabilities," *IEEE Robotics and Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.

[158] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, Context-aware Robotic System for Assisted Reaching and Grasping," 2018.

[159] E. Wästlund, K. Sponseller, and O. Pettersson, "What you see is where you go: Testing a gaze-driven power wheelchair for individuals with severe multiple disabilities," in *Eye Tracking Research and Applications Symposium (ETRA)*, 2010.

[160] M. Subramanian, N. Songur, D. Adjei, P. Orlov, and A. A. Faisal, "A.Eye Drive: Gaze-based semi-autonomous wheelchair interface," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2019.

[161] I. Tong, O. Mohareri, S. Tatasurya, C. Hennessey, and S. Salcudean, "A retrofit eye gaze tracker for the da Vinci and its integration in task execution using the da Vinci Research Kit," in *IEEE International Conference on Intelligent Robots and Systems*, 2015.

[162] Z. Li, I. Tong, L. Metcalf, C. Hennessey, and S. E. Salcudean, "Free Head Movement Eye Gaze Contingent Ultrasound Interfaces for the da Vinci Surgical System," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2137–2143, 2018.

[163] D. P. Noonan, G. P. Mylonas, A. Darzi, and G.-Z. Yang, "Gaze contingent articulated robot control for robot assisted minimally invasive surgery," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1186–1191, 2008.

[164] G. P. Mylonas, K.-W. W. Kwok, A. Darzi, and G.-Z. Z. Yang, "Gaze-contingent motor channelling and haptic constraints for minimally invasive robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 5242 LNCS, pp. 676–683, Springer, 2008.

[165] M. Visentini-Scarzanella, G. P. Mylonas, D. Stoyanov, and G.-Z. Z. Yang, "i-BRUSH: A gaze-contingent virtual paintbrush for dense 3D reconstruction in robotic assisted surgery," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5761 LNCS, pp. 353–360, 2009.

[166] N. T. Clancy, G. P. Mylonas, G.-Z. Yang, and D. S. Elson, "Gaze-contingent autofocus system for robotic-assisted minimally invasive surgery," *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5396–5399, 2011.

[167] S. M. Ali, L. A. Reisner, B. King, A. Cao, G. Auner, M. Klein, and A. K. Pandya, "Eye gaze tracking for endoscopic camera posi-

tioning: an application of a hardware/software interface developed to automate Aesop.," *Studies in health technology and informatics*, 2008.

[168] K. Fujii, A. Salerno, K. Sriskandarajah, K. W. Kwok, K. Shetty, and G. Z. Yang, "Gaze contingent cartesian control of a robotic arm for laparoscopic surgery," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3582–3589, 2013.

[169] E. R. Morales, D. Brasset, and P. Invernizzi, "Robotized surgery system with improved control," 6 2016.

[170] M. Unger, D. Black, N. M. Fischer, T. Neumuth, and B. Glaser, "Design and evaluation of an eye tracking support system for the scrub nurse," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 15, no. 1, p. e1954, 2019.

[171] M. D. Ballard, "System and method of tracking surgical sponges," 2004.

[172] P. Köhler, B. Six, and J. S. Michels, "Industry 4.0 : an overview from the perspective of a German-headquartered firm," *Robotica*, vol. 105, pp. 8–12, 2016.

[173] Wikipedia The Free Encyclopedia, "Industry 4.0." https://en.wikipedia.org/wiki/Industry_4.0 (accessed: 20 July 2019).

[174] S. Hirides and P. Hirides, "Surgery 4.0 vs . 4th Generation Robots : Clash of the Titans in the Near Future of Robotic Surgery," *International Journal of Surgery : Open Access*, pp. 1–3, 2018.

[175] M. Chand, N. Ramachandran, D. Stoyanov, and L. Lovat, "Robotics, artificial intelligence and distributed ledgers in surgery: data is key!," *Techniques in Coloproctology*, vol. 22, no. 9, pp. 645–648, 2018.

[176] F. Kühn, M. Leucker, and A. Mildner, "OR.NET - Approaches for risk analysis and measures of dynamically interconnected medical devices," in *OpenAccess Series in Informatics*, 2014.

[177] B. Andersen, H. Ulrich, A. K. Kock, J. H. Wrage, and J. Ingenerf, "Semantic interoperability in the OR.NET project on networking of medical devices and information systems - A requirements analysis," in *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, 2014.

[178] S. Li, *NOVEL INTUITIVE HUMAN-ROBOT INTERACTION USING 3D GAZE*. PhD thesis, Colorado School of Mines, 2017.

[179] A. Kar, S. Member, and P. Corcoran, "A Review and Analysis of Eye-Gaze Estimation Systems , Algorithms and Performance Evaluation Methods in Consumer Platforms," pp. 16495–16519, 2017.

[180] A. J. Larrazabal, C. E. García Cena, and C. E. Martínez, "Video-oculography eye tracking towards clinical applications: A review," 2019.

[181] K. Essig, M. Pomplun, and H. Ritter, "A neural network for 3D gaze recording with binocular eye trackers," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 21, pp. 79–95, 4 2006.

[182] T. Pfeiffer, M. E. Latoschik, and I. Wachsmuth, "Evaluation of Binocular Eye Trackers and Algorithms for 3D Gaze Interaction in Virtual Reality Environments," *Journal of Virtual Reality and Broadcasting*, vol. 5, no. 16, 2009.

[183] T. Pfeiffer, "Measuring and visualizing attention in space with 3D attention volumes," p. 29, 2012.

[184] G. P. Mylonas, A. Darzi, and G.-Z. Yang, "Gaze Contingent Depth Recovery and Motion Stabilisation for Minimally Invasive Robotic Surgery," in *Medical Imaging and Augmented Reality*, no. October, pp. 311–319, 2004.

[185] Z. Wan, X. Wang, L. Yin, and K. Zhou, "A Method of Free-Space Point-of-Regard Estimation Based on 3D Eye Model and Stereo Vision," *Applied Sciences*, vol. 8, no. 10, p. 1769, 2018.

[186] S. Wibirama and K. Hamamoto, "3D gaze tracking system for NVidia 3D Vision¡sup¿®¡/sup¿," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3194–3197, IEEE, 2013.

[187] S. Wibirama, U. G. Mada, and K. Hamamoto, "Error correction in geometric method of 3d gaze measurement using singular value decomposition," in *Proc. of The 5th Indonesia Japan Joint Scientific Symposium*, no. October, pp. 554–558, 2012.

[188] H. Craig, L. Peter, C. Hennessey, and P. Lawrence, "Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 790–799, 2009.

[189] Y.-M. Kwon, K.-W. Jeon, J. Ki, Q. M. Shahab, S. Jo, and S.-K. Kim, "3D Gaze Estimation and Interaction to Stereo Display," Tech. Rep. 3, 2006.

[190] Y. Kwon, J. S. I. C. o. T. f. E, and U. 2006, "Experimental researches on gaze-based 3d interaction to stereo image display," *Springer*.

[191] Y.-m. Kwon, "Gaze Computer Interaction on Stereo Display,"

[192] J. Ki, Y.-M. M. Kwon, and K. Sohn, "3D gaze tracking and analysis for attentive Human Computer Interaction," in *Frontiers in the*

Convergence of Bioscience and Information Technologies, pp. 617–621, IEEE, 2007.

[193] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee, and K. R. Park, "3D gaze tracking method using Purkinje images on eye optical model and pupil," *Optics and Lasers in Engineering*, vol. 50, no. 5, pp. 736–751, 2012.

[194] W. Sewell and O. Komogortsev, "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network," p. 3739, Association for Computing Machinery (ACM), 4 2010.

[195] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *Proceedings - International Conference on Pattern Recognition*, pp. 1167–1172, Institute of Electrical and Electronics Engineers Inc., 12 2014.

[196] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 329–341, 2013.

[197] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3D gaze estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1821–1828, 2014.

[198] F. Lu and X. Chen, "Person-independent eye gaze prediction from eye images using patch-based features," *Neurocomputing*, vol. 182, pp. 10–17, 3 2016.

[199] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised Gaussian process regression," pp. 17–23, Association for Computing Machinery (ACM), 9 2013.

[200] S. M. Munn and J. B. Pelz, "3D point-of-regard, position and head orientation from a portable monocular video-based eye tracker," *Proceedings of the 2008 symposium on Eye tracking research \& applications - ETRA '08*, vol. 1, no. 212, p. 181, 2008.

[201] K. Takemura, K. Takahashi, J. Takamatsu, and T. Ogasawara, "Estimating 3-D Point-of-Regard in a Real Environment Using a Head-Mounted Eye-Tracking System," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 531–536, 2014.

[202] K. Takemura, Y. Kohashi, T. Suenaga, J. Takamatsu, and T. Ogasawara, "Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements," *Proceedings of the 2010 Symposium on Eye-Tracking Research \& Applications - ETRA '10*, vol. 1, no. 212, p. 157, 2010.

[203] H. Wang, J. Pi, T. Qin, S. Shen, and B. E. Shi, "SLAM-based localization of 3D gaze using a mobile eye tracker," pp. 1–5, 2018.

[204] L. Paletta, K. Santner, and G. Fritz, "An Integrated System for 3D Gaze Recovery and Semantic Analysis of Human Attention," *arXiv preprint arXiv:1307.7848*, pp. 6–9, 7 2013.

[205] L. Paletta, "3D Attention : Measurement of Visual Saliency Using Eye Tracking Glasses," pp. 199–204, 2013.

[206] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. Saddik, "Evaluating and improving the depth accuracy of Kinect for Windows v2," *IEEE Sensors Journal*, vol. 15, no. 8, pp. 4275–4285, 2015.

[207] M. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras: Principles, Methods and Applications*. 2013.

[208] NaturalPoint, "OptiTrack™ Motion Capture System." https://optitrack.com/.

[209] NaturalPoint Inc., "Optitrack™ Prime 13." https://optitrack.com/products/prime-13/.

[210] T. Wiedemeyer, "IAI Kinect2." https://github.com/code-iai/iai_kinect2.

[211] G. Bradski, "The OpenCV Library," *Dr Dobbs Journal of Software Tools*, 2000.

[212] K. Harezlak, P. Kasprowski, and M. Stasch, "Towards accurate eye tracker calibration -methods and procedures," *Procedia Computer Science*, vol. 35, no. C, pp. 1073–1081, 2014.

[213] J. J. Moré, "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical analysis*, pp. 105–116, Springer, 1978.

[214] S. Agarwal, K. Mierle, and Others, "Ceres Solver." http://ceres-solver.org.

[215] S. Se, D. Lowe, and J. Little, "Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks," *Robotics Research, The International Journal of*, vol. 21, no. 8, pp. 735–758, 2002.

[216] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[217] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[218] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2011.

[219] "PCL - Statistical Outlier Removal." http://pointclouds.org/documentation/tutorials/statistical_outlier.php.

[220] R. Schnabel and R. Klein, "Octree-based Point-Cloud Compression," *Spbg*, pp. 111–120, 2006.

[221] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[222] J. J. Leonard and H. F. Durrant-Whyte, "Mobile robot localization by tracking geometric beacons," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 376–382, 1991.

[223] B. Preim and C. Botha, "Image-Guided Surgery and Augmented Reality," in *Visual Computing for Medicine*, 2014.

[224] E. Kowler, "Eye movements: The past 25years," *Vision research*, vol. 51, no. 13, pp. 1457–1483, 2011.

[225] T. J. C. O. Vrielink, J. G.-B. Puyal, A. Kogkas, A. Darzi, and G. Mylonas, "Intuitive Gaze-Control of a Robotized Flexible Endoscope," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1776–1782, IEEE, 2018.

[226] M. S. Mould, D. H. Foster, K. Amano, and J. P. Oakley, "A simple nonparametric method for classifying eye fixations," *Vision Research*, vol. 57, pp. 18–25, 2012.

[227] T. Moller and B. Trumbore, "Fast, minimum storage ray-triangle intersection," Tech. Rep. 1425, 1998.

[228] D. Mardanbegi and D. W. Hansen, "Parallax error in the monocular head-mounted eye trackers," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, p. 689, 2012.

[229] A. A. Laviana, S. B. Williams, E. D. King, R. J. Chuang, and J. C. Hu, "Robot assisted radical prostatectomy: the new standard?," *Minerva urologica e nefrologica = The Italian journal of urology and nephrology*, vol. 67, p. 47—53, 3 2015.

[230] M. A. Makary and M. Daniel, "Medical error-the third leading cause of death in the US," *BMJ (Online)*, 2016.

[231] G. E. H. El-Shallaly, B. Mohammed, M. S. Muhtaseb, A. H. Hamouda, and A. H. M. Nassar, "Voice recognition interfaces (VRI) optimize the utilization of theatre staff and time during laparoscopic cholecystectomy," *Minimally Invasive Therapy and Allied Technologies*, 2005.

[232] D. A. Padilla, J. A. B. Adriano, J. R. Balbin, I. G. Matala, J. J. R. Nicolas, and S. R. R. Villadelgado, "Implementation of eye gaze tracking technique on FPGA-based on-screen keyboard system using verilog and MATLAB," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2017.

[233] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[234] L. C. Ebert, G. Hatch, G. Ampanozi, M. J. Thali, and S. Ross, "You Can't Touch This: Touch-free Navigation Through Radiological Images," *Surgical Innovation*, vol. 19, no. 3, pp. 301–307, 2012.

[235] H. Alemzadeh, J. Raman, N. Leveson, Z. Kalbarczyk, and R. K. Iyer, "Adverse events in robotic surgery: A retrospective study of 14 years of fda data," *PLoS ONE*, 2016.

[236] N. Hong, M. Kim, C. Lee, and S. Kim, "Head-mounted interface for intuitive vision control and continuous surgical operation in a surgical robot system," 2018.

[237] M. R. Treat, S. E. Amory, P. E. Downey, and D. A. Taliaferro, "Initial clinical experience with a partly autonomous robotic surgical instrument server," *Surgical Endoscopy and Other Interventional Techniques*, 2006.

[238] J. D. Van Der Laan, A. Heino, and D. De Waard, "A simple procedure for the assessment of acceptance of advanced transport telematics," *Transportation Research Part C: Emerging Technologies*, 1997.

[239] A. J. Rivera-Rodriguez and B. T. Karsh, "Interruptions and distractions in healthcare: Review and reappraisal," 2010.

[240] T. Gillie and D. Broadbent, "What makes interruptions disruptive? A study of length, similarity, and complexity," *Psychological Research*, 1989.

[241] C. G. L. Cao and P. Milgram, "Disorientation in minimal access surgery: A case study," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, pp. 169–172, SAGE Publications Sage CA: Los Angeles, CA, 2000.

[242] H. Imaeda, N. Hosoe, K. Kashiwagi, T. Ohmori, N. Yahagi, T. Kanai, and H. Ogata, "Advanced endoscopic submucosal dissection with traction," *World journal of gastrointestinal endoscopy*, vol. 6, no. 7, p. 286, 2014.

[243] L. L. Swanstrom, R. Kozarek, P. J. Pasricha, S. Gross, D. Birkett, P.-O. Park, V. Saadat, R. Ewers, and P. Swain, "Development of a new access device for transgastric surgery," *Journal of gastrointestinal surgery*, vol. 9, no. 8, pp. 1129–1137, 2005.

[244] G. O. Spaun, B. Zheng, and L. L. Swanström, "A multitasking platform for natural orifice translumenal endoscopic surgery (NOTES): a benchtop comparison of a new device for flexible endoscopic surgery

and a standard dual-channel endoscope," *Surgical endoscopy*, vol. 23, no. 12, p. 2720, 2009.

[245] T. J. C. Oude Vrielink, M. Zhao, A. Darzi, and G. P. Mylonas, "ESD CYCLOPS: A new robotic surgical system for GI surgery," in *Robotics and Automation (ICRA), 2018 IEEE International Conference on*, IEEE, 2018.

[246] L. Zorn, F. Nageotte, P. Zanne, A. Legner, B. Dallemagne, J. Marescaux, and M. de Mathelin, "A Novel Telemanipulated Robotic Assistant for Surgical Endoscopy: Preclinical Application to ESD," *IEEE Transactions on Biomedical Engineering*, 2017.

[247] S. J. Phee, S. C. Low, V. A. Huynh, A. P. Kencana, Z. L. Sun, and K. Yang, "Master and slave transluminal endoscopic robot (MASTER) for natural orifice transluminal endoscopic surgery," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 1192–1195, IEEE, 2009.

[248] M. Remacle, V. M. N. Prasad, G. Lawson, L. Plisson, V. Bachy, and S. der Vorst, "Transoral robotic surgery (TORS) with the Medrobotics Flex™ System: first surgical application on humans," *European Archives of Oto-Rhino-Laryngology*, vol. 272, no. 6, pp. 1451–1455, 2015.

[249] C. C. Nduka, P. A. Super, J. R. Monson, and A. W. Darzi, "Cause and prevention of electrosurgical injuries in laparoscopy.," *Journal of the American College of Surgeons*, vol. 179, no. 2, pp. 161–170, 1994.

[250] B. Zheng, E. Rieder, M. A. Cassera, D. V. Martinec, G. Lee, O. N. M. Panton, A. Park, and L. L. Swanström, "Quantifying mental workloads of surgeons performing natural orifice transluminal endoscopic

surgery (NOTES) procedures," *Surgical endoscopy*, vol. 26, no. 5, pp. 1352–1358, 2012.

[251] K. Kume, T. Kuroki, T. Sugihara, and M. Shinngai, "Development of a novel endoscopic manipulation system: The Endoscopic operation robot," *World journal of gastrointestinal endoscopy*, vol. 3, no. 7, p. 145, 2011.

[252] H. J. M. Pullens, N. Van Der Stap, E. D. Rozeboom, M. P. Schwartz, F. van der Heijden, M. G. H. Van Oijen, P. D. Siersema, and I. A. M. J. Broeders, "Colonoscopy with robotic steering and automated lumen centralization: a feasibility study in a colon model," *Endoscopy*, vol. 48, no. 03, pp. 286–290, 2016.

[253] E. Rozeboom, J. Ruiter, M. Franken, and I. Broeders, "Intuitive user interfaces increase efficiency in endoscope tip control," *Surgical endoscopy*, vol. 28, no. 9, pp. 2600–2605, 2014.

[254] V. K. Dik, I. T. C. Hooge, M. G. H. van Oijen, and P. D. Siersema, "Measuring gaze patterns during colonoscopy: a useful tool to evaluate colon inspection?," *European journal of gastroenterology & hepatology*, vol. 28, no. 12, pp. 1400–1406, 2016.

[255] G. P. Mylonas, K.-W. W. Kwok, D. R. C. James, D. Leff, F. Orihuela-Espina, A. Darzi, and G.-Z. Z. Yang, "Gaze-Contingent Motor Channelling, haptic constraints and associated cognitive demand for robotic MIS," *Medical image analysis*, vol. 16, no. 3, pp. 612–631, 2012.

[256] R. Reilink, G. de Bruin, M. Franken, M. A. Mariani, S. Misra, and S. Stramigioli, "Endoscopic camera control by head movements for thoracic surgery," in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, pp. 510–515, IEEE, 2010.

[257] R. N. Barker and S. G. Brauer, "Upper limb recovery after stroke: the stroke survivors' perspective," *Disability and rehabilitation*, vol. 27, no. 20, pp. 1213–1223, 2005.

[258] "Kinova Robotics." https://www.kinovarobotics.com/.

[259] "Exact Dynamics." https://www.exactdynamics.nl/.

[260] C. Grau, R. Ginhoux, A. Riera, T. L. Nguyen, H. Chauvat, M. Berg, J. L. Amengual, A. Pascual-Leone, and G. Ruffini, "Conscious brain-to-brain communication in humans using non-invasive technologies," *PLoS One*, vol. 9, no. 8, p. e105225, 2014.

[261] K. Tanaka, K. Matsunaga, and H. O. Wang, "Electroencephalogram-based control of an electric wheelchair," *IEEE transactions on robotics*, vol. 21, no. 4, pp. 762–766, 2005.

[262] H. A. Shedeed, M. F. Issa, and S. M. El-Sayed, "Brain EEG signal processing for controlling a robotic arm," in *8th Int'l Conf. Computer Engineering Systems (ICCES)*, pp. 152–157, 2013.

[263] E. Albilali, H. Aboalsamh, and A. Al-Wabil, "Comparing brain-computer interaction and eye tracking as input modalities: An exploratory study," in *Current Trends in Information Technology (CTIT), 2013 International Conference on*, pp. 232–236, IEEE, 2013.

[264] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[265] E. A. Curran and M. J. Stokes, "Learning to control brain activity: a review of the production and control of EEG components for driving brain–computer interface (BCI) systems," *Brain and cognition*, vol. 51, no. 3, pp. 326–336, 2003.

[266] E. B. Kelly, "Encyclopedia of human genetics and disease," 2013.

[267] "Eyedrivomatic." https://www.eyedrivomatic.org/.

[268] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 858–865, IEEE, 2011.

[269] R. c. Willow Garage, "ORK: (O)bject (R)ecognition (K)itchen." https://github.com/wg-perception/object_recognition_core.

[270] I. A. Sucan and S. Chitta, "Moveit!," *Online at http://moveit. ros. org*, 2013.

[271] C.-S. Chung, H. Wang, and R. A. Cooper, "Functional assessment and performance evaluation for assistive robotic manipulators: Literature review," *The journal of spinal cord medicine*, vol. 36, no. 4, pp. 273–289, 2013.

[272] A. J. Knulst, R. Mooijweer, F. W. Jansen, L. P. Stassen, and J. Dankelman, "Indicating shortcomings in surgical lighting systems," *Minimally Invasive Therapy and Allied Technologies*, 2011.

[273] J. Y. Zhang, S. L. Liu, Q. M. Feng, J. Q. Gao, and Q. Zhang, "Correlative Evaluation of Mental and Physical Workload of Laparoscopic Surgeons Based on Surface Electromyography and Eye-tracking Signals," *Scientific Reports*, 2017.

[274] J. Miseikis, K. Glette, O. J. Elle, and J. Torresen, "Automatic Calibration of a Robot Manipulator and Multi 3D Camera System." 2016.