A 3D camera-based system concept for safe and intuitive use of a surgical robot system

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Philip Matthias Nicolai

aus Heidelberg

Tag der mündlichen Prüfung:	9. Juni 2016
Erster Gutachter:	Prof. Dr.-Ing. Dr. h.c. Heinz Wörn
Zweiter Gutachter:	Prof. Paolo Fiorini, PhD

Credits

This thesis would not have been possible without the support of numerous people. First, I'd like to thank my supervisor, Prof. Dr.-Ing. Dr. h.c. Heinz Wörn, for providing me with the opportunity to work in the exciting and fast-moving research field of surgical robot systems and 3D camera systems. My sincere thanks also goes to Prof. Paolo Fiorini, PhD, who led the European research project *Patient Safety in Robotic Surgery (SAFROS)*, during which I had the chance to develop many concepts of this thesis, and who graciously agreed to be the second reviewer of my thesis. A further big "Thank you" goes to Dr. Jörg Raczkowsky, leader of the medical research group MeGI at IAR-IPR, for his support, discussions and for his trust that allowed me to freely explore research topics.

In parallel with this thesis, the OP:Sense system has been created as a collaboration with multiple colleagues from both medical research groups at IAR-IPR. OP:Sense aims to integrate the results from different projects and theses, and I could not have finished this thesis in its final scope without relying on many parts of OP:Sense contributed by my colleagues. I'd like to thank my (ex-)colleagues for the great working atmosphere and collaboration both on scientific endeavours and otherwise: Tim Beyl, Thorsten Brennecke, Julien Mintenbeck, Holger Mönnich, Luzie Schreiter, Yaokun Zhang, Andreas Bihlmaier and Jessica Hutzl. Special thanks is reserved for Mirko Kunze for his general humour throughout the day and for meticulously proof-reading my thesis draft even in the final stages of his own thesis.

Most important, I'd like to thank Simone Rupp for supporting me throughout the course of this thesis, even if at times we only saw each other awake when I came home from work at the same time as she had to get up early.

The final thanks is reserved for my mother Helga Nicolai. She selflessly supported me throughout my whole life and worked hard to open up all possibilities to me, but always respected my decisions and never pressured me to follow any specific path in life she might or might not have wished for.
This thesis is dedicated to her.

Abstract

Within the last decades, surgical robot systems have been integrated operation rooms worldwide. However, in current robotic procedures, the surgical personnel has to devote a significant part of attention in order to ensure and monitor seamless functioning of the robot system.

To overcome this limitation, this thesis explores the feasibility of developing a system for safe and intuitive use of surgical robots, based on state-of-the-art range imaging cameras and newly developed algorithms. A novel concept for an Operating Room (OR) monitoring system is proposed that can perceive the environment of a surgical robot using multiple 3D cameras and detect potentially harmful situations between the robot, its surroundings and the persons in its vicinity, i.e. the OR personnel and the patient. Such a system is realized in a generic way in order to be applicable to different surgical robot systems. It is optimized for a low spatial footprint for not interfering with the OR personnel and their actions in already crowded ORs. Furthermore, the system provides intuitive feedback to the OR personnel whenever safety-critical events are detected, without drawing on their attention otherwise.

The realized system was extensively evaluated using the OP:Sense surgical research platform. Based on the proposed approach of establishing a virtual safety zone around each robot arm, the system was shown to reliably detect and therefore avoid impending collisions, without requiring information about the trajectory of the robot. To ensure the applicability of use within the operating room, the effects of sterile draping on range imaging cameras were analyzed. A filtering method was put forward to eliminate these effects within the realized ToF camera system, allowing for successful detection of impending collisions even for draped robots.

The results indicate that a 3D-camera-based supervision system can effectively contribute to the safety of use of surgical robot systems in the OR, allowing the OR personnel to completely focus on their medical tasks. The proposed methods contribute to scene supervision for human-robot cooperation and show the feasibility of the approach.

Zusammenfassung

Motivation

Seit die ersten kommerziell erhältlichen Chirurgierobotersysteme Mitte der Neunziger Jahre klinisch eingesetzt wurden, haben sich sowohl der Einsatz derartiger Systeme als auch die Vielfalt der abgedeckten Operationen vervielfacht. Im Jahr 2014 wurden alleine mit dem am Markt führenden da Vinci System über $500\,000$ Operationen durchgeführt.

Insbesondere im Bereich der minimal-invasiven Chirurgie hat der Einsatz von Chirurgierobotern jedoch neben den medizinischen Aspekten noch weitere Konsequenzen: Der Chirurg führt die Operation von außerhalb des sterilen Bereichs per Telemanipulation des Robotersystems durch, was direkte Auswirkungen auf die Arbeitsabläufe und Rollenverteilung im Operationssaal hat. Da aktuelle und in der Entwicklung befindliche chirurgische Robotersysteme größtenteils keine Sensorik enthalten, mit der sie ihre Umgebung erfassen könnten, muss das OP-Personal sowohl die korrekte Funktionsweise als auch die Kollisionsfreiheit der Robotersysteme fortlaufend überwachen.

Ziel dieser Arbeit ist daher die Entwicklung und Umsetzung eines 3D-Kamera-basierten Überwachungssystems für den Operationssaal, das die sichere Anwendung chirurgischer Robotersysteme überwacht und mittels räumlicher erweiterter Realität (AR) Informationen z.B. zum Roboterzustand intuitiv für das OP-Personal erfassbar macht. Hierdurch soll das OP-Personal unterstützt und entlastet werden, um ihm die uneingeschränkte Konzentration auf die medizinischen Aspekte der Operation zu ermöglichen.

Methoden

Im Mittelpunkt dieser Arbeit steht die Konzeption und Realisierung eines Überwachungssystems bestehend aus mehreren, räumlich verteilten 3D Kameras, das die unmittelbare Umgebung der Operationsliege in Echtzeit als Punktwolke erfasst und in Bezug auf verschiedene Sicherheitsaspekte analysiert. Um jeden Roboterarm wird eine virtuelle, dynamische Sicherheitshülle gebildet, mittels derer die erfasste Szene segmentiert wird. Bei Verletzungen der virtuellen Sicherheitshülle wird zunächst eine räumliche Analyse durchgeführt, um die Ursache der Verletzung zu ermitteln. Falls diese im Eindringen eines externen Objektes oder eines

Menschen besteht, werden je nach Zustand des Robotersystems verschiedene Reaktionen ausgeführt, um beispielsweise Kollisionen zwischen Roboter und Mensch zu vermeiden.

Grundlage für dieses Konzept ist der entwickelte Shape Cropping Algorithmus, mittels dessen die sicherheitshüllenbasierte Segmentierung der erfassten Szene umgesetzt wird. Weiterhin ermöglicht dieser Algorithmus die Detektion des Robotersystems in der Szene, um den korrekten Aufbau des Robotersystems in Übereinstimmung mit einer präoperativen Planung zu überprüfen.

Das entwickelte Kamerasystem gliedert sich in zwei separate Teilsysteme, deren Kameras die Szene anhand unterschiedlicher Technologien dreidimensional erfassen. Das erste Teilsystem besteht aus sieben Time-of-Flight (ToF) Kameras aus dem industriellen Bereich, das zweite Teilsystem aus vier Kinect v1 Kameras.

Aufgrund der unterschiedlichen Vor- und Nachteile der verwendeten Kamerasysteme bezüglich Auflösung, Latenz und externer Kontrolle der Kameraparameter werden die erfassten Informationen in ein zweistufiges Szenenmodell integriert. Die erste Ebene des Szenenmodells beinhaltet niedrig aufgelöste Punktwolken, die von den Time-of-Flight-Kameras mit einer geringen Latenz erfasst werden und über keine semantischen Informationen verfügen. Diese Daten bilden die Grundlage für sicherheitsrelevante Funktionen wie die Erkennung potentieller Kollisionen. Die zweite Ebene beinhaltet höher aufgelöste räumliche Daten inklusive semantischer Informationen, insbesondere der Erkennung von Menschen in der Szene. Diese Daten sind jedoch erst mit einer größeren Latenz verfügbar.

Zur Überbrückung dieser semantischen Lücke wurde ein Algorithmus entwickelt, der die Vorwärtspropagation von semantischen Annotationen ermöglicht. Dieser erlaubt es, die auf der zweiten Ebene des Szenenmodells vorhandenen semantischen Annotationen fortlaufend vorwärts zu berechnen, so dass sie auf der ersten Ebene des Szenemmodells direkt zur Verfügung stehen. Somit kann beispielsweise bei der Detektion einer bevorstehenden Kollision des Roboters anhand der vorwärts-berechneten semantischen Information erkannt werden, ob der Roboter mit einem Menschen oder der Umgebung kollidieren würde, so dass unterschiedliche Reaktionen ausgelöst werden können.

Um Informationen zum Zustand des Robotersystems oder andere Hinweise intuitiv für das OP-Team verfügbar zu machen, wurde ein entsprechendes Konzept entworfen und umgesetzt, das auf projektorbasierter räumlicher erweiterter Realität basiert. Dies ermöglicht (i) die Visualisierung des Zustandes des Chirurgieroboters durch eine farblich kodierte, passgenaue Projektion auf den realen Roboterarm, (ii) die Lenkung der Aufmerksamkeit des OP-Teams im Falle sicherheitskritischer Situationen sowie (iii) im Falle von minimal-invasiven Eingriffen die Augmentierung des Patienten mit den Posen der laparoskopischen Instrumente.

Ergebnisse

Das vorgestellte Konzept für ein Überwachungssystem wurde im Rahmen dieser Arbeit vollständig umgesetzt und die Effektivität des Systems konnte in verschiedenen Versuchen anhand der Forschungsplattform OP:Sense nachgewiesen werden. Die realisierten Kamerasysteme gewährleisten eine redundante Abdeckung des Arbeitsraums rund um die Operationsliege, so dass die virtuelle Sicherheitshülle um die Roboter jederzeit von mehreren Kameras überwacht wird. Ein speziell für den Einsatz im OP entworfener, projektionsbasierter Algorithmus zur Registrierung der verschiedenen 3D-Kamerasysteme wurde erfolgreich evaluiert.

Basierend auf der erfolgten Registrierung der Kamerasysteme konnte gezeigt werden, dass das Konzept der virtuellen Sicherheitshülle um den Roboter die Erkennung potentieller Kollisionen ermöglicht, so dass diese sicher vermieden werden können. Um die Nutzung dieses Konzepts im Operationssaal zu ermöglichen, wurden die Auswirkungen der sterilen Schutzhülle, die Chirurgierobotern im Operationssaal übergestülpt wird, auf die verschiedenen 3D-Kamerasysteme analysiert. Eine Methode zur Entfernung entsprechender Artefakte in den ToF-Kamerabildern wurde entwickelt und evaluiert.

Experimente zur Bewertung des Algorithmus zur Vorwärtspropagation semantischer Annotationen zeigten, dass der entwickelte Algorithmus eine zuverlässige und präzise Methode darstellt, um die in der zweiten Ebene des Szenenmodells enthaltenen semantischen Informationen trotz Latenz auf die erste Ebene des Szenenmodells zu übertragen und somit bei drohenden Kollisionen des Roboters zwischen Menschen und Umgebung unterscheiden zu können.

Diskussion, Ausblick und Fazit

Das in der Arbeit vorgestellte Konzept für ein 3D-kamerabasiertes Überwachungssystem des Operationssaals wurde erfolgreich realisiert und anhand der Forschungsplattform OP:Sense evaluiert. Die Ergebnisse zeigen, dass das Systemkonzept sich zur sicheren und redundanten Überwachung chirurgischer Robotersysteme eignet und somit das Potential hat, das OP-Personal bei roboter-assistierten Operationen zu unterstützen und zu entlasten.

Aufbauend auf dem vorgestellten Überwachungssystem bieten sich künftige Arbeiten im Bereich der räumlichen und semantischen Integration in den Operationssaal, der Verbindung des Überwachungssystems mit einer Wissensbasis und der Transfer des Konzepts in ein industrielles Umfeld an.

Zusammengefasst leistet die Arbeit wissenschaftliche Beiträge in den Bereichen der redundanten 3D-Überwachung und Einsatz von 3D-Kameras im Operationssaal sowie sicherer Mensch-Roboter-Kooperation.

Contents

1. Introduction

In the early 1970s, the *National Aeronautics and Space Agency* (NASA) proposed the first concept for telerobotic healthcare. In the mid-1980s, first surgical robot systems were developed. In the mid-1990s, the first commercially available surgical robot systems were clinically used.

Since then, the field of surgical robot systems has come a long way: Surgical robot systems are nowadays used in a wide field of interventions, ranging from minimally invasive procedures to orthopedics, pediatrics, neurosurgery and radiosurgery. Medical robots in general span an even wider field that encompasses surgery, but also rehabilitation and imaging.

In 2000, the da Vinci™ surgical robot system was the first such system to receive clearance from the Food and Drug Administration (FDA) for general Minimally Invasive Robotic Surgery (MIRS) and since then it has clearly dominated the market. In 2015, *Intuitive Surgical* reported a total of almost $3\,500$ installations of the da Vinci™, used in over half a million interventions in the year before [72]. However, there are several drawbacks to the system, such as the sheer size and volume of the monolithic robot, its lack of force sensing capabilities and it's closed nature.

Currently, multiple new systems for MIRS are in advanced research stages or close to commercialization that aim to overcome these limitations [60]. The trend of development and research clearly points to smaller, more modular systems that allow for a higher flexibility in the OR, such as the *MiroSurge* system developed by German Aerospace Center (DLR), Germany. A further trend is the application of preoperative planning in order to find the optimal configuration of a surgical robot system, depending on the specific patient anatomy and clinical indication.

Further, there is a clear trend towards automation and integration of the OR. Currently, many vendors only market closed solutions for integrated operating rooms and do not allow for interoperability with other systems. Apart from the commercial interests of each medical company, this is also due to a lack of standards for exchange of information between medical devices in the OR. Creating a technical basis for vendor-independent safe integration and networking of medical devices is therefore an active field of research. One example is the OR.NET project that brings together vendors, both of integrated ORs and medical equipment, with research institutes and clinics to develop open standards for interconnection of medical devices [97].

1.1. Motivation

The design and development of surgical robot systems is naturally focused on the medical aspects of the respective system and domain, e.g. the specific requirements of a certain type of intervention. Therefore, most surgical robot systems do not include exteroceptive sensors that would allow them to perceive and react to their environment. It is up to the OR personnel to supervise the motions of the robot throughout an intervention and intervene in case of hazardous situations. However, during MIRS the surgeon is spatially removed from the situs and cannot notice such situations. This can lead to hazardous situations, which are illustrated by reports of OR theatre teams that had "to quickly tell the surgeon to stop because the robot arms were going to hit the patient" [151].

As more and more surgical robot systems enter the market or are close to commercialization, it would be beneficial to implement a generic sensor system for the OR that allows to supervise the safe use of different surgical robot systems. This way, an additional, redundant safety layer could be provided that operates independently of the respective surgical robot system. Further, such a system contributes to allowing the surgical team to completely focus on their medical tasks without having to divide attention to the correct working of the surgical robot system.

Based on the ongoing research on interoperability between systems in the OR, which is expected to come to fruition in the near future, and the development of new surgical robot systems that allow more flexibility than the current systems, now is the right time for thinking ahead and asking the question: If the robot can be integrated with the room, can the room monitor safety of the robot?

1.2. Aim of this thesis and research questions

To answer this question, this thesis aims to explore the feasibility of developing a supervision system, based on state-of-the-art range imaging cameras and newly developed algorithms. Therefore, a novel concept for an OR monitoring system is proposed that can perceive the environment of a surgical robot using multiple 3D cameras and detect potentially harmful interactions between the robot and the humans, i.e. the OR personnel and the patient. Such a system (i) needs to be realized in a generic way in order to be applicable to different surgical robot systems, (ii) has to be optimized for a low spatial footprint for not interfering with OR personnel and its actions in already crowded ORs and (iii) must provide intuitive feedback to the OR personnel whenever safety critical events are detected, without drawing on their attention otherwise.

Multiple open research questions are addressed that have not been investigated before:

- How can a system be realized that adds a layer of safety to different surgical robots without requiring hardware modifications?
- How does sterile draping, in which robots are covered during interventions, affect scene acquisition by current range imaging cameras?
- Can a 3D camera based supervision system detect the positions of surgical robots in an intraoperative setting and verify the correct setup of the robot system?
- Is it possible to reliably monitor the performance of the surgical robot and detect potential collisions with its surroundings before they occur?
- How can feedback about the safety state of the surgical robot system be provided to the OR personnel in an intuitive and distraction-free manner?

1.3. Outline

This thesis is structured as follows:

State of the art A brief introduction is given on relevant norms and definitions, followed by an overview of current and upcoming surgical robot systems. Different principles of range imaging are discussed and supplemented by an outlook of current and upcoming 3D cameras. Previous research on the application of 3D cameras and augmented reality in the OR is discussed as well as side effects and non-clinical performance characteristics of surgical robot systems. An overview of different approaches to safe human-robot collaboration is given with a focus on external optical sensing. Finally, open research questions are discussed and related to the work performed in this thesis.

System concept Technical and clinical prerequisites of the modular supervision system for OR monitoring are discussed and the realized camera system as well as corresponding algorithms are introduced. The safety concept of establishing a safe zone around the robot is detailed with the according Shape Cropping algorithm and the proposed safety features. Clinical motivation for establishing a Spatial Augmented Reality (SAR)-based feedback system are discussed and the augmentation concept is presented.

Realization The hardware and software components used in this thesis are detailed as well as the realized distributed system architecture. An in-depth description of the supervision system is given, discussing the camera placement and the implementation of the different camera systems. Two developed algorithms closely related to the supervision system are described,

namely the projector-based registration method and the algorithm for forward propagation of semantic labelling. For the safety system, the Shape Cropping algorithm and its application to robot localization, detection of impending collisions and continuous pose supervision are explained. The chapter closes with a description of the concrete setup of the SAR system for feedback to the OR personnel and the realized feedback features.

Results Evaluation and results of the realized camera systems are presented in terms of interference, performance, registration accuracy and analysis of the achieved coverage. The use cases and results obtained for forward propagation of semantic labelling are discussed, followed by results of the safety concept including performance of the Shape Cropping algorithm itself and its applications to robot localization, collision avoidance and continuous pose supervision. Last, a description and the results of a user study on the implemented SAR features are given.

Conclusions The results obtained in the previous chapter are discussed in relation to the research questions brought forward in the introduction. The contributions of this thesis are listed and directions for potential future research are indicated.

2. State of the Art

This chapter provides an overview of background knowledge and research performed on the topics covered in this thesis. It starts by highlighting the relevant norms and definitions for general and surgical robotics as well as human-robot collaboration in section 2.1, followed by short introductions of the main concepts of intuitive use and Augmented Reality (AR).

Section 2.2 gives a brief overview of technologies relevant for this thesis, starting with current and upcoming surgical robot systems for MIRS. The main sensing principles of range imaging cameras are discussed and an extensive overview of current and upcoming range imaging cameras is presented.

Section 2.3 details existing research on applications of the previously introduced technology and concepts. Works in the surgical domain are presented first, followed by works on safe human-robot interaction in a general context.

Based on the discussed fundamentals and research works, open research questions are identified in section 2.4. Finally, the research hypotheses are presented that have been explored in this thesis.

2.1. Norms and definitions

In the following, a short overview of relevant norms and definitions for robot systems and their safe application in human-robot collaboration is given. The term *intuitive use* is then introduced and a short taxonomy of augmented reality is presented.

2.1.1. Norming bodies and entities

Worldwide standards are developed and published by three international organizations who, as an alliance, form the *World Standards Cooperation (WSC)*:

- The *International Organization for Standardization (ISO)* develops "market-relevant International Standards that support innovation and provide solutions to global challenges" [73]. ISO standards cover a wide range of topics from quality management to social responsibility and occupational health and safety.

- The *International Electrotechnical Commission (IEC)* provides "International Standards for all electrical, electronic and related technologies" [66].

- The *International Telecommunication Union (ITU)* is responsible for "the technical standards that ensure networks and technologies seamlessly interconnect" [80]. It operates with a global perspective, e.g. by allocating global radio spectrum and satellite orbits.

ISO and IEC have published and/or are working on standards that concern robotic surgery or safety in human-robot collaboration, which will be referenced in either their international or German version (as published by *Deutsches Institut für Normung (DIN)* and *Verband der Elektrotechnik (VDE)*). Due to its nature, the ITU has not defined any standards that concern safety of robotic surgery. This might change if telesurgery gains widespread adoption and the question of prioritized data transmissions for such procedures becomes more prompting.

2.1.2. Robotics

To distinguish between different kinds of actuated systems, ISO 8373:2012 "Robots and robotic devices – Vocabulary" [77] gives the following definitions:

- *robot*: "actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks". It is further noted that "a robot includes the control system and interface of the control system".

- *robotic device*: "actuated mechanism fulfilling the characteristics of an industrial robot or a service robot, but lacking either the number of programmable axes or the degree of autonomy", for example a "teleoperated device".

- *robot system*: "system comprising robot(s), end effector(s) and any machinery, equipment, devices, or sensors supporting the robot performing its task".

In line with these definitions, the FDA considers Robot Assisted Surgical Devices (RASDs) as "technically not robots, since they are guided by direct user control" [83]. In fact, they were first cleared by the FDA as being equivalent to laparoscope holding devices.

Concerning medical applications, the Robotic Consensus Group of the *Society of American Gastrointestinal and Endoscopic Surgeons and Minimally Invasive Robotic Association* defines *robotic surgery* as "a surgical procedure or technology that adds a computer technology-enhanced device to the interaction between surgeon and the patient during a surgical operation and assumes some degree of freedom of control heretofore completely reserved for the surgeon." [59].

2.1.3. Safety in industrial robotics

The technical committee 184/SC 2 of the ISO has published several standards that deal with safety requirements for the usage of industrial robots. In the scope of this thesis, the most notable such standard is ISO 10218:2011 "Robots and robotic devices – Safety requirements for industrial robots". It consists of two parts that detail basic safety requirements for the robot itself [75] and for the full robotic system and its integration [76].

Concerning human-robot collaboration, ISO 10218 puts forward the following definitions:

- *Collaborative robot:* "robot designed for direct interaction with a human within a defined collaborative workspace" [75, 3.3]
- *Collaborative workspace:* "space in which purposely designed robots work in direct cooperation with a human within a defined workspace" [76, 3.4]
- *Collaborative operation:* "workspace within the safeguarded space where the robot and a human can perform tasks simultaneously during production operation" [75, 3.5]

Furthermore, ISO 10218-1 distinguishes four different methods that can be applied to collaborative operations, as given in Table 2.1. These types of collaborative operations will be further detailed by the upcoming technical specification ISO/TS 15066 "Robots and robotic devices – Safety requirements for industrial robots – Collaborative operation". Following the ISO standardization process, this specification will be valid for three years after publication and might then be integrated into ISO 10218-2 if it is deemed suitable for a standard.

For classification of safety risks in collaborative operations, two kinds of potential contact between human and robot are distinguished in both ISO 10218 and ISO 15066:

- *Quasi-static contact*: Part of the body of the operator is clamped between a moving part of the robot system and some part of the environment.
- *Transient or dynamic contact*: The operator is not clamped by the contact and can retract or recoil freely.

In both ISO 10218-1 and ISO 15066, a maximal allowed velocity for the robot's end-effector of 250 $\mathrm{mm/s}$ is specified for human-robot interaction to limit the energy that can be transferred to the human in case of either kind of contact. In order to allow for a more accurate estimation of potential injuries caused by adverse contact events, biomechanical limits of the human body are researched in different projects and will be included in ISO 15066 [61, 117].

Clause	Type	Main means of risk reduction
5.10.2	Safety-rated monitored stop	No robot motion when operator is in collaborative space
5.10.3	Hand guiding	Robot motion only through direct input of operator
5.10.4	Speed and separation monitoring	Robot motion only when separation distance above minimum separation distance
5.10.5	Power and force limited by inherent design or control	In contact events, robot can only impart limited static and dynamic force

Table 2.1.: Methods for collaborative operation and according means of risk reduction according to ISO 10218-1 [75].

2.1.4. Safety in medical robotics

Contrary to the domain of industrial robotics and the domain of personal care robots, for which a safety standard was published in 2014 [78], there exists no such standard for medical robots yet. In 2011, the *Joint Working Group (JWG) 9* was formed between ISO and the IEC. JWG 9 is working on a technical report that will define medical robots as *Medical Electric Equipment* with a *Degree of Autonomy (DoA)*. This technical report will offer guidance on methodologies for risk assessment and basic safety and essential performance for such systems.

In addition to defining a general safety standard for medical robotics, JWG 9 proposed to define particular standards for surgical robots and rehabilitation robots [68]. Since April 2015, the new JWG 35 is working on a standard for medical robots in surgery. This standard is planned to be completed in November 2018 as IEC 80601-2-77 [79].

2.1.5. Intuitive use

Naumann et al. regard *intuitive use* as a characteristic of human-machine-systems. They postulate that intuitive use cannot be applied to a technical system per se but has to be tied to a specific context, e.g. to achieve a certain objective with a technical system. This results in the following definition: *"A technical system is, in the context of a certain task, intuitively usable while the particular user is able to interact effectively, not-consciously using previous knowledge"* [130].

This definition depends on the term *effectiveness*, which is defined by ISO as *"accuracy and completeness with which users achieve specified goals"* [74]. Therefore effectiveness can be used as a metric to rate or assess the procedure of intuitive interaction according to Naumann et al.

In [62], Hurtienne mentions the notion of intuitive use *"as a sub-concept of usability with a strong focus on the mental demands in using technology"* and subsequently

emphasizes the importance of the mental load for the term intuitive use. This leads to a slightly altered definition compared to Naumann et al. : *"[...] intuitive use is defined as the extent to which a product can be used by subconsciously applying prior knowledge, resulting in an effective and satisfying interaction using a minimum of cognitive resources"*.

2.1.6. Augmented reality

Milgram et al. were the first to define a taxonomy of *Mixed Reality (MR)* visual displays. They introduce the idea of a virtuality continuum that spans between two extrema: the *real environment* and the *virtual environment* (see Figure 2.1). A MR environment is defined as *"one in which real world and virtual world objects are presented together within a single display"* [121].

Carmigniani et al. build upon the work of Milgram and define *AR* as *"a real-time direct or indirect view of a physical real-world environment that has been enhanced/augmented by adding virtual computer-generated information to it"* [19]. Contrary to Virtual Reality (VR), which puts the user into a fully synthetic world without a glimpse of their real surroundings, AR aims to enhance the real environment of the user by adding augmentations such as virtual objects or cues. These augmentations can refer to different senses besides vision, like smell, touch and hearing.

According to [19], three main modalities for AR-displays can be distinguished:

1. *Head-Mounted Displays (HMDs)* are attached to the users head, for example in the form of glasses or a helmet, and overlay virtual images over their perception of the real world. Two main concepts for HMDs are

    a) *video-see-through*: The user perceives the world via displays in front if their eyes that show a live (stereoscopic) video stream, taken from their point of view. The augmentations are directly superimposed onto the video stream. This allows for optimal synchronization between the augmentations and the perception of the world, but has long suffered from low resolution of the displays and the technical effort in terms of necessary cameras and computation.

    b) *optical-see-through*: Virtual images are superimposed over the users own view of the physical world, e.g. by using half-silver mirror technology. This allows for an optimal perception of the real world, but can quickly lead to jitter if the augmentations lag behind the real world.

2. *Handheld Displays* also follow the video-see-through concept as they capture the environment via built-in camera and add augmentations to the live video stream that is shown on their display. Depending on the application, they often employ additional sensors such as GPS, compass and gyroscopes. Prime examples of HMDs are smartphones and tablet PCs.

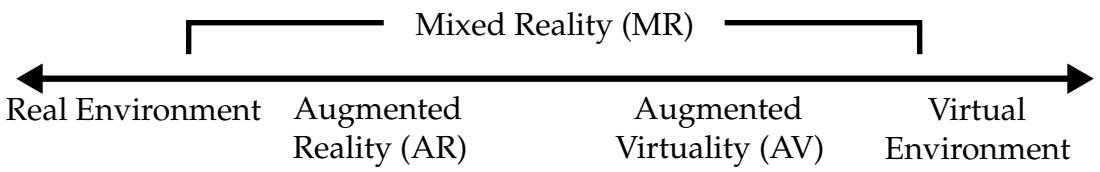


Figure 2.1.: Virtuality Continuum accoding to Milgram: Augmented Reality (AR) and Augmented Virtuality are both subsets of Mixed Reality (MR), in between the real and virtual world [121].

3. *SAR* means that graphical information is directly displayed onto physical objects without the need for users to wear or carry an AR-device. This requires to integrate the according technology, such as projection systems, into the environment. The benefit of this modality is that SAR is visible for multiple users at the same time as they do not need a special device each. Similar to HMDs, SAR can be further separated into different approaches:

   a) *video-see-through*: Screen-based SAR-systems that are similar in concept to the corresponding HMD approach. A drawback of this method is that it is completely stationary.

   b) *optical-see-through*: SAR-systems that employ spatial optical combiners, like mirror beam splitters or transparent screens to superimpose augmentations. Like video-see-through systems, these are stationary and offer one major direction of view; in addition, their position has to be carefully calibrated in order to provide accurate overlays.

   c) *direct augmentation*: Projector-based spatial displays that project the augmentation directly onto the surface of an object. These systems offer a seamless augmentation and allow for multiple simultaneous users from different angles of view. A disadvantage is that projections are by their nature prone to occlusions, i.e. shadows being caused by persons or objects moving between the projector and the augmented surface.

In addition to these AR modalities, the MR modality *Screen-based Mixed Reality* denotes virtual augmentations onto live video streams, presented on a fixed monitor without necessary registration to the environment.

Portable modalities for AR such as HMDs and handheld displays can also benefit from external information about their location, in addition to their integrated sensors. This information can be provided in an inside-out way, for example by fitting a room with easily trackable AR-markers at known locations, based on which the device can infer its position. In contrast, the outside-in approach is to actively track the device, e.g. using an external marker-based tracking system that directly provides the location of the device.

2.2. Technologies

2.2.1. Surgical robot systems

2.2.1.1. Main concepts

In [60], Hoeckelmann et al. give an extensive overview of surgical robot systems and their capabilities. They distinguish between two fundamental concepts in robot-assisted surgery:

1. *Teleoperated systems*: The surgical robot is controlled by the surgeon from an operating console which is usually located in the OR.
2. *Image-guided systems*: The surgical robot executes a preoperatively defined surgical plan, based on intraoperative geometric information acquired by a navigation system or other tracking system.

An additional control modality is *hands-on guidance*: The surgeon manually guides the robot by hand contact to move the surgical instrument attached to the robot into the desired position. Hands-on guidance can be used with both concepts, e.g. for initial positioning of the robotic instruments before the procedure starts.

2.2.1.2. da Vinci™

The most prominent example of a teleoperated surgical robot is the *da Vinci™* system by *Intuitive Surgical* [48]. With 3 477 installations worldwide and over 570 000 procedures in 2015 (as reported by Intuitive Surgical [72]), it clearly dominates the market for general MIRS.

Since the first version of the da Vinci™, which was cleared by FDA in 2000, there have been several iterations of the system. While each generation introduced changes in the hardware and software capabilities, the basic concept of the robot system has remained unchanged since its first iteration. As a teleoperated system, it features a surgeon console, from which the robot is controlled, and a movable patient cart with three or four robotic arms (see Figure 2.2), to which the instruments and camera are attached. Before the start of an intervention, a so called *docking procedure* is performed during which the patient cart is positioned close to the OR table and the arms are manually positioned so that their mechanical pivot points correspond to the trocar positions.

The da Vinci™system is a strictly teleoperated system that does not include any sensors that allow to collect information about its environment. Especially, it does not offer force-torque sensors in the robotic arms or any other modalities that can perceive the proximity of surgical personnel.

As the da Vinci™ has been the de facto standard for MIRS for over 15 years, most studies on the effects of MIRS evaluate interventions performed with the da

Figure 2.2.: Patient cart of the da Vinci™ with three arms for endoscopic instruments and one arm for the endoscopic camera. *Left:* da Vinci® Si™ (2009) [70], *right:* da Vinci® Xi™ (2014) [71]. ©*2016 Intuitive Surgical, Inc.*

Vinci™. Concerning the robot system, an often-mentioned drawback is the possibility of collisions between the external parts of the robot arms [118]. Goldstraw et al. name this as a "serious problem" as it can lead to prolonged intervention times due to necessary re-docking of the robot [42]. Another disadvantage is the sheer size and weight of the da Vinci™. Apart from necessitating structural enhancements for some OR to support the robot's weight, it can also literally make the intervention revolve around the robot [42], as some clinics have found that moving the OR table around the robot is in fact more practical than moving the robot to the correct location at the OR table [118].

2.2.1.3. MiroSurge

The *MiroSurge* system was developed by the *DLR* as "a versatile system for research in endoscopic surgery" [50]. Similar to the da Vinci™, it is also mainly targeted at teleoperated usage for MIRS. However, the design of the MiroSurge robot system is very different from the da Vinci: Instead of a monolithic structure that holds the robotic arms, MiroSurge consists of multiple small, independent robotic arms that are directly attached to the OR table (see Figure 2.3).

The lightweight MIRO arms have been specifically designed with the goal of operating in an unstructured environment and interacting with humans. They feature integrated joint torque sensors that allow for sensing external forces, which enables the system to react to its environment, e.g. by detecting collisions and offering compliant control for hands-on guidance [179].

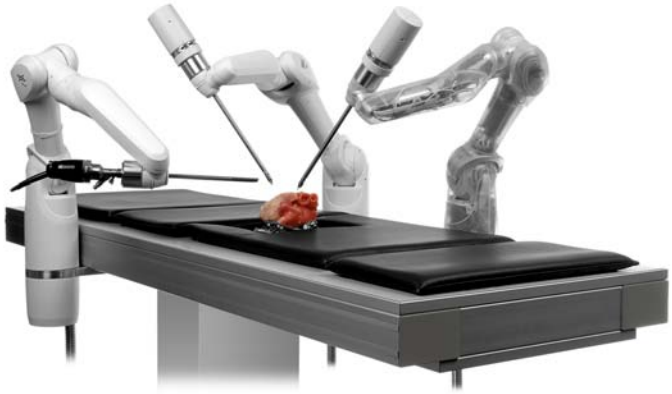


Figure 2.3.: The MiroSurge system with three lightweight MIRO arms that are directly attached to an OR table [37].

To aid the OR personnel in the correct setup of the robot system, an additional tool called VR-Map has been developed that can be mounted to a MIRO arm and integrates a 3D scanning system and a laser projector [94]. The intended usage is to first register the patient body by scanning its surface, followed by a re-optimization of the planned setup. The resulting trocar point positions and robot positions alongside the OR table are then projected onto the patient and the OR table. However, no intraoperative projections can be realized using this system, as it is used only preoperatively and replaced by a surgical instrument before the start of the intervention.

2.2.1.4. Further systems

There are numerous research systems as well as systems that are currently being developed for commercialization. In the following, a selection of such systems that are targeted at MIRS is presented very shortly. For further reference, an exhaustive list can be found in [60]. For the purpose of this work, the listed systems are categorized into two groups based on how the robotic arms are positioned at the OR table, either on stand-alone carts or directly attached to the OR table.

Systems using stand-alone carts:

- Two Italian systems are in the process of being commercialized: The *ALF-X*, developed by *Sofar S.p.A.*, and the *Surgenius* by *Surgica Robotica S.p.A.*. Both systems feature multiple independent robotic arms that are each mounted on dedicated carts. The ALF-X aims to lessen the impact of the robot system on the available space around the situs by extending the arm lengths and height so that the carts can be positioned farther away from the OR table (see Figure 2.4). It has already completed first human applications in a Phase II study [32] and was acquired by *TransEnterix*, USA, in September 2015.

Figure 2.4.: Upcoming surgical robot systems with different concepts. *left*: The ALF-X system features extendable stand-alone carts that are designed to free up space around the situs [165]). *right*: The SOFIE system is directly attached to the OR table and requires no floor space [30].

- The *BIT* system (to be renamed in *Cordoba*) makes use of standard industrial *UR5* robots by *Universal Robots* that are mounted onto a specific cart each, which also houses an industrial PC. Connection between these PCs controlling the robots and the surgeon console is established via Robot Operating System (ROS) [8].
- The *Bitrack* was developed at the *Institute for Bioengineering of Barcelona* and is now in the process of being commercialized by *Rob Surgical S.L.*. It consists of one tall cart that holds two robotic arms and was designed for minimal space requirements.

Systems with direct attachment of the robotic arms to the OR table:

- The *SOFIE* system was developed at the *Eindhoven University of Technology* and focuses on force feedback in multiple dimensions [186]. It is an example for a research system that uses small robotic arms which are not mounted on stand-alone carts. Rather, they are attached to a single frame which is in turn attached to the the OR table (see Figure 2.4).
- The *Raven II* robot from *Applied Dexterity* is an open source surgical research robot system that is being used in multiple universities worldwide [53]. It employs small, cable-driven robotic arms that are mounted onto the OR table.
- The *Robin Heart* robot is developed at the *Foundation for Cardiac Surgery Development* since 2000 and is targeted towards minimally invasive cardiovascular interventions [110]. Multiple small robotic arms that have specifically been designed for a small footprint are directly mounted to the OR table.

Each of these systems focuses either on a specific medical target application in the domain of MIRS, e.g. cardiovascular abdominal interventions, or on a more technical research direction, such as providing force feedback or enabling an open

source research platform. However, none of these systems includes any sensors that allow to perceive and react to the immediate surroundings of the robot.

2.2.1.5. GestoNurse

A perioperative nurse, often called scrub nurse, is responsible e.g. for preparing the surgical instruments before an intervention and for assisting the surgeon intraoperatively. On demand, e.g. spoken or via gesture, scrub nurses need to quickly pass the surgeon the correct instrument and clean and place it back afterwards. They also have to count all instruments as well as sponges and other tools after surgery to make sure that none was left in the situs.

Purdue University, USA, developed the *GestoNurse* as a robotic scrub nurse with a multimodal user interface. By speech or gesture command, the surgeon can request instruments which are handed over by the robot. Contrary to the systems mentioned above, the GestoNurse is equipped with sensing capabilities, e.g. in the form of 2D/3D cameras and microphones, to detect and interpret commands as well as detect the surgeon's hand position for safe collaboration. Jacob et al. report a command recognition accuracy of over 97% and time reductions using the multimodal interface over speech-only of 14.9% in a mock-up intervention [81, 82].

2.2.2. Range Imaging

There exist various sensing principles and systems for acquiring range images. The following sections give an overview of the conceptual as well as technical background of such systems, focusing on range imaging cameras with active illumination.

2.2.2.1. Background

Range Imaging is the process of acquiring three-dimensional information of a surface or a scene from a certain viewpoint. A range image is a 2D image in which the pixel value at each coordinate corresponds to the distance to the object at that coordinate from the perspective of the viewpoint. Combined with the intrinsic parameters of the range imaging device, range images can be used to reconstruct a *point cloud* that consists of multiple 3D points where each point corresponds to one pixel in the range image.
A point cloud reconstructed from a single range image technically represents a 2.5D model: While it contains 3D information, this information is only available from a single viewpoint, in a viewer-centered coordinate frame. In contrast, a 3D model representation would use an object-centered coordinate frame and contain

volumetric information and surface primitives for the whole model [112]. However, 3D models can be created by using range images acquired from multiple viewpoints.
For easier reading, range imaging devices will be referred to as *3D cameras* throughout this thesis. The range image itself will also be referred to as a *distance map*, if each pixel value corresponds to the distance between the camera sensor and the scene geometry, or as a *depth map*, if each pixel value corresponds to the distance between the plane of the camera sensor and the scene geometry.

2.2.2.2. Structured Light

Sensing principle The basic principle of structured light consists of projecting a light pattern into a scene which is then captured by a camera. Based on the known geometry ("structure") of the projected light pattern and the known spatial relation between camera and projector, a 3D view of the scene can be reconstructed. Many implementations of structured light systems exist, often consisting of a standalone projector and a traditional RGB camera, that use different algorithms for calibration and acquisition of 3D information.

Structured light measurement systems rapidly gained wide-spread adoption in robotics research when the first version of the *Microsoft Kinect* was brought to market [51]. It contains a complete structured light system in a small housing which offered unprecedented 3D sensing capabilities in this price range. In the following, the structured light sensing principle will be further detailed, mainly based on the Kinect implementation.

Figure 2.5.: Optical components of the Microsoft Kinect for XBox 360.

The Kinect is based on a proprietary structured light system by *PrimeSense* that has been licensed by Microsoft. It consists of three optical components as depicted in Figure 2.5:

- *Infrared laser projector*: A static speckle pattern is projected into the scene using a $780\,\mathrm{nm}$ infrared laser diode combined with a custom holographic diffraction grating. The pattern is constructed based on *spatial-multiplexing*

light coding: As each point in the projected pattern needs to be robustly detectable ("coded") and no information can be encoded by temporal changes, each point has a unique spatial neighborhood which is used as its *spatial multiplexing window support*. More specifically, the projected pattern features an uncorrelated distribution across each row [23]. The pattern is depicted in Figure 2.6.

- *Infrared sensor*: A grayscale sensor combined with an Infrared (IR)-pass-filter is used in combination with an astigmatic lens with different focal lengths for x-axis and y-axis to acquire an infrared image of the scene. The different focal lengths result in different shapes of the detected points of the projected pattern, depending on their distance to the camera, and thereby provide additional clues to distinguish near and far points.
- *RGB sensor*: An off-the-shelf RGB sensor provides a 2D color image. Using the extrinsic calibration between the IR sensor and the RGB sensor, the color information can be mapped onto the acquired depth data.

To reconstruct a 3D representation of the scene, correspondences between the emitted IR light pattern and the acquired IR image are determined. Taking into account additional effects such as perspective distortion, the disparity d at each pixel is calculated. According to Khoshelman et al. [88], in PrimeSense devices the disparity is stored as a (de)normalized 11 bit integer d' with supposedly linear normalization coefficients m and n such that $d' = (d-n)/m$. One bit of d' is further reserved to encode invalid pixels where no disparity could be determined, so $2^{10} = 1\,024$ discrete values remain for encoding the disparity.
Taking into account the focal length f of the camera as well as the base length b between the optical centers of the projector and the camera, the depth resolution Δz for a given disparity d' can be calculated as

$$\Delta z(d') = Z(d') - Z(d'-1), \tag{2.1}$$

which leads to the following formula for calculating the depth error at a certain distance Z:

$$\Delta z = (\frac{m}{fb} Z^2). \tag{2.2}$$

Based on their calibration results, Khoshelham et al. determine the factor $|\frac{m}{fb}|$ to $2.85\,10^{-5}$. Using Equation 2.2, this results in a maximal depth resolution of $2.8\,\mathrm{mm}$, $25.7\,\mathrm{mm}$ and $71.3\,\mathrm{mm}$ at respective distances of $1\,\mathrm{m}$, $3\,\mathrm{m}$ and $5\,\mathrm{m}$.

While the Kinect is manufactured as a low-cost consumer device, Martinez and Stiefelhagen investigate the depth reconstruction pipeline of the integrated *PS1080* chip with the focus on determining potential improvements on the depth reconstruction [113]. By estimating depth at the projected points only and performing

no interpolation between these points, they report achieving a higher depth accuracy than the original algorithm. The drawbacks of this reconstruction pipeline are a lower lateral resolution of about $28k$ points and a reduced frame rate of $1.75\,\mathrm{fps}$ compared to the original algorithms [114].

Characteristics Structured light systems such as the Kinect exhibit special characteristics:

- *Depth resolution*: The depth reconstruction error is a quadratic function of the measured distance. This leads to an increasing uncertainty in measurements at larger distances which impedes e.g. reconstruction of angled surfaces as they are "projected" onto discrete distance values.
- *Sensitivity to external light*: Structured light range imaging depends on detecting a projected pattern. Strong external illumination can lower the contrast of the detected speckle pattern which leads to an increasing number of invalid distance measurements. For the Kinect, experiments have shown that measurement errors occur when using external light sources with an irradiance of $6 - 7\,W/m^2$ [102]. As sunlight has an irradiance of about $75\,W/m^2$ in the relevant wavelengths between $800\,\mathrm{nm}$ and $900\,\mathrm{nm}$, this negatively influences or completely prevents the usage of the Kinect in environments with direct sunlight, e.g. in outdoor applications or in rooms with large windows.
- *Influence by scene properties*: Color and reflectivity inhomogeneities of materials in the scene can influence the perceived contrast and detection quality of the projected structured light pattern, leading to incorrect or missing depth reconstruction [23]. Furthermore, the geometry of the scene plays an important role as heavily slanted surfaces introduce a strong perspective distortion of the structured light pattern as seen by the IR sensor. This prevents the correct determination of disparities and results in failures in depth reconstruction.
- *Depth discontinuities and occlusions*: Due to the baseline between the IR projector and the IR sensor, objects closer to the camera occlude points farther away from the camera, making them invisible to either the IR projector or the IR sensor. This results in missing depth estimations in the background of a scene at the boundaries of foreground objects. In the case of the Kinect, heuristics are in place to attempt reconstruction of this missing information. However, these can in turn lead to misalignments between real and estimated depth discontinuities [23].

Camera models The original technology behind the Kinect was also licensed to *Asus* and subsequently put on the market as *X-tion Pro* and *X-tion Pro Live*. Regarding the 3D sensing capabilities, both cameras are identical to the Kinect. In November 2013, PrimeSense was acquired by *Apple*, USA, which subsequently led to the discontinuation of the X-tion product line.

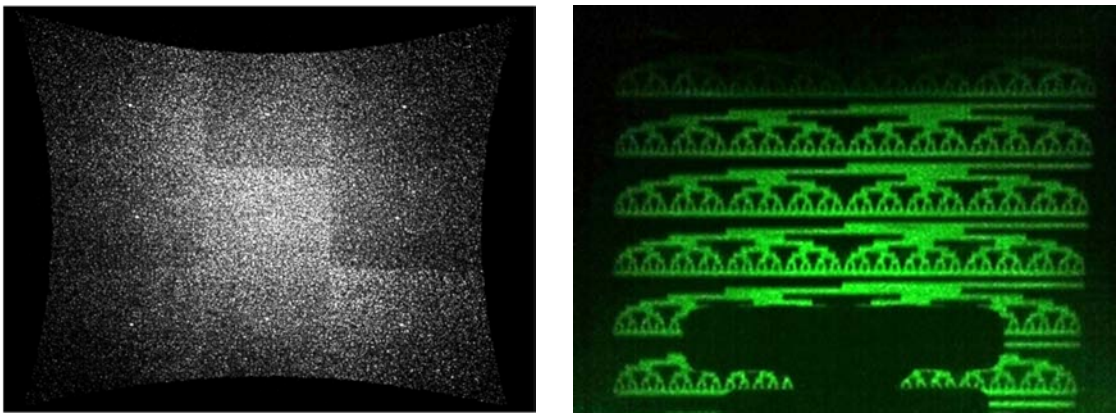


Figure 2.6.: Examples for structured light patterns. *Left*: Microsoft Kinect projects a static pattern that contains enough detail to uniquely identify the spatial neighborhood of each point [23]. *right*: Intel RealSense F200 projects multiple patterns like the depicted one. Point identification is possible only due to temporal information [177].

Orbbec, USA, produces the *Astra* cameras, which are based on structured light technology developed by Orbbec that offers a lateral resolution of up to $1\,280 \times 1\,024\,px$. Contrary to PrimeSense-based devices, the Astra cameras offer an SDK that allows to control the camera on a technical level, e.g. adjusting the IR sensor gain to change the sensing range and controlling the laser projection to enable crosstalk-free temporal multiplexing of multiple cameras. In late 2015, the new *Persee* camera was announced that will integrate a host computer in the camera housing at a similar size to e.g. the Kinect. It features a quad core ARM processor at up to $1.8\,\mathrm{GHz}$ with a dedicated GPU as well as an Ethernet port and wifi connectivity, thereby facilitating the easy distribution of multiple cameras in a room [139].

In 2015, *Intel* introduced two models of its *RealSense* camera family as development kits: The *F200* is a close-range cameras designed for detecting a user' s head and hands, the *R200* is a medium-range camera aimed at environmental sensing. Both provide a lateral resolution of $640 \times 480\,px$ at up to $60\,\mathrm{fps}$. While the F200 features the same types of components as the Kinect (IR projector, IR sensor, RGB sensor), it employs a temporal-multiplexing projection approach based on a MEMS micro-mirror by *STMicroelectronics*, Switzerland [167]. The amount of patterns used per frame can be configured and influences achievable frame rate as well as accuracy [69]. For illustration of the difference to static patterns, one of the patterns is depicted in Figure 2.6. In contrast to the F200, the R200 features two IR cameras and employs a stereoscopic scene reconstruction approach based on the disparity between both IR cameras [22].

2.2.2.3. Time-of-Flight

Sensing principle Time-of-Flight (ToF) cameras acquire distance information by sending out (near-)infrared light and measuring the time delay until the light reflected by the scene is being detected by the according pixel on the sensor (see Figure 2.7). As ToF sensors consist of an array of such pixels that are read out in parallel, ToF cameras capture the full scene at the same time, instead of scanning it line-by-line like e.g. laser scanners.

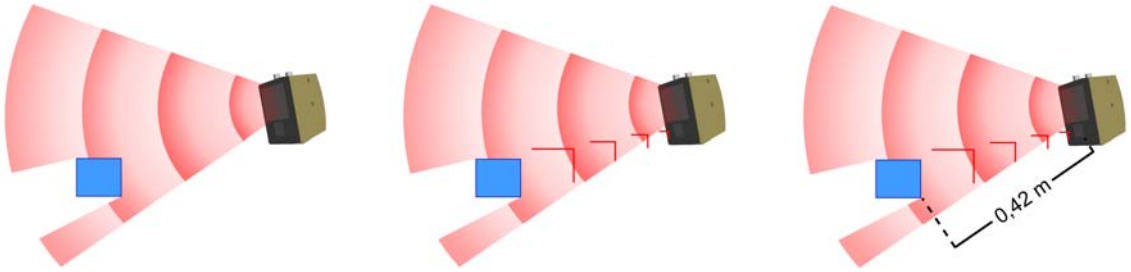


Figure 2.7.: Illustration of the ToF measurement principle. The camera emits modulated infrared light *(left)*, which is reflected by objects in the scene *(center)*. Distance to the object is calculated based on the phase shift between emitted and reflected light *(right)*.

This work mainly uses ToF cameras with Photonic Mixer Device (PMD) sensors by ***pmd****technologies*, Germany. Therefore, in the following, the amplitude-modulated ToF measurement principle is described based on the technical implementation used in PMD sensors [156].

At each pixel of the PMD-sensor, the time delay between emission and detection of the reflected light directly corresponds to the phase difference $\Delta\phi$ between the emitted and reflected infrared signal. $\Delta\phi$ can be calculated based on the relation of four phase images, acquired using electric charge values controlled by four phase control signals with $90°$ phase delay between each other, that determine the collection of electrons from the received IR light (see [54] for further details).

With the known speed of light c and the signal modulation frequency f, the distance d to the reflecting object can be calculated as

$$d = \frac{c}{2f}\frac{\Delta\phi}{2\pi}. \tag{2.3}$$

On industrial-grade ToF cameras, multiple adjustable settings influence the quality and performance of the distance measurements:

- *Integration time*: The electric charges used to calculate the phase difference are integrated over a fixed period of time. Increasing this integration time leads to a better Signal-to-Noise-Ratio (SNR), but decreases the achievable maximum frame rate as less measurements can be taken per time interval.

- *Frame rate*: Depending on the camera model and cooling solution, thermal constraints can further restrict the available combinations of integration time and frame rate [14].

- *Modulation frequency*: Due to the four-charge implementation for calculating the phase difference, the phase ϕ of the modulation signal can only be measured in the range between $[0, 2\pi)$ (see [54] for further details). Therefore, the maximum distance d_{max} which can be reliably measured is

$$d_{max} = \frac{c}{2f}. \tag{2.4}$$

  For a more intuitive understanding, this means that light can travel at most one full wavelength from the camera to an object and back before its phase becomes indistinguishable from the next wave cycle. Therefore, d_{max} is equal to half the wavelength λ_{mod} of the modulated light. For common modulation frequencies between 20 MHz and 30 MHz, this corresponds to an unambiguity range of up to $\approx$7.5 m and $\approx$5 m, respectively.

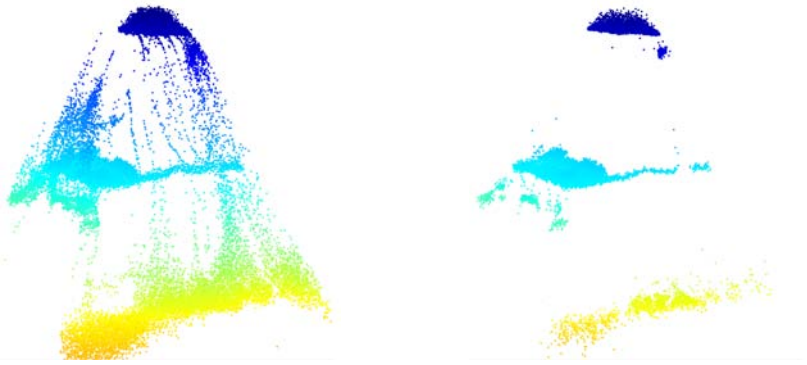


Figure 2.8.: Flying pixel phenomenon visible as color-coded side view of scene captured by a ceiling-mounted [pmd]vision CamCube 2.0. *Left:* Unfiltered point cloud exhibiting a significant amount of flying pixels; *right:* Filtered point cloud.

Characteristics Compared to 2D grayscale or RGB cameras, ToF cameras offer only a low lateral resolution (see Table 2.2). As a consequence of their measurement principle, they furthermore exhibit special characteristics that have to be taken into account:

- *Signal interference*: If multiple ToF cameras with the same illumination parameters (i.e. modulation frequency and IR wavelength) are operated in the same environment, reflections of light emitted by one camera are also received by other cameras. This crosstalk results in heavily distorted measurements and therefore has to be carefully avoided for multi ToF camera systems.

- *Flying pixels*: Due to the low lateral resolution of ToF cameras, depth inhomogeneities, which occur for example at the boundaries of objects, often lead to incorrect measurements as light reflected by an object and light reflected by the background are both integrated in the same pixel. This results in an incorrectly measured distance value that is between the distances of the object and the background. This phenomenon is commonly referred to as flying pixels ([18, 104]) and is depicted in Figure 2.8.
- *Multi-path reflections*: ToF distance calculation is based on the assumption that the emitted light is reflected directly back to the sensor. If there are multiple reflections instead, e.g. in the case of a sharp inside corner junction of a wall, the distance is incorrectly measured.
- *Motion artifacts*: The four phase images, which are used in PMD sensors to sample the autocorrelation function and calculate the phase shift, are taken successively. If motion occurs between any of these sample images, this leads to incorrect distance values at object boundaries [93].
- *Intensity-related distance errors*: Especially with early generations of PMD-sensor based ToF cameras, the reflectivity of the object surface influences the distance measurement due to physical effects of both the semiconductor detector and the camera electronics [93].
- *Temperature drift*: Due to the high responsivity of semiconductor materials to temperature changes, temperature variations within ToF cameras can affect the measurement accuracy. According to [148], temperature changes due to internal warming up of ToF cameras can account for maximum variations of about 120 mm for the measured mean distance. Operating the camera for a device-dependent warm up period of up to 40 min before conducting measurements can bypass these effects.
- *Multi frequency acquisition*: In order to increase the unambiguity range, multiple depth acquisitions with different modulation frequencies can be combined. This decreases the achievable frame rate and disables the use of frequency multiplexing schemes, as each camera utilizes multiple frequencies.

Further, ToF cameras capture the amplitude of the received signal at each pixel. It is used as an indicator of the validity and quality of the according measurement and internally evaluated in the camera. For industrial-grade ToF cameras, the amplitude value per pixel is usually accessible as an *amplitude map* while it is not accessible with consumer-grade ToF cameras.

Camera and sensor models For a number of years, ToF cameras have been targeted solely at industrial use due to a number of reasons including technical limitations, high price due to low volume manufacturing and limited lateral resolution. Common examples for older and current stand-alone industrial ToF camera models with PMD sensors are the [pmd]vision® CamCube 2 and 3 as

well as the [pmd]vision® S3, both by **pmd**technologies, and the *Argos^{3D} P100* by *Bluetechnix GmbH*, Austria. The well-known *SwissRanger 4x00* series by *MESA Imaging AG*, Switzerland, has been discontinued as of November 2015 after the acquisition of MESA Imaging AG by *Heptagon*, Singapore [58].

Texas Instruments (TI), USA, offers different ToF sensors with standard specifications. These are employed e.g. in the short-range, consumer-grade web camera *Senz3D* by *Creative*, Singapore, and the (discontinued) short-range *DepthSense 325* by Belgian company *SoftKinect*, which has been acquired by *Sony* in October 2015.

ESPROS Photonics AG, Switzerland, rolled out the *epc660* System on a Chip (SoC) in 2015. At a lateral resolution of $320 \times 240\,px$, it features modulation frequencies between $0.625\,\mathrm{MHz}$ and $20\,\mathrm{MHz}$, which correspond to unambiguity ranges of up to $240\,\mathrm{m}$. The epc660 provides so-called "full 3D TOF frames", that are combined from four consecutive phase images called "Differential Correlation Samples" (DCS), at a rate of $65.5\,\mathrm{fps}$. As a minimum of two DCS per frame is necessary for distance calculation, the frame rate can be doubled to $131\,\mathrm{fps}$. The frame rate is limited by the read-out bandwidth, so reducing the Region of Interest (ROI) further increases the frame rate proportionally to the reduction factor up to a maximum full 3D TOF frame rate of $1\,048\,\mathrm{fps}$ at a ROI of $\frac{1}{16}$ of the full resolution. In addition to the image sensor, four temperature sensors are integrated to be used for drift compensation [31].

Microsoft switched the range imaging method for their consumer-grade Kinect cameras from structured light to ToF with the introduction of the second generation Kinect. While Microsoft officially names the first generation "Kinect for XBox 360" and the second version "Kinect for XBox One", there exist various naming schemes in different publications. In the following, the cameras will be referred to as *Kinect v1* and *Kinect v2* for the sake of simplicity.
The Kinect v2 provides depth information at $30\,\mathrm{fps}$ with a lateral depth resolution of $512 \times 424\,px$, which is comparatively high for a ToF camera. It employs multi-frequency acquisition using three widely spread modulation frequencies of approx. $16\,\mathrm{MHz}$, $80\,\mathrm{MHz}$ and $120\,\mathrm{MHz}$ [18] to increase the non-ambiguity range up to $18.74\,\mathrm{m}$ by using frequency-based phase unwrapping. According to a joint paper by Microsoft Research and Technische Universität München (TUM) [168], the Kinect v2 internally works with an acquisition frequency of $300\,\mathrm{Hz}$. At the beginning of each measurement, internally nine frames are acquired in rapid succession at different combinations of three frequencies of laser illumination and the three modulation frequencies. These are followed by a tenth infrared frame captured without active illumination which is used to correct for external lighting effects.
According to Breuer et al. [18], the depth resolution of the Kinect v2 is comparable to that of its predecessor both in quantity and quality as it also decreases quadratically with the distance. Furthermore, they note that the Kinect v2 has a temperature drift which they evaluated to approx. $2.5\,\mathrm{cm}$ in the first 25 minutes of operating time.

odos imaging, Scotland, manufactures the *Real.iZ™ OI-VS-1000* that operates with pulsed illumination instead of amplitude modulation, i.e. it directly measures the time delay of a non-modulated light pulse instead of calculating the delay based on the phase shift of a continuous modulated signal. The camera offers a high resolution of $1\,280 \times 1\,024\,px$ at up to $30\,\mathrm{fps}$ with the additional option of a high-speed mode with up to $450\,\mathrm{fps}$ that can only be stored internally. As the emitted light is not modulated, multiple cameras can only be used in a time-multiplexing setup [136].

Since December 2015, multiple large companies have announced new generations of ToF cameras and sensors:

- **pmd**technologies and *Infineon*, Germany, announced the new *REAL3™* image sensor family with a resolution between 19 k and 100 k pixels. It is advertised as improving the photo-sensitivity by a factor of two compared to the previous sensor generation due to optimized micro-lens technology. Acquisition speed is given as up to $100\,\mathrm{fps}$ using modulation frequencies of up to $100\,\mathrm{MHz}$ [67].

- TI announced the new *OPT8320* SoC, that, according to its specifications, far surpasses currently available ToF sensors in terms of acquisition speed. At a relatively low resolution of $80 \times 60\,px$, it offers a frame rate of up to $1\,000\,\mathrm{fps}$, which is in turn based on an internal raw frame rate of to $4\,000\,\mathrm{fps}$. The modulation frequency is specified as $10\,\mathrm{MHz}$ to $100\,\mathrm{MHz}$ from which a native unambiguity range of about $15\,\mathrm{m}$ to $1.5\,\mathrm{m}$ can be calculated. Additional information such as auxiliary depth data representing the amplitude of the received signal are available [176].

- In January 2016, Heptagon announced the *TARO 3DRanger™* as successor to the discontinued SwissRanger series. While no technical improvements in terms of resolution or frame rate are advertised, an optional wireless interface will be available that allows for connectivity via wifi, bluetooth and zigbee [57].

- *Basler*, Germany, announced their new engineering sample of a ToF camera that is expected to be available in 2017. It features a high lateral resolution of $640 \times 480\,px$ at a rather low frequency of $15\,\mathrm{fps}$ [6].

Further technical information about specifications of a selection of the aforementioned ToF cameras can be found in Table 2.2.

2.2.2.4. Further sensing principles and research

In addition to structured light and ToF, there are several other technologies available that offer 3D sensing. As they are currently not suitable and/or available for usage in medical scenarios, they will only be described shortly.

	PMD S3	CamCube 2.0	Argos 3D P100	Mesa SR 4000	Kinect v1	Kinect v2
Lateral resolution (px)	64×48	204×204	160×120	176×144	640×480	512×424
Field of view	$30° \times 40°$	$40° \times 40°$	$90° \times 90°$	$69° \times 56°$	$43° \times 57°$	$70° \times 60°$
IR wavelength (nm)	850	870	850	850	830	860
Connection	Ethernet	USB 2.0	USB 2.0	USB 2.0 / Ethernet	USB 2.0	USB 3.0
Frame rate (fps)	up to 20	up to 25	up to 160	up to 54	30 (fix)	30 (fix)
Operating range (m)	$0.2 - 6.0$	$0.2 - 7.0$	$0.1 - 3.0$	$0.3 - 5.0$	$0.4 - 4.5$	$0.5 - 4.5$
Modulation frequency (MHz)	20.4, 20.6, 23.0	18.0, 19.0, 20.0, 21.0	$5-30$ (adjustable)	$29 - 31$ (adjustable)	–	$10 - 130$ (non-adjustable)

Table 2.2.: Technical specifications for different 3D cameras [14, 119, 142]

Plenoptic cameras *Plenoptic* or *light field* cameras offer monocular, passive 3D sensing and require no artificial illumination. Contrary to conventional camera systems, where a single lens is used to focus the incoming rays of light so they converge on one pixel on the sensor, plenoptic cameras usually employ an array of microlenses, often placed between the main lens and the image sensor, to split incoming light by angle of its direction. Contrary to e.g. ToF cameras, expensive computation is necessary for reconstruction of depth information. Therefore plenoptic cameras usually employ highly parallel algorithms on according hardware such as Graphics Processing Units (GPUs) or Field Programmable Gate Arrays (FPGA). As both angular and positional structure of incoming light rays have to be recorded simultaneously, the achievable lateral resolution is smaller than the native resolution of the sensor [146, 158].

Plenoptic cameras following this sensing principle are commercially available from the companies *Lytro*, USA, and *Raytrix*, Germany. The first generation of Lytro cameras is reported to have a sensing range of approx. $210\,\mathrm{mm}$ with a depth reconstruction error of $30\,\mathrm{mm}$ under laboratory conditions [153]. The second generation Lytro *Illum* is a handheld camera focused on the consumer market with the application of changing focus in a picture after it was taken. While it offers a 2D export resolution of $2\,450 \times 1\,634\,px$, acquisition speed is limited to $3\,\mathrm{fps}$ [108]. In April 2016, Lytro announced that it would not develop its consumer camera lineup any further and focus on light-field capturing for business customers instead. In contrast, Raytrix offers various models of its *R* series targeted at industrial and scientific use. Depending on the model, they offer effective resolutions between 1 megapixel at $180\,\mathrm{fps}$ and 10 megapixel at $7\,\mathrm{fps}$ and support industrial standard lens mounts. An analysis of a first generation Lytro camera and the Raytrix R5 are published in [41] and [197], respectively.

Venkataraman et al. developed a different approach to a plenoptic camera, the so-called "array camera" *PiCam*. Rather than using one image sensor, it consists of an array of separate, optically isolated image sensors that are sensitive to a single spectral color each. Using a 4×4 array of sensors with a resolution of $1\,000 \times 750\,px$ each, they derive an 8 megapixel depth image at a camera package height of only $3.5\,\mathrm{mm}$. However, the error of estimated depth increases with increasing distance from $1.1\,\%$ at $0.2\,\mathrm{m}$ to $38.9\,\%$ at a distance of $5\,\mathrm{m}$ [187].

Shape from Polarization (SfP) *Photon-X*, USA, have developed a proprietary, passive 3D volumetric imaging technology called *Spatial Phase Imaging (SPI)* that simultaneously captures color, 3D coordinates and 3D shape (e.g. normals) per pixel. According to [5], it uses a single, enhanced camera and requires no artificial illumination or coherent light as it operates based on the spatial phase properties of any light source, including ambient light. Incoming light is broken down into spatial phases with different orientations, out of which a phase difference between adjacent pixels can be calculated. Based on the known Field of View (FoV) and pixel size, depth information can then be geometrically reconstructed up to a depth resolution that directly depends on the lateral resolution upon the object. Photon-X claims that SPI scales with the optical system used and can therefore be arbitrarily configured in terms of resolution, sensing range and frame rate. Since Photon-X has carried out mostly military and defense work, no further details are publicly available.

Kadambi et al. have proposed to combine coarse depth maps obtained by ToF cameras with SfP cues in order to improve the resulting depth reconstruction [84]. Using a Kinect v2 and a DSLR with linear polarizer, they were able to recover features in the size of $300\,\mu\mathrm{m}$ from short distances. However, their method requires three consecutive images throughout which the scene cannot change, takes about one minute for full depth reconstruction and has specific requirements for the surface materials. Therefore, it is not yet applicable to real world scenarios.

Multi-Image sensing *photoneo*, Slovakia and *ximea*, Germany, reported to have developed a laser-projection based 3D sensing technology called "Multi-Image sensing" that can supposedly deliver motion-blur free 3D reconstruction and will be available on the market in the *PhoXi Cam++* in 2016. With a resolution of up to 2.7 million pixels and up to $60\,\mathrm{fps}$, it is advertised to feature a very high depth accuracy of up to $50\,\mu\mathrm{m}$ at $1\,\mathrm{m}$ and 2σ [195].

2.3. Applications

2.3.1. Surgical applications

2.3.1.1. Range imaging

In recent years, there was a growing interest in the application of range imaging for health care purposes. While this is partially due to technological progress, e.g. the availability of small and low-cost 3D cameras, there are several benefits with regard to medical practice: Based on real-time dense measurements, range imaging offers the possibility of touch-free interaction in sterile environments and marker-free tracking which can save effort in setup.

Research on range imaging in surgical interventions can be categorized into three main applications:

- *Guidance in computer-assisted interventions*, e.g. 3D laparoscopy, which is detailed further in section 2.3.1.1.
- *Monitoring of the OR*, which is covered in the main part of this thesis.
- *Touch-less interaction*, which is discussed in section 2.3.1.1.

Bauer et al. present an overview of range imaging in the general domain of health care, focusing on ToF-based research [7]. Some of their findings will be summarized in the following.

Concerning the usage of ToF cameras for monitoring OR safety, they put forward a set of requirements on both the characteristics of the cameras and their spatial setup: Cameras need to offer a low latency with the exact timing dependent on the movement speeds in the supervised scene. The accuracy of the whole system depends on the accuracy of each single sensor, so the accuracy of the single sensors has to be known and matched with the target application. For collision avoidance, they suggest a safety margin in the centimeter range as motions cannot be stopped instantaneously. In order to avoid or mitigate occlusions and provide redundancy, a multi-camera setup is recommended. Work performed in the scope of this thesis [125, 133] is cited as examples of such systems.

Further, Bauer et al. discuss current issues and limitations of range imaging in medical applications. These include systematic errors, such as temperature-dependent distance measurements and the flying pixel phenomenon, and possible (partial) solutions such as a preoperative warm-up phase and a fix integration time to keep the thermal conditions constant. If multiple cameras are to be used in the same environment, a multiplexing scheme has to be devised to avoid cross-talk. While these considerations are fairly generic and hold true for various domains, they emphasize specific requirements for allowing a proper integration into clinical routine. These include the necessity of a simple and reliable process for (re)calibration

of multi-camera systems as well as choosing a form-factor that can be used in e.g. an OR setting without obstructing the personnel.

3D laparoscopy *Lucidux*, USA, is in the process of developing a 3D laparoscope based on SPI technology, naming Intra-procedural Detection of Sub-Surface Cancer Nodules in Lung as an application [147]. In addition to 3D imaging for purely visual purposes, i.e. better visualization, the product might also provide capabilities for reconstruction of tissue properties. This is indicated by a patent that has been filed in 2011 and generally describes a system and methods for measuring mechanical properties of deformable materials. It explicitly names different medical applications that involve probing tissue for determining its internal properties as well as minimally invasive surgery [55].

In [115], Herrera et al. present a prototype of a 3D laparoscope using SfP. They extend a standard laparoscope shaft with a rotatable polarizer at the tip. Reconstruction of the shape of the surface is performed based on three images acquired with different polarizer orientations. For ex vivo experiments using lamb organs, an angular reconstruction error of the surface normals below $15°$ is reported.

Richard Wolf, Germany, manufactured a prototype for a ToF-based 3D laparoscope that offers depth information at a frame rate of $20\,\mathrm{fps}$ and a lateral resolution of $64 \times 48\,px$ [144, 145]. While it is not commercially available, the prototype has been extended and used in several research projects [46, 92].

Touch free interaction During surgery, a surgeon often needs to review 2D or 3D pre- or intraoperative imaging data. This can be a cumbersome undertaking as mouse and keyboard that are used to control the display of said imaging data are in a non-sterile zone of the OR. Directing an assistant by voice to show the right parts of the image on screen may not result in the optimal view that the surgeon desires, especially for 3D data. If surgeons need direct hands-on control over the images, they have to leave the sterile zone to browse the imaging data and then spend additional time on re-sterilization, i.e. rescrubbing, before they can resume the intervention.
Touch-free control of e.g. DICOM viewers in the OR is a modality that can help surgeons overcome these delays as they can control the displaying of imaging data directly from the sterile zone by themselves. Various works have therefore focused on using e.g. the Kinect as an input device to enable such gesture-based interaction (see e.g. [137]).

Karl Storz, Germany, offered a first commercial system for touch free control of patient records under the name *MI Report* that used an infrared stereo camera for hand segmentation and was based on research by *Fraunhofer HHI* [20]. An evaluation carried out during 51 surgical interventions and a survey upon 25 surgeons showed high interest and positive results. However, the activation gesture for starting touch-free control failed in several cases, resulting in $31\,\%$

failed attempts where usage was cancelled by the surgeon before the system was activated [28].

As of 2015, the *TedCube* is available as a commercial solution by *TedCas*, Spain. It is compatible with different gesture input devices such as Kinect v1, Kinect v2, SoftKinetic DS325 and Leap Motion and acts as a bridge between the input device and a DICOM viewer plugin for gesture control. The FDA has stated that the TedCube itself is not considered a Medical Device and therefore does not need to be certified as such [174].

2.3.1.2. Augmented Reality

There are various works on the application of different forms of AR in the OR. The vast majority of these publications focuses on direct medical benefits of applying AR, i.e. by assisting the surgeon with display of medical information [106]. In the following, a short overview of related works and their specific AR-modality is given, using the classification introduced in section 2.1.6.

Kersten-Oertel et al. review the state of the art of visualization in mixed reality image-guided surgery. Concerning the display modality, they report that taking into account all surgical domains, monitors have been used in $47\,\%$ of the analysed works, followed by HMD ($20\,\%$), microscopes ($16\,\%$) and projectors ($9\,\%$). These numbers vary by the respective surgical domain: In neurosurgery, microscopes are used in $52\,\%$ of the analysed works; in endoscopic interventions, monitors are used in $57\,\%$; in maxillofacial surgery, projectors are the most used modality with $38\,\%$ [87].

While HMDs seem to be a preferred display modality according to these numbers, Wen argues that they have several inherent drawbacks. These include limits to the field of view of the surgeon, ergonomic limitations imposed by the headsets and the lack of multiple observers. Furthermore, potential adverse effects of HMDs are listed: separation of the surgeon from the medical scene due to indirect view, need for manual registration and issues with tracking and timing synchronization [190].

Eggers et al. study the usage of intraoperative AR in preclinical trials with a HMD-based and a two-projector-based system [85]. They conclude that wearing glasses for a whole intervention would be too cumbersome, but that see-through concepts offer a higher flexibility as well as the possibility for stereoscopic AR. Projection-based systems have the advantage of not requiring additional hardware close to the surgeon and providing an augmentation which is visible to all of the staff. In total, they find that both modalities complement each other, while the "projector based system is more comfortable and integrates better into the surgical workflow." [29].

HMDs Dickey et al. use the *Google Glass®* as an optical-see-through HMD to aid in inflatable penile prosthesis. As an interactive modality, a remote physician could see the live video stream captured by the HMD and provide visual (and auditive) information to the surgeon or trainee. The method received high ratings in terms of usefulness and ease of navigation, but medium ratings concerning the caused distraction in the OR [25].

The upcoming *HoloLens* by *Microsoft*, USA, is also an optical-see-through HMD. Opposed to past and current embodiments of HMDs, it integrates a miniaturized computer as well as multiple sensors that allow for both inside-out-tracking of the HoloLens pose and user interactions based on gestures and gaze tracking [120]. This enables spatially stable AR overlays without the need for tethering to external devices for either tracking or calculation of the AR overlays and could counter some of the drawbacks listed by Wen and Eggers et al.. Medical companies such as *Stryker*, USA, are partners for HoloLens pilot projects. However, a drawback is the reportedly low FoV of the augmentations which is estimated to be about $30° \times 17°$. At the time of this thesis, neither the exact FoV nor the existence or nature of upcoming medical applications was confirmed.

In a recent review, Mitrasinovic et al. analyse the use of smart glasses in healthcare based on 241 selected articles. They see great further potential, but also give several current limitations such as *"(...) effects of divided attention and cognitive tunneling, and spatial disorientation (...)"*. They voice concerns about the *"(...) risk that the use of augmented reality may draw focus away from the operating field (...)"* [123].

Handheld Displays Rodas et al. present an intraoperative system for analysing the scattered radiation by a robotized X-ray imaging device. The result is provided as a reconstructed 3D scene view, augmented by the illustration of the radiation per person, with the aim to make personnel aware of the radiation. To facilitate the usage in the OR, they propose a tablet with markerless inside-out tracking for visualization [107, 157].

Müller et al. propose using a tablet for intraoperative augmentation of preoperatively acquired patient anatomy in percutaneous interventions. Colored markers are attached to the skin of the patient that are captured by the camera of an iPad. The camera stream of the tablet is relayed via wifi to a stationary computer that performs inside-out tracking of the tablet position relative to the patient, calculates the augmentation overlay and streams the resulting images back to the tablet. The approach was evaluated successfully with the drawback of a lack of depth information of the augmentation [128].

SAR In 2001, Wörn and Hoppe were among the first to present a concept for a surgical spatial AR system, consisting of a projector and two cameras, that targets the transfer of surgical planning data into the OR. The patient is initially registered using structured light, without the need for artificial screw markers, and then

tracked using optical markers [194]. Marmulla et al. further develop the system and use it in maxillofacial interventions to augment the face of a patient with geometric planning data such as osteotomy lines as depicted in Figure 2.9. As the concept does not require tracked instruments, the surgeon can use all available surgical instruments to cut along the projected trajectory [111]. In a clinical study with 10 patients, Krempien et al. demonstrate the applicability of such a system to interstitial brachytherapy with a reported projection error of about $1\,\mathrm{mm}$ [96].

Tardif et al. present a concept and realization of a calibration-free projector-based augmented reality system. Using structured light projection, they create a mapping of pixel-to-pixel correspondences between the projector and a camera that designates the Point of View (POV) of the surgeon. Using this mapping, they pre-warp the projected content such that its projection on the patient appears undistorted from the surgeon's POV [172]. As this concept is based on a static POV of the surgeon and cannot project spatially dependent information (e.g. anatomy of the patient) due to lack of registration, it is not applicable to a real surgical scenario.

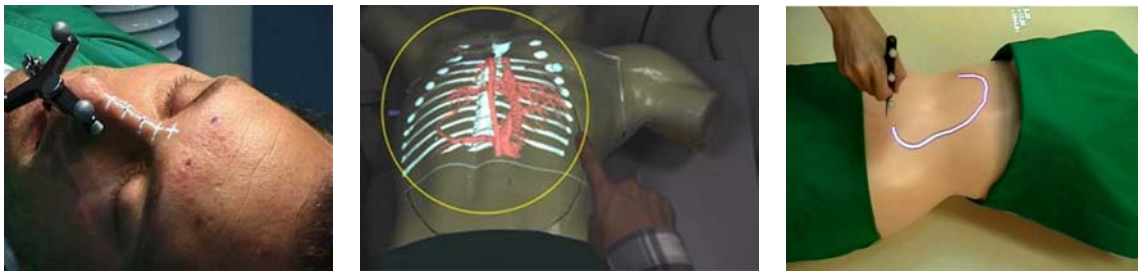


Figure 2.9.: Examples of different applications of medical spatial augmented reality. *Left:* Projection of preoperatively planned osteotomy lines onto patient [111]; *center:* visualization of anatomical features [192]; *right:* live annotation using a tracked probe [164].

Seo et al. propose a system concept that combines a projector with a camera and a tracking system. After acquiring the patient's surface using structured-light, they offer the surgeon the possibility to draw markings on the patient's skin using not an ink pen, but a tracked probe whose tip trajectory is projected onto the patient [164] as depicted in Figure 2.9.

Wen et al. also employ a direct augmentation system, consisting of a projector and a stereo camera system, to enable direct projection of anatomical features on the patient's body and visualize a robot-assisted needle-insertion process [191]. In later works, they add a Kinect to allow for gesture-based interaction with the projection [192] and gesture-based control of the needle-insertion robot. They demonstrate successful control of the robot both in manual and in semi-automatic mode and report clinically acceptable insertion errors of below $2\,\mathrm{mm}$ [193]. No details are given as to how the robot's position was determined in relation to the patient and camera/projector system.

Kocev et al. present a direct augmentation system that consists of a projector and a Kinect and is targeted at touch-free interaction with imaging data. The Kinect is used for detecting gesture input by a surgeon to control the projection of 3D anatomical features on arbitrary surfaces. No registration of the projector to the scene and/or the patient is performed [91].

Gavaghan et al. design a marker-tracked, handheld device that includes a projector for direct spatial augmentation onto the liver in a tumor ablation procedure. Due to the direct projection of anatomical structures and needle guidance information, they report a positive evaluation, citing that the system creates an "immersive and intuitive scene" [39, 40].

Zhang et al. present an optical-see-through SAR system and demonstrate its applicability in a brain phantom experiment in which data acquired in a CT scan is overlaid over the phantom. The augmentation is however only visible for one person at a time as a fix point of view is required [198].

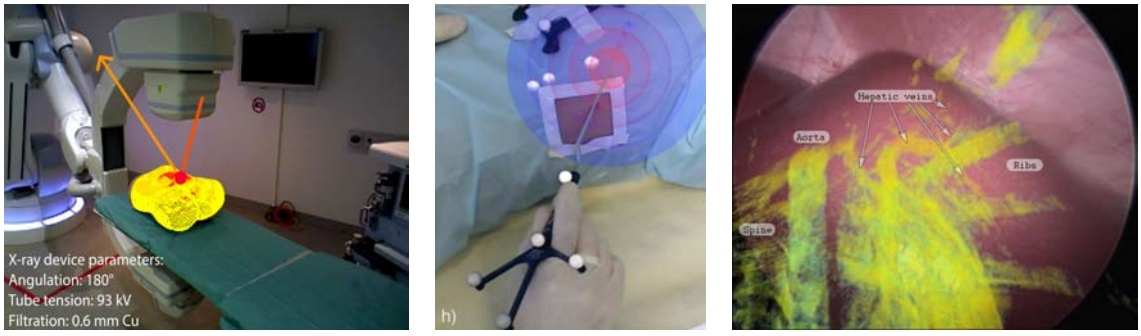


Figure 2.10.: Examples of different applications of medical mixed reality. *Left:* Tablet-based intraoperative visualization of patient exposure to X-ray [157]; *center:* 2D screen-based visual feedback of depth information in needle biopsy [15]; *right:* screen-based overlay of anatomical features onto video stream from laparoscopic camera [34].

Screen-based Mixed Reality One of the major use cases for computer-generated overlays in surgery is the scenario of Minimally Invasive Surgery (MIS), where the surgeon can see the intervention field inside the body only through the video stream taken by the laparoscope. An extensive account of such works is given in the PhD thesis of Feuerstein [34]. While many publications refer to this modality as AR, using the classification by Milgram introduced in section 2.1.6 it is actually a class 1 MR display. In the following, a selection of of recent works that employ Screen-based Mixed Reality in a surgical scenario is given. Examples of the resulting graphical representations are depicted in Figure 2.10.

Bork et al. study the effects of using an AR overlay on a video stream in combination with auditory features to guide users in the task of a simulated needle

biopsy procedure. They report a high improvement in accuracy, albeit with a huge increase of completion time for the task [15].

Habert et al. physically attach an Asus X-tion Pro Live camera to a C-arm so that the camera views the patient through a mirror. Using a manual calibration procedure, it is ensured that the camera's FoV and center are aligned to those of the X-ray source. X-ray images acquired by the C-arm can then be mapped onto the 3D surface of the patient as captured by the camera, thereby offering the surgeon an augmented view of the patient [49].

Navab et al. present a freehand *Single-photon emission computed tomography (SPECT)* system in which a tracking system and a 2D camera both view the situs. A marker-tracked handheld radiation counter is used to acquire data for a 3D reconstruction of the target anatomy. For visualization, the reconstruction is overlaid onto the patient and visualized in an AR view. The system has been commercialized and has received FDA approval [131].

Lamata et al. give an overview of various projects on medical AR, including the process applied and results obtained by the European research project ARIS*ER [101].

Dixon et al. investigate the effect of AR surgical navigation, in the form of contour augmentations on the endoscopic camera image, on attention, efficiency and accuracy of the surgeon. In cadaveric exercises with 50 participants, they report that the group that had to look up the contours on an external monitor had a significantly better recognition rate of a foreign body as compared to the group that used the augmented display, whereas task completion time and accuracy remained stable between both groups [27].

On a side note, Barad notes that while most AR devices include some kind of voice recognition-based control, this has several undesirable consequences in practice, due to general noise in the OR and the lack of fine control achievable. In the experience of Barad, these limitations result in rare intraoperative usage of this modality [4].

2.3.1.3. Robot usage in the Operating Room

Side effects of robot usage in the OR Lai et al. [100] study the effects of MIRS on the roles and communication of the OR team. Robot-assisted surgery is usually performed with teams of five persons: Two nurses, an anesthesiologist, the operating surgeon and an assisting surgeon. Almost all of these roles in the OR are affected as their tasks in MIRS are extended in comparison to traditional laparoscopic interventions:

- *Surgeons* are placed farther away from the situs, outside the sterile field, while conducting the intervention. They are therefore more dependent on the OR team to communicate information about the status of robot and patient. Being physically separated from the sterile field has even been reported to lead to a sense of isolation from the patient.

Figure 2.11.: Team positions during an intervention performed with the da Vinci™ surgical robot system. *Left:* The OR personnel works in close proximity to the patient (shown during docking of the robot); *right:* The surgeon is located outside the sterile zone, physically distant from the patient.

- *Anesthesiologists* need to check that while the robot is docked to the patient, the robot does not harm the patient, e.g. by knocking out the airway or hitting the patient.
- *Nurses* face a more complicated task as they have to conduct the changes of the robotic tools, which need to be coordinated with the surgeon. In interviews with stakeholders in MIRS, Lai et al. noted that nurses "must spend energy and attention attending to the needs of the robotic system, which can divert their attention from the patient." [100]

Both Lai et al. and Randell et al. emphasize that, due to the changes in the OR that result from MIRS, especially the remote location of the surgeon as shown in Figure 2.11, it is important that the whole team has access to the same information and a "shared situation awareness" [100, 152].

Concerning the impact of robotic surgery, it is further noted by Healey et al. that the surgeon faces considerable demands to "select and filter information from noise whilst attending to multiple concurrent tasks" [56]. This underlines the need for non-intrusive modalities to convey information to the surgeon in a way that does not further increase the high cognitive load.

Non-clinical performance characteristics At a recent FDA workshop on challenges and opportunities of RASDs [185], Taylor put forward several technical characteristics, including the following:

- Safety-related considerations:
  - The system shall not make unintended motions. This has to be guaranteed by redundant means to detect failure.

  - The system should stop or pause in case of detection of an error or unexpected sensor event. The remaining robot motion after receiving the stop command has to be limited to a small, application-specific amount, which in turn limits the maximum actuator speeds.
  - Unintended contact with either patient, OR personnel or OR equipment has to be avoided, especially with parts of the robot not directly visible to the surgeon.
- Human factors and interfaces: The surgeon needs to interact with information infrastructure without disengaging from the patient.

Looking into the future, Taylor postulates that "Surgical robots will be only one element in an increasingly information-intensive environment. In many respects, the "robot" may more appropriately be thought of as the room." [173].

2.3.2. Safety in human-robot interaction

2.3.2.1. General concepts

To allow for shared workspaces between humans and robots, the robot either has to be constructed in a compliant way so that it cannot physically harm the human, even in case of a collision, or some sort of sensing capability is required that allows to apprehend and react to potentially unsafe situations.

The former concept, which is employed in several new collaborative industrial robots under the term *soft robotics*, inherently allows for the robot's end effector to deviate from its planned position, e.g. if the robot arm is pushed away by a human in case of a collision. Especially in the scenario of MIRS, the concept of soft robotics is not applicable as each instrument motion needs to be performed in accordance with a remote center of motion (the trocar point) and deviations of the instrument tip poses from those intended by the surgeon can lead to injuries. In addition, in soft robotics human-robot collaboration scenarios, special diligence has to be paid to the the objects handled by the robot in order to eliminate the possibility of injuries caused e.g. by accidental contact between human and handled sharp objects such as medical instruments.

For sensor-based safety concepts, different kinds of sensors are available. A broad categorization of robot's sensors is the distinction between *proprioceptive* sensors, which are used to perceive the internal state of the robot, e.g. encoder values and joint angles, and *exteroceptive* sensors that offer information about the external world, such as distances to objects, sound or light levels [116]. Sensors can be further divided into *contact* and *non-contact* sensing. In addition, external sensors that offer information about the environment, such as 3D cameras, will also be classified as exteroceptive sensors in the following.

The following sections give an overview of relevant works on safe human-robot interaction, based on different kinds of sensor concepts.

2.3.2.2. Integrated and attached sensors

The latest generation of light-weight robots developed by DLR, the *LWR 4* and the MiroSurge arms, feature integrated torque-sensors in every joint. The force and torque that are measured at each joint are the combination of forces based on the robot's structure and motion, including static forces like gravity and dynamic forces caused e.g. by acceleration, and potential external forces that are applied to the robot. By calculating the current gravity-induced forces based on the known pose of the robot, external forces can be calculated per joint. This allows for contact-based proprioceptive sensing, i.e. collisions between robot and environment can be detected as soon as the contact force is detectable by the robot [24]. However, due to the contact-based sensing principle, it is not possible to preempt collisions as they cannot be detected prior to the time of their impact.

Escaida et al. study the potential of capacitive tactile proximity sensors that might be fitted to robot arms, similar to the concept of an artificial skin. As the sensors offer detection of both touch and proximity of objects, they enable preemptive braking or trajectory re-planning of the robot before a collision occurs. Early experiments show the feasibility of using such sensors for tracking objects close to the sensor matrix, even under occlusions, and preemptive collision avoidance [132].

In the PhD thesis of Ostermann, a norm-compliant sensing concept is developed based on multiple small ultra-sound sensors that cover the whole surface of one or multiple segments of an industrial robot. The realized prototype offers redundant sensing due to low inter-sensor distances and achieves a latency of 30 ms [140].

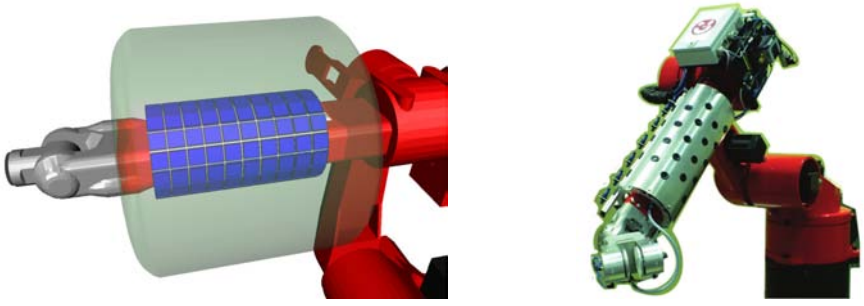


Figure 2.12.: Examples of attached sensors for distance-sensing in the context of safe human-robot collaboration. *Left:* Concept rendering of matrix of capacitive tactile proximity sensors and their sensing range [132]; *right:* complete realization of norm-compliant, ultra-sound-based range sensing [140].

The key benefits of non-contact-based sensing concepts such as both aforementioned systems are that they allow for a reasonably high spatial resolution and completely avoid occlusions, as the measurements originate at the robot itself. However, the sensors have to be either integrated into the robot during the development phase or added at a later time, which requires extensive hardware modifications in order to attach both the sensors and the necessary power and data cables (see Figure 2.12). In both cases, the feasibility of fitting a robot with such a sensor matrix heavily depends on the complexity of the surface geometry of the robot and has to be mechanically adapted for each type of robot. Furthermore, sensor integration increases the volume of the robot arms, which contradicts the surgeons' wish for smaller robot arms, as voiced e.g. in a survey of the FDA Medical Product Safety Network [33]. Independent of these considerations, it is unclear how either sensing modality would be affected by the sterile drapes in which robot arms are covered during surgical interventions.

2.3.2.3. External optical sensing

2D camera based approaches The *SafetyEYE* by *Pilz*, Germany, is a norm-compliant, commercially available safety system for industrial applications based on optical workspace supervision. The sensing component consists of a housing with three integrated 2D cameras which is ceiling-mounted to offer a top-down view of the robot cell. The acquired 3D data is continuously checked for violation of predefined, static safety zones. Depending on the zone in which such a violation is detected, different reactions can be executed, such as limiting the movement speed or emergency-stopping the robot [149].
Due to the top-down view of the scene, a high mounting point of the camera is required to achieve a sufficiently large observable area as depicted in Figure 2.13. Also, close human-robot interactions cannot be sensed accurately.

The PhD thesis of Ladikos focuses on multi-view 3D reconstructions for interventional environments. Using a configuration of 16 synchronized 2D cameras, they first perform histogram-based background subtraction for determining objects of interest and then reconstruct the 3D shape of these objects using the Visual Hull approach. The safety aspect of the system is evaluated in the context of using a C-arm in a laboratory setting with the applications of collision avoidance and radiation monitoring specific to human body parts.
Ladikos concludes that the transfer of the system to a real intervention room faces multiple challenges, especially in relation to the background segmentation which is not robust enough for changing backgrounds, homogeneously-colored environments and changing lighting conditions. The usage of ToF cameras is proposed to overcome several of these limitations [99].

Stengel et al. similarly employ a five camera system for reconstructing the shape of a human freely moving within a robotic work cell. A Single Gaussian-based background model is learned offline and used for online detection of foreground objects, taking into account dynamic occlusions by the robot. Based on the foreground

classification by all cameras, a voxel-based scene reconstruction is performed [166]. As this work is targeted at industrial applications where the environment can be strictly controlled, it assumes a static background where humans and robots are the only moving entities.

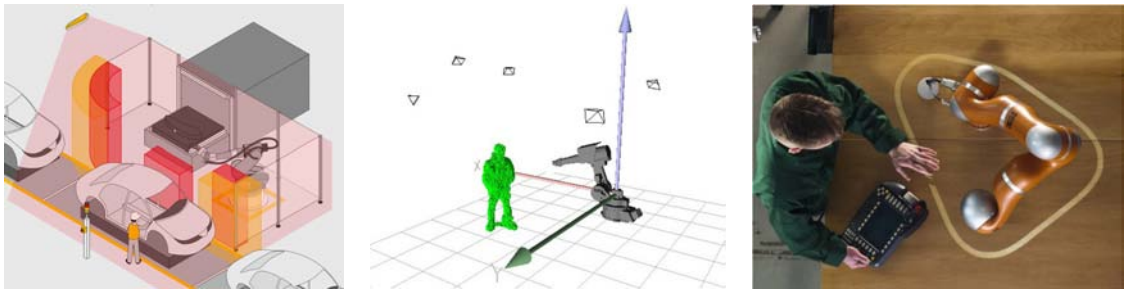


Figure 2.13.: Examples of (multi) 2D camera based concepts for safe human-robot collaboration. *Left:* SafetyEYE surveillance concept with color-coded safety zones [149]; *center:* Virtual scene with voxel-based human reconstruction and virtual robot [166]; *right:* Projection-based safety space around an industrial robot [189].

Tan and Arai focus on human-robot collaboration in a cellular manufacturing setting where a shared workspace exists on a workbench which is located between human and robot. In addition to standard industrial safety measures such as light fences, the human's upper limbs are tracked based on color information applied to the clothing, using a 2D stereo vision system. While no further details are given as to how the human posture is taken into account for safety considerations, it is hinted that a rule-based system is used [170].
In later works, the human tracking is extended to a three camera configuration, evaluated as three independent stereo vision systems, to prevent occlusions. As the extension focuses on the camera system and tracking accuracy, no further information about the safety aspect is given [171].

While also researching safe human-robot interaction using 2D cameras, Vogel et al. employ a radically different approach compared to the aforementioned works. Instead of reconstructing 3D information from 2D images, they aim to reduce the problem of detecting potential collisions into 2D space. They calculate a top-down 2D view of a safety space around the robot based on its current pose and a simplified geometrical model of the robot. Using a ceiling-mounted projector, they project this view into the scene and segment the resulting projection in each of four ceiling-mounted cameras. The real projection is then checked against an ideal virtual projection, taking into account occlusions by the robot itself. Objects within the safety space result in differences between real and virtual projection and can therefore be detected [188, 189].
This approach is very effective in terms of computational cost and intuitively understandable for users, however it is not applicable to crowded and changing unknown environments as they influence the camera's view of the projection.

3D camera based approaches Graf et al. present a first approach to safe interaction between a human and an industrial robot based on a single, ceiling-mounted ToF camera. Based on the human posture, which is estimated from the acquired depth data, potential collisions with a known robot trajectory are detected and prevented by re-planning the robot's trajectory [43].
Continuing these works, Dittrich et al. extend the original sensing setup with a Kinect v1 for full skeleton tracking of the interacting human. They also introduce fuzzy logic for estimating the risk of a situation, based e.g. on the view direction of the human, as well as a motion planner that takes the estimated risk into account. Further, different command and translation gestures are implemented that allow for a gesture-based control of the robot [26].

Rybski et al. combine stereo vision with 3D cameras to supervise a scene in an industrial work cell. Similar to the works of Stengel et al., they first estimate a background model of the empty scene and then classify all moving foreground objects in the scene as persons, represented by a 3D occupancy grid. By surrounding the robot with a *danger zone* as well as a larger *warning zone* and similarly surrounding persons with a *safety zone*, intersections between the different zones can be checked for potential collisions [160].
Due to the initialization on an empty scene and the assumption that the scene does not change, this approach is also not feasible for crowded or changing scenes such as in the OR.

Schmidt et al. present an approach to online trajectory adaptation based on detected humans. While using two Kinect v1, they employ a custom detection of moving persons that consists of background subtraction, again from a previously learned background model, and filtering of the acquired foreground objects. To efficiently calculate the distance between the point clouds of detected persons and the virtual robot model, they create a voxel grid and calculate distances to the center of each occupied voxel only. To avoid collisions, the trajectory of the robot is adapted when the distance of the end-effector to the closest human would decrease beneath some predefined threshold [162].
This is only feasible if the robot's task consists of reaching discrete positions with its end-effector, e.g. in pick and place tasks, without restrictions on the behaviour in between these goals.

Morato et al. present a multi-camera system for safe human-robot collaboration in a simulated assembly task. They fuse the human skeleton tracking from four Kinect v1 using a Kalman filter and abstract the human body with spheres of different sizes virtually attached to the detected joint positions. Assuming a known trajectory of the robot, they propose a simple bi-modal control strategy where robot motion is stopped if the detected human comes within a certain distance to the planned trajectory [126].

Reardon et al. present a different approach, targeting a healthcare scenario in which both mobile and stationary robots are envisioned to cooperate in order to transport and sterilize non-sterile instruments after an intervention. In their experiments, a humanoid soft robot, the *Baxter* by *Rethink Robotics*, USA, is used to both supervise

Figure 2.14.: Examples of (multi) 3D camera based concepts for safe human-robot collaboration. *Left:* Different risk estimation for robot poses based on viewing direction of user [26]; *center:* Approximation of user with spheres based on multi-camera skeleton tracking [126]; *right:* Intersection of danger zone around robot and safety zone around user based on occupancy grids [160].

the scene and give visual feedback to users. Supervision is performed using a Kinect v1 mounted on the Baxter's head, allowing for dynamic rotation of the Kinect towards the ROI. Using the distance between the user and the mobile robot as a safety metric, the Baxter signals the safety state to the user by displaying different "faces" and raising its arms in case of safety-critical situations [155].

2.4. Summary and open research questions

Due the holistic approach of this thesis, the proposed system addresses several topics that have been researched in both the surgical and the industrial domain. These are listed in the following with a short description on the respective research gap related to safe and intuitive usage in a surgical scenario.

Apart from the GestoNurse [82], which is not a surgical robot system in a narrower sense, current and upcoming surgical robot systems do not integrate any sensing modalities that allow for contact-less perception of their environment. For general robotics, different approaches of equipping robots with exteroceptive sensors have been proposed [132, 140]. However, especially in the medical context, it is not feasible to retroactively apply such sensing concepts to existing robot systems due to their requirements in terms of space, wiring and control. Further, it is unclear if and how these kinds of sensors would be affected by sterile draping in which robots are covered during interventions.

In industrial scenarios, the best practice for the usage of robots has long been a complete spatial separation between humans and robots, enforced by fencing. This is slowly being replaced by optical supervision systems that establish static safety zones in lieu of fencing [149]. Common research approaches for human-robot collaboration in a shared workspace focus on dynamic adaptation of the trajectory of the robot [26, 43, 162]. However, this is not applicable for robot-assisted surgery

as it requires both a priori knowledge about the trajectory of the robot and the possibility of altering the trajectory of the robot to avoid collisions.

Further, many approaches to safety in human-robot collaboration assume a static, known environment that does not change apart from the motions of humans and the robot. This is often employed for background segmentation as a principal step in the algorithmic pipeline, for example with 2D cameras in the medical [99] and industrial scenario [166], but also with 3D cameras [162, 160] in an industrial scenario. It is especially important to note that all works assume a static, known position of the robot in the scene which needs to be established beforehand. This is not the case for surgical interventions, where the positioning of the surgical robot varies with each patient.

The usage of 3D cameras in the OR has primarily been researched regarding applications to touch free interaction by gesture based control [137] and 3D laparoscopy [92, 115, 145]. Recent works on monitoring of the OR focus on extraction and interpretation of human poses, but do not take the surgical robot itself and the non-human environment into account [9].

Works on applications of various forms of augmented reality to a surgical setting have focused on medical applications. A prime example is the augmentation of a view of the patient with anatomical features, which has been realized using different modalities such as endoscopic video streams as mixed-reality screen based AR [27, 34], handheld displays as see-through AR [128] and projection-based approaches as SAR. The latter are usually standalone approaches that do not integrate a surgical robot system [91, 111, 172, 194]. Employing AR for visualizing the state of the robot and giving feedback to the OR team has not been investigated yet.

To summarize, there are clear gaps in research concerning the monitoring of surgical robots for safety purposes:

- No concepts exist for intraoperatively detecting and verifying the robot's position.
- The effects of sterile draping on scene perception by range imaging cameras have not been investigated.
- Industrial approaches to safe human-robot collaboration by altering the robot's trajectory are not applicable to teleoperated surgical robot systems, where no trajectory is known prior to the robot's motion.
- Approaches based on background segmentation and/or known environments are not applicable to the surgical scenario.

3. System Concept

This chapter gives an overview of the proposed system, both in terms of concept and algorithms. First, the perception channel is presented by deriving the clinical prerequisites and technical design criteria of the multi-camera supervision system and establishing the two-level scene model. The algorithm for forward propagation of semantic labelling is then introduced which allows to transfer semantic information between both levels of the scene model. After these steps, all necessary information for safety considerations is available.

The safety concept is discussed next, starting with the Shape Cropping algorithm for spatial segmentation of the perceived scene into different zones around the robot. Based on Shape Cropping, different safety measures are derived that target the setup of the robot system as well as safety during online usage.

In the last part of this chapter, requirements for a feedback channel that relays information back to the OR personnel are put forward. The concept for such a visual feedback channel based on SAR is introduced together with a secondary usage of the projector for registration of the 3D cameras of the supervision system.

3.1. Operating Room Monitoring

For allowing a safe and intuitive usage of a robot system, a modality for perceiving the environment of the robot is required. While such a modality can employ proprioceptive and/or exteroceptive sensors, this thesis focuses on exteroceptive sensors, i.e. 3D cameras, due to the drawbacks of proprioceptive sensors for usage in the OR (see section 2.3.2.2).

3.1.1. Prerequisites

3.1.1.1. Clinical prerequisites

In the OR, there are many challenges for using 3D cameras in the context of safe human-robot interaction. The ever-changing positions of the OR personnel during and between interventions lead to dynamic occlusions around the OR table and the robot system. Ceiling-mounted medical equipment, such as OR lamps or monitors which tend to be moved occasionally during interventions, presents another source for occlusions, albeit from a different angle. Therefore,

a supervision system needs to be robust against occlusions and cannot rely on a single point of view of the clinical scene.

In order to keep the OR free from contamination, it is kept under slightly positive pressure during interventions and the air is recommended to be completely exchanged between 15 and 20 times per hour [109]. This is achieved by integrating a ventilation system into the ceiling, located over the sterile zone. Many such ventilation systems produce a laminar air flow, i.e. air streams out in a homogeneous flow without turbulences, which has been shown to reduce the risk for Surgical Site Infections (SSIs) for certain types of interventions, e.g. orthopaedic surgery [16]. To not interfere with laminar air flow, no devices that cause air turbulence, i.e. due to an integrated fan, may be mounted above the OR table.

Another important aspect of bringing additional technical systems into the OR is their requirements in terms of space. For decades, it has been reported that the OR becomes increasingly crowded, e.g. due to the introduction of laparoscopic surgery and its needs for additional equipment compared to open surgery [3]. This trend continued with the advent of surgical robot systems, which also require a significant amount of space and have been reported to lead to additional long travel distances in the OR [2]. Any additional technical systems that do not require access to the patient therefore need to have the smallest footprint possible in the direct vicinity of the OR table, without requiring additional floor space.

3.1.1.2. Technical design criteria

The technical design criteria that have to be taken into account by the supervision system stem from different aspects: oclinical prerequisites, safety requirements and future-proofing.

The clinical prerequisites described above necessitate realizing the supervision system as a multi-sensor setup to reliably monitor the ROI, even if the line-of-sight of one or more sensors is blocked. Due to the requirements of not disturbing laminar air flow and occupying a minimum amount of space close to the situs, it is not possible or desirable to place any computers into the OR. Instead, the sensing devices, i.e. 3D cameras, have to be spatially separated from the processing devices.

From a safety-oriented perspective, the system needs to be redundant on all layers in order to cope with possible problems caused by the hardware itself, e.g. malfunctions of single sensors, by the infrastructure, e.g. networking components, or by circumstances which cause a whole group of sensors to malfunction.

Furthermore, the system requires a high modularity in order to be device-agnostic, i.e. not tied to a specific set of components, and therefore being extendable with new sensors and components.

3.1.2. Supervision system

Taking into account the technical and clinical constraints as well as the characteristics of sensors as described in section 2.2.2, this thesis proposes a modular supervision system, consisting of multiple 3D cameras, to perform real-time monitoring of the OR environment around the surgical robot system. The proposed supervision system consists of two independent subsystems: The first subsystem is based on industrial-grade PMD cameras, the second subsystem is based on Kinect v1 cameras.

PMD cameras and Kinect v1 cameras employ different range imaging methods, namely ToF and structured light (see section 2.2.2). Combining two subsystems with different sensing principles to one supervision system offers several key advantages:

- Both types of cameras display diametrically opposing strengths and weaknesses that complement each other. Kinect cameras have the drawback of being "blackbox"-devices that do not offer external control, configuration or triggering, but offer a high lateral resolution and basic semantic interpretation, i.e. human tracking. PMD cameras can be fully configured, but offer only a low lateral resolution, which increases the difficulty of semantic scene interpretation.
- By combining two subsystems with different sensing principles, the robustness of the complete system against sensor-specific measurement deficits is increased.
- ToF cameras heavily suffer from interferences which limits the number of cameras that may be used in the same volume (see section 2.2.2.3). Extensive tests conducted in this thesis have shown that Kinect cameras neither suffer from nor cause interferences with ToF cameras (see section 5.1.1), which enables the use of additional cameras in the same volume.

The cameras of the supervision system are placed close to the operating table, just over head height, to perceive the direct environment of the surgical robot system and minimize the impact of occlusions caused by persons, e.g. the scrub nurse, standing close to the situs. Figure 3.1 illustrates the spatial layout of the supervision system.

3.1.3. Two-level scene model

The low level, geometrical information gathered by the various 3D cameras of the supervision system needs to be organized and interpreted on a higher level. For this, a two-level scene model is proposed where each level corresponds to one of the subsystems of the supervision system. On both levels, all information acquired by the according camera subsystem is fused into a common scene representation.

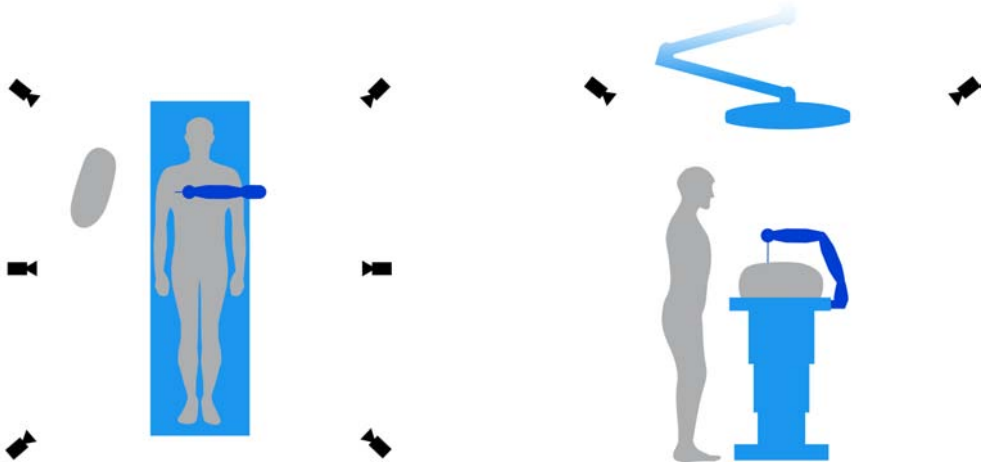


Figure 3.1.: Illustration of the spatial layout of the proposed supervision system. *Left:* Top view with six camera positions located sideways of the operating table; *right:* Side view depicting the height of the camera positions.

The first level consists of purely geometrical information without semantic interpretation and serves as the basis for all safety-related algorithms. On this level, all data acquired by the PMD camera subsystem is integrated to establish a scene model with a lower latency, albeit at a low spatial resolution. As PMD cameras are designed for industrial use, there is full control of the data acquisition process, including synchronization, integration time and modulation frequency per camera.

On the second level, geometrical information from the Kinect v1 subsystem is semantically interpreted, e.g. by segmentation of persons or OR devices in the scene. As described above, the Kinect cameras offer no external control over the data acquisition process, but deliver a high resolution point cloud with color information. Due to this lack of control and the higher latency compared to the first level, the second level cannot be used for safety-critical algorithms, but for semantic scene interpretation.

3.1.4. Forward propagation of semantic labelling

It is desirable to combine the different advantages of both levels of the scene model, namely low latency on the first level and semantic information on the second level, to enable a low latency scene representation with semantic information.

To achieve this, a general algorithm is proposed for forward propagation of semantic labelling between two independent data sources with different timing characteristics, such as latency and frame rate, that only requires a known mapping between each data source. The algorithm consists of two parallel processing pipelines: One pipeline processes all data from the faster, semantic-free data

source by calculating the optical flow and providing a tracking estimate in every step based on the last tracking information available. The second pipeline processes information from the slower, semantically enriched data source to update a background model of the scene and inject new tracking information into the first pipeline.

As both levels of the scene model contain geometrical information in the form of point clouds and there is a known spatial relation between each pair of cameras, this algorithm can be applied to the scene model to forward calculate the semantic annotations from the second level to the first level. This results in a low-latency scene representation which is annotated by e.g. human tracking information.

3.2. Safety concept

Based on the supervision system presented in section 3.1.2, this section proposes different safety features with the overall goal of establishing a safe, redundant monitoring capability for surgical robot systems in the OR. Each proposed safety feature addresses one of the non-clinical performance characteristics put forward by Taylor (see section 2.3.1.3).

3.2.1. Differences from industrial settings

In industrial applications, the exact task that a robot has to perform is known beforehand and human-robot interaction usually only involves one or more operator(s) who are acquainted with the specific scenario. The manipulated object is inanimate and the manipulation task is mostly either not time critical or, if time critical, a failure caused by a delay does not result in human injuries. Therefore, it is viable to prevent collisions between humans and robots by altering the trajectory of the robot. Furthermore, the layout of robot cells as well as human-robot collaboration workspaces can be optimized in advance to allow for optimal placement of sensors. For a given task, e.g. assembly, the position of non-mobile robots is fixed, so the spatial relation between safety sensors and the robot only needs to be calibrated once and can be assumed to be static afterwards.

The clinical scenario of an OR, however, has many differences compared to an industrial setting. It is usually crowded by no less than five people who need to focus on their primary task, the medical care of the patient. The position of the patient and surgical robot system varies between each intervention, so no fixed relation between robot and sensors can be assumed and it is not possible to optimize the placements of sensors to a specific task. In the case of MIRS, the robot is telemanipulated and is in continuous contact with the patient, i.e. it has to strictly adhere to the position of the trocar points. For these reasons, avoiding external collisions between OR personnel and robot by altering the robot's trajectory is not feasible in MIRS scenarios.

3.2.2. Robot safety zone

A safety zone is established around each arm of the surgical robot system to ensure that potential collisions between the robot and its surroundings can be detected before they occur, thereby allowing for initiating an appropriate reaction. Each safety zone is dynamically updated in real time in order to match the current pose of the corresponding robot arm, based on the robot's joint angles and an internal kinematic model. While the safety zone is clear of any obstacles, there is no danger of collision with the OR personnel, the patient or inanimate objects in the scene. If the safety zone is violated, the robot is in close proximity to a nearby object which can necessitate an according reaction. Figure 3.2 illustrates the safety zone around a robot arm and the detection of violations by OR personnel or environment.

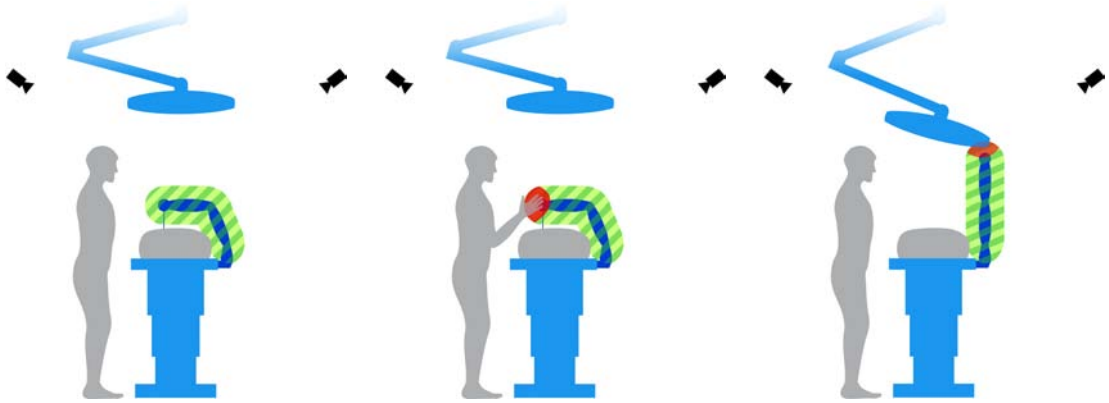


Figure 3.2.: Illustration of the robot safety zone in different situations. *Left:* Clear safety zone without obstacles; *center:* Violation of safety zone by person; *right:* Violation of safety zone by object, e.g. OR lamp.

In contrast to multi-zone approaches as presented e.g. by Rybski et al. [160], where intersections are calculated only between detected users and the robot, the proposed single safety zone approach takes into account the whole environment. As a result, even if semantic information is missing or incorrect, i.e. a person is not correctly classified and tracked, potential collisions are still detected.

3.2.3. Shape Cropping algorithm

To implement the safety zone approach, two requirements have to be met: The robot has to be removed from the virtual scene and the remaining scene needs to be segmented based on the safety zone boundaries. As the virtual scene is represented as a point cloud in the proposed scene model, it is desirable to use a geometric segmentation approach that can fulfill both requirements. This is achieved by the proposed *Shape Cropping* algorithm[1].

[1] There is a similarly named algorithm developed at the French National Institute for computer science and applied mathematics (INRIA) which is called "3D Shape Cropping" [38]. Contrary

Algorithm 1: Shape Cropping

input : Point cloud of scene, segments of inner and outer hulls as meshes, robot joint angles
output: Segmentation of scene into inlier, outlier and neutral

foreach *segment* $s \in$ (inner hull $\cup$ outer hull) **do**
  s $\leftarrow$ `fk_transform`(*s*, joint angles)
end
roi $\leftarrow$ `crop`(scene, `aabb`($\bigcup s, s \in$ outer hull))
foreach *segment* $s \in$ inner hull **do** inlier, roi $\leftarrow$ `crop`(roi, *s*);
foreach *segment* $s \in$ outer hull **do** outlier, neutral $\leftarrow$ `crop`(roi, *s*);
scene $\leftarrow$ scene $\cup$ neutral

The Shape Cropping algorithm requires two pre-computed meshes per segment of the robot which are both derived from the Computer Aided Design (CAD) model of the according segment and have been geometrically simplified. One mesh represents an *inner hull*, which is a volume that encompasses the according segment of the robot. Compared to the original CAD model, the mesh is slightly enlarged. The other mesh represents an *outer hull*, which corresponds to the safety zone as described above. Its enlargement as compared to the original segment model therefore depends on the desired margin of the safety zone.

In each iteration of the algorithm (see Algorithm 1 for formal description), all segments of the inner and outer hull are positioned according to the robot's current pose based on the joint values and the forward kinematics of the robot. A ROI is extracted from the point cloud that represents the virtual scene based on an Axis-Aligned Bounding Box (AABB) around all segments of the outer hull. The ROI is then spatially segmented into three separate classes as depicted in Figure 3.3:

- *Inlier*: Points within any segment of the inner hull, i.e. belonging to the robot.
- *Outlier*: Points within any segment of the outer hull, but outside of any inner hull segment, i.e. points that violate the safety zone.
- *Neutral*: Points outside of any hull segment.

When applied to a real world scenario with 3D measurements obtained by actual 3D cameras, this geometric segmentation often does not correspond to a semantically correct segmentation: Due to noise present in the point cloud representing the virtual scene, caused by incorrect data acquired by the 3D cameras, points are often classified wrongly. For example, points corresponding to the robot surface can be classified as outlier points, if the distance error of the measurement is large enough to place them outside of the inner hull (see Figure 3.3). This needs to be taken into account in all subsequent processing steps.

to the algorithm proposed in this work, where a known shape is segmented from a 3D scene, it aims to calculate a polyhedral bounding surface of an unknown object based on multi-perspective views.

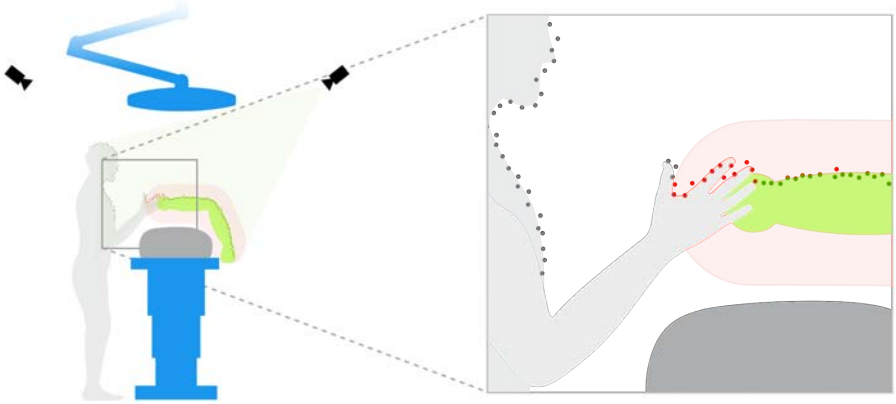


Figure 3.3.: Illustration of the Shape Cropping algorithm. The point cloud representing the scene, depicted as colored points on object boundaries observed by one overhead camera, is segmented into different classes based on the shape and pose of the robot: Inlier points inside the inner hull belong to the robot (green), outlier points in the outer hull violate the safety zone (red), neutral points outside either hull are distant parts of the scene.

3.2.4. Safety Features

3.2.4.1. Robot localization

For traditional laparoscopic interventions, the locations of the access ports, through which the instruments are inserted into the patient, are determined by anatomical constraints only: Following general medical guidelines, the exact position of each access port is determined by the OR personnel during an intervention. For MIRS, the mechanical design of the robot presents additional constraints: The trocars need to be placed in a way that prevents external collisions of the robotic arms, which often results in trocar positions that are different from the traditional port location [118]. While this can be performed somewhat intuitively by experienced personnel, an even bigger challenge is to optimize the trocar positions so that dexterity of the surgical instruments inside the patient is maximized. As the reachable workspace depends on both the configuration of the robot and the location of the pivot point relative to the robot, the correct planning of both the trocars and the position of the robot arms can be critical to guarantee full dexterity for the surgeon [65].

For these reasons, an intraoperative setup verification is proposed as a novel safety measure that detects the positions of the robot arms and checks them against the preoperative planning. As the proposed system has the goal to work with different types of robots in an OR setting, the detection of the robot arms cannot rely on color cues: Apart from potentially varying colors between different models and manufacturers, most current and upcoming surgical robot systems are white,

which is the same color as most other OR devices. Therefore, the robot arms need to be detected based on their coarse shapes. However, to guarantee sterility, robot arms are covered with sterile drape before they can be positioned close to the patient. This can influence the shape of the robot as perceived by the camera, thereby preventing the detection of the robot arms based on 3D keypoints as estimated by keypoint detection algorithms such as NARF or 3D SIFT.

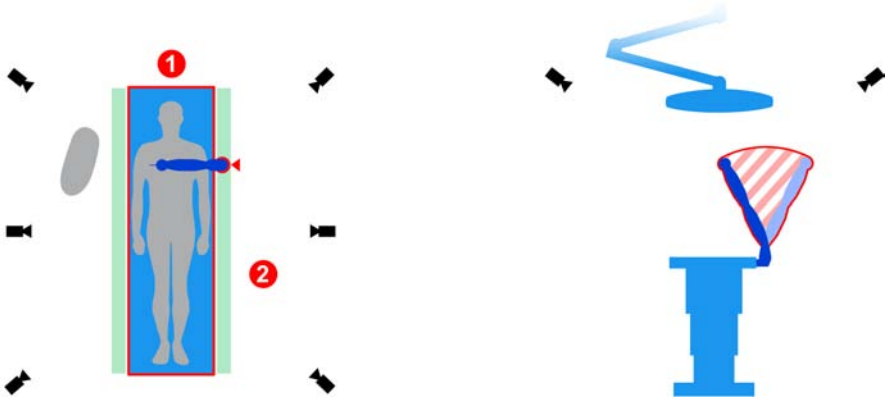


Figure 3.4.: Illustration of the robot localization. *Left:* With passive localization, first a landmark such as the OR table is detected and second the robot is localized in the smaller resulting search space; *right:* Active localization is based on spatial change detection of a motion performed by the robot.

In the proposed system, the localization of each robot arm is therefore performed in two steps as illustrated in Figure 3.4:

- *Initial localization*: The position of each robot arm is detected with a lower accuracy, using one of the following methods.
  - *Passive localization*: To minimize the search space, localization is performed based on known landmarks in the scene. For robot arms mounted to the OR table, the OR table itself can be used as such a landmark as it limits the search space to the space directly adjacent to the OR table rails.
  - *Active localization*: Each robot arm performs a pre-defined motion which is detected by thy supervision system using *spatial change detection*. Based on the performed motion and the shape of the detected volume, the position of the robot is calculated.
- *Localization optimization*: The detected position is refined to improve the accuracy. In this step, the Shape Cropping algorithm is applied to transform the localization task into an optimization problem. To evaluate the quality of a detected robot position, different measures are defined as optimization goals based on the number of inliers and outliers.

As the localization optimization step evaluates discrete locations, starting at the position originally detected by the initial localization, a heuristic is required for finding the correct position. First, the steepest ascent hill climbing algorithm is applied by sampling equally-spaced positions in a sphere around the initially located position. This is repeated with increasingly smaller spheres and therefore increasingly fine granularity of the detection. However, the criteria evaluated by the localization based on shape cropping can lead to systematic errors which have to be corrected specifically. When for example only a single camera is used, the shape cropping algorithm exhibits a tendency to detect the robot too close the camera, as this leads to a higher $\frac{inlier}{outlier}$ ratio, which is one of the optimization criteria. This is corrected by evaluating additional positions along the camera-robot axis with an increasing distance between robot and camera. Another such example is an erroneous initial detection which leads to classifying most of the robot as outliers. This can be remedied by an outlier-based correction step that shifts the estimated localization towards the direction in which most outliers are detected.

3.2.4.2. Collision handling

Detection of impending collisions Detection of impending collisions is the basis for any further situation-specific reaction, such as collision avoidance and/or warning the OR personnel. Detection of impending collisions is again based on shape cropping: The safety zone, represented by the volume between inner and outer hull, is continuously monitored for violations. A violation is detected if the number of outliers in the safety zone of one segment increases significantly over a short period of time. Based on euclidean clustering, the shape of the violating object and its position relative to the robot is then determined. This allows for both reacting to the impending collision, e.g. by stopping the robot motion, and communicating information about the potential collision and colliding object to the OR personnel.

Collision classification Contrary to the industrial setting, where avoiding any contact between human and robot is considered the gold standard for safe interaction, the surgical domain requires a more detailed consideration. For the purpose of this thesis, a *collision* shall be defined as *an event of establishing contact between two entities that were spatially separated before*. Therefore, the term collision will be used as a general term to refer to both voluntary and involuntary contacts without specifying involved forces or limits thereof. Concerning the OR domain, potential collisions with the robot can be classified as belonging to one of three main categories:

- *Collision with patient*: During an intervention, robot arms are not allowed to come into contact with a patient, except for the attached surgical instruments when used to perform the intervention. Involuntary contact can lead to harm

to the patient, ranging from bruising to serious harm if e.g. an airway is knocked out. It is usually the task of the anesthesiologist to ensure this (see section 2.3.1.3).

- *Collision with environment*: Collisions with the environment include all cases where a robot arm comes into contact with or pushes against any object other than a person in the OR, i.e. other than patient or OR personnel. While this does not pose a direct threat to either patient or OR personnel, it has adverse effects on the performance of the robot or on the object to which the robot applies pressure. Both cases can indirectly lead to harm to the patient, e.g. if the robot does not perform the correct motion or if the object abruptly moves under continued pressure from the robot arm.
- *Collision with OR personnel*: There are different situations during which the robot can come into contact with the OR personnel. *Active collisions* occur when the robot performs a motion that leads to contact with a person, whereas *passive collisions* are caused by a person coming into contact with a robot arm that is not moving. A further distinction needs to be made between voluntary contact, for which the person is aware of the impending collision or is actively causing it, and involuntary contact.

Rule-based collision prevention For the scope of this thesis, a rule-based strategy for collision handling has been implemented. As detailed above, collisions between a robot arm and the patient or the environment can lead to patient harm and therefore need to be prevented. If an impending collision between a person and a robot arm is detected, knowledge about the current situation may be required to enable an appropriate reaction.

As standard collision avoidance methods such as altering the trajectory of the robot are not applicable for robot assisted surgery, especially in case of MIRS, an emergency stop of the robot is used for collision avoidance within this thesis. This is in line with the proposal of Taylor that "[a surgical robot] system should stop or pause motion [...] on system detection of error or sensor event" [173].

Table 3.1 shows the resulting decision matrix for all combinations of the current robot state and collision class. The possible reactions are:

- *stop*: An emergency stop is triggered to prevent a collision.
- *situation-based*: Reaction of the impending collision is depending on the current situation.
- "-": No reaction will be performed, i.e. the collision is allowed to occur.

Impossible combinations are designated *n/a*. As situation-based reactions require online detection and reasoning about the surgical situation and its medical aspects, which is outside the scope of this work, they were simplified as follows: When a robot arm is in hands-on mode, collisions are allowed to occur, as the OR personnel is in full charge of the robot movements at this time. Active collisions with the OR

	passive collision		*active collision*
	non-moving	**hands-on**	**moving**
Patient	n/a	-	stop
Environment	n/a	-	stop
OR personnel	-	situation-based	situation-based

Table 3.1.: Decision matrix for rule-based collision prevention based on collision class and robot state.

personnel always lead to an emergency stop when the according rule is triggered. Therefore, the proposed decision matrix is in line with the cited requirements of the respective ISO norms (see subsection 2.1.3). It adheres to clause 5.10.2, 5.10.3 and 5.10.4 of ISO 10218-1 and does not allow for quasi-static contacts to happen between robot and either OR personnel or patient.

3.2.4.3. Continuous pose supervision

Especially for medical applications, potential malfunctions of robot systems can have severe consequences. In MIRS, the surgeon directly controls all motions of the robot and guides the instruments based on their relation to the target region in a way of visual servoing. As the surgeon will unconsciously compensate for slightly incorrect robot motions and no autonomous motions are performed by the robot, it might slip the attention of the OR personnel if the real position of the robot slightly differs from the desired one, e.g. due to an erroneous actuator in one of the joints. However, if the trocar constraint for the robot arm is implemented in software and not mechanically based on the robot's kinematic structure, the trocar constraint may not be complied with in case of malfunctioning or miscalibrated joints. This can result in undesired forces being exerted on the abdominal wall of the patient via the trocar ports.

Continuous supervision of the robot's pose by a redundant external system, i.e. the supervision system presented above, is proposed as a measure to mitigate this risk. Again, the surrounding of the robot is segmented into inliers and outliers using shape cropping. If the real pose of the robot deviates from the desired one, it also deviates from the shapes of the inner and outer hull as these are calculated based on the desired joint values. This results in violations of the safety zone, caused not by external influence, but by the robot itself. To be able to differentiate between both, the spatial distribution of the detected violation is analysed via euclidean clustering: If the cluster that causes the violation is connected to the robot cluster in the inner hull, the root cause for the detected violation is a dysfunction of the robot arm. If the violating cluster is not connected to the robot or belongs to an outside cluster, it is interpreted as an impeding collision as described above. Figure 3.5 illustrates these cases.

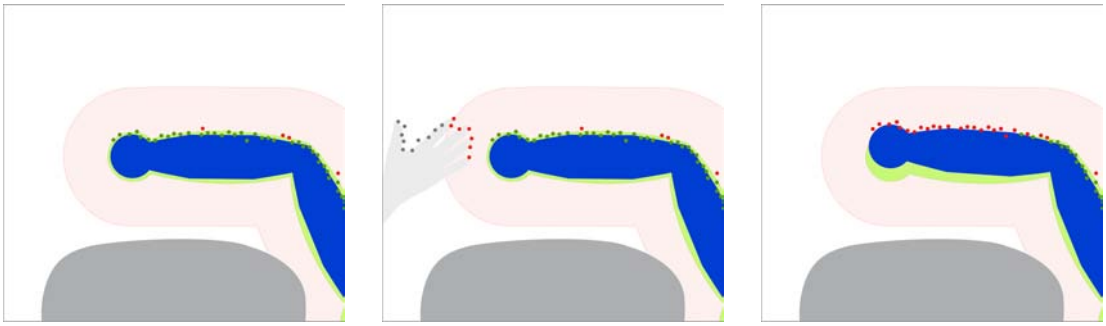


Figure 3.5.: Illustration of Shape Cropping applied to collision detection and continuous pose supervision. *Left:* No violation of the safety zone is detected; *center:* The safety zone is violated by an external object; *right:* The safety zone is violated by the robot itself due to a slightly incorrect pose of the robot.

3.3. Feedback to OR personnel

3.3.1. Clinical considerations

As detailed in section 2.3.1.3, the introduction of MIRS has altered the responsibilities and communication between the different persons in the OR, as the surgeon is now separated from the patient and outside of the sterile zone. This results in an increased demand for communication and explicit sharing of information. As Randell et al. note, this is especially true in case of complications, where the OR team is even more involved and therefore has to be aware of both the current situation and the current progress of the intervention [152]. This holds true especially for the assistant surgeon, who is inside the sterile zone and needs to be constantly aware of the actions and needs of the main surgeon outside the sterile zone [89].

Traditionally, feedback from medical devices in the OR is given using auditory signals, which enable drawing the attention of the OR personnel to a specific device, and/or using monitors, which have to be regularly checked. However, studies have shown that increased noise level in the OR can lead to increased SSI [98]. Additionally, there is the phenomenon of *alarm fatigue*, where a high amount of false alarms results in the medical personnel failing to respond to all alarm sounds. In hospitals in general, studies have estimated the percentage of false alarms at $72\,\% - 99\,\%$, especially in intensive care, with the additional problem that it is often unclear to the medical personnel which alarm is sounding. While the percentage of false alarms in ORs is not comparable, occasions have been reported where muting of alarms in an OR have led to the death of patients [163].

In a study on human factors engineering for improving patient safety in the domain of cardiovascular operating rooms, Gurses et al. also state that high noise levels can pose immediate perioperative cardiac surgery hazards. Furthermore, various problems with tools and technologies are listed as potential hazards,

including poor usability, inadequate safety features and safety features that do not fit to the users' needs and tools [47]. In a study with very similar scope, Pennathur et al. identified four major sources of technology-related hazards. One of these is "design factors", which include: Lack of feedback, information not being available "at a glance" and no dynamic information presented. Especially a lack of feedback is related to undesired cognitive load, caused i.e. by uncertainty about the status of a device and the required attention to deal with it [143].

Concerning the optimal presentation of such feedback, Rayo et al. propose that data can be presented in a continuous, non-interrupting way, using an alternative sensory modality, in order to *"reduce the overall mental workload required for directing attention and interpretation"*. Using alternative sensory modalities other than sound can also serve to reduce the overall mental workload [154].

For the reasons given above, it is not advisable to communicate additional information about the state of the robot system using sound as the main modality. Rather, a visual approach is required that can dynamically present information in the scene without distracting the personnel from their medical tasks.

3.3.2. Advantages of spatial augmented reality

As can be seen from the technical and human factors detailed above, it is crucial that the information gathered by the supervision system and especially the resulting safety measures are communicated to the OR personnel in an intuitive and non-intrusive way. The proposed usage of Spatial Augmented Reality (SAR) fits these requirements and offers several advantages for providing feedback for using surgical robot systems in the OR:

- Usage of SAR does not increase the noise level and the amount of different acoustic signals the OR personnel needs to discern.
- SAR allows to project information onto one or multiple specific locations in the scene, offering color properties (such as hue and brightness) as well as arbitrary shapes to convey information. In contrast, auditory signals always originate at a fixed source and can be modulated only in pitch and frequency.
- Information presented by SAR is visible to all OR personnel alike, thereby facilitating a shared situation awareness.
- Presenting information directly in the scene, e.g. by projection onto the robot or the situs, allows the OR personnel to maintain focus on their medical tasks without the need to e.g. regularly check a secondary monitor.

During an intervention there are many factors that can negatively influence the quality of projections. Among others, these include: Light being present in the scene, e.g. coming from the OR lamp in open interventions or illuminating the situs from within in case of minimally invasive interventions; drapes and covers with different colors and opaqueness, which crumple and thereby change shape during

the intervention; different surface reflectivity, especially in case of open surgery. Therefore, very detailed projections such as text or smaller elements cannot be relied on to be easily perceivable by the OR personnel and should be avoided. For these reasons, the SAR system proposed in this thesis does not employ intricate shapes or text, but instead focuses on conveying information by spatially correct projection, aided by non-complex shapes and a distinct color palette.

3.3.3. Augmentation concept

In the context of the proposed system, SAR is used to augment the scene with the following information:

- *State of robot arm*: Depending on its current mode, each robot arm is illuminated in a specific color that corresponds to its current state, allowing the user to discern if e.g. the robot arm is currently teleoperated by the surgeon or if it is decoupled and will not perform motions.

- *Instrument poses in MIRS*: Based on the known poses of the robot arms, the spatial configuration of the instruments can be projected onto the situs. This allows e.g. to quickly see if an instrument is in view of the endoscope and serves to further a shared awareness of the situation for all OR personnel involved around the operating table.

- *Feedback in case of adverse events*: If the expected behaviour of the surgical robot is overridden for safety reasons, e.g. if the supervision system detected an impending collision and the robot was stopped, according feedback is given directly in the scene, e.g. by illuminating an obstacle with which the robot would have collided.

The points listed above are directly addressing the subject of this thesis: safe and intuitive usage of a surgical robot system. Each can be realized based solely on data which is provided either by the supervision system, by the safety features or by the surgical robot system itself.

If the according information is available and its visibility in close proximity of the situs is desired, further projections could easily be added from a technical point of view. One example is the projection of information for tracked objects, e.g. as instruments or pointers, if the SAR-projector is extrinsically calibrated w.r.t. the tracking system. By projecting the tracked position as well as the tracking quality onto the scene, the user can focus on the respective task, e.g. marking points of interest, without having to consult a secondary screen to check the visibility of the tracked tool. For development purposes, this has been realized in this thesis and is widely used in OP:Sense development.

Further, physiological patient data or planning data such as trocar positions can be projected, if the patient has been registered. However, it has to be expected that projecting a multitude of information onto a confined surface leads to distractions and therefore has adverse effects on cognitive load.

3.3.4. Projection-based registration

All cameras integrated in the supervision system and the SAR-projector need to be extrinsically calibrated in order to establish a common coordinate system. As noted by Bauer et al. [7], this requires a simple procedure which can be included in the clinical routine in order to be applicable to real-world usage. Due to the number of cameras of the proposed supervision system, a pairwise registration scheme, e.g. based on a checkerboard, would require too much manual work for calibrating the system. This is especially true in the domain of an OR, where a high calibration accuracy is required, but the effort needs to be kept minimal.

To provide a simple and quick registration procedure, an active extrinsic calibration procedure is proposed that projects features as artificial landmarks into the scene using the SAR-projector. The 2D pixel position of each projected feature is stored as well as its 3D positions as detected by each camera, thereby building a set of known correspondences between all cameras and the projector. After collecting a sufficient amount of correspondences, the extrinsic calibration of all cameras and the projector as well as the calibration of the intrinsic projector parameters is calculated using bundle adjustment.

This registration procedure can be run without the need for continuous user interaction which is required in many other registration methods e.g. for replacing the checkerboard. The only required activity is to change the scene layout at specific times during the registration procedure to enable projections at different heights. This can e.g. be achieved by raising or lowering the OR table. Registration of optical tracking systems for navigation cannot be performed automatically, as these cannot detect visible light in the scene. The proposed projection-based extrinsic calibration method supports manual registration of such systems by reprojecting the previously projected features, which can then be annotated manually e.g. using a pointer. All manual annotations are then added to a correspondence list and are therefore also processed by the bundle adjustment step.

4. Realization

This chapter describes the realization of the system concept presented in section 3. It starts with giving an overview of OP:Sense, the platform for research on surgical robotics at the KIT IAR-IPR which has been co-developed throughout this thesis. Components of OP:Sense that are used and integrated by this thesis are discussed in terms of hardware and software. The system architecture is presented, including information about the required data and interfaces that are needed to integrate the proposed system with surgical robot systems.

The realization of the supervision system is outlined in terms of implementation of the different camera subsystems and discussion of optimal camera placement to maximize redundant coverage. The implementation of both the projection-based registration process and the algorithm for forward propagation of semantic labelling are presented in detail.

For the safety concept, the Shape Cropping algorithm is described, focusing on its implementation on the main processor and the subsequent parallelization using a GPU. This is followed by a description of the implementation and realization of the different safety features, such as active and passive robot localization, detection of impending collisions and continuous pose supervision.

Lastly, the system design of the SAR subsystem is described, taking into account both the physical setup and the software implementation. General use-cases are derived and their exemplary application to the OP:Sense platform is discussed.

4.1. OP:Sense

OP:Sense is a modular platform for research on new concepts for surgical robotics that is being developed at the KIT IAR-IPR. While original development was based on Windows and Matlab using a custom CORBA-based communication stack [124], the system was later ported to Linux and ROS. As OP:Sense was developed during the same time frame as this thesis was conducted, large parts of OP:Sense were created or contributed in this thesis, concerning both the supervision system and the general OP:Sense-architecture and components.

For further reference, a general overview about the architecture and implementation of OP:Sense using ROS has recently been published in the Springer book *Robot Operating System (ROS) - The Complete Reference* [95] in the chapter *ROS-based Cognitive Surgical Robotics* [12].

4.2. Components

4.2.1. Software

To allow for a flexible combination and extension of the different parts of the proposed system, its components are developed as lightweight, modular software nodes. The open source Robot Operating System (ROS) [150] is used as both a communication framework and ecosystem for development and debugging. After initial prototyping in Matlab, all processing of 3D data is realized using the Point Cloud Library (PCL) [159] whereas processing of 2D data is implemented using the Open Source Computer Vision (OpenCV) library [17].

The Ceres Solver [1] is used for performing the bundle adjustment required for the projector-based extrinsic calibration.

openFrameworks [196] is used for projecting information onto the scene and was extended with a ROS-interface to allow querying transformations and receiving instructions for graphical output via ROS, encoded e.g. as Scalable Vector Graphics (SVG) markup.

4.2.2. Sensors

3D cameras The supervision system developed within this thesis is composed of multiple 3D cameras, grouped into two independent subsystems as described in section 3.1.2. The PMD camera subsystem consists of six [pmd]vision® S3 cameras that feature a lateral resolution of $64 \times 48\,px$ and are connected and triggered via a standard Ethernet connection. Additionally, a [pmd]vision® CamCube 2.0 is integrated with a lateral resolution of $204 \times 204\,px$, connected and triggered via USB 2.0. The Kinect v1 subsystem consists of four Kinect v1 cameras with a lateral resolution of $640 \times 480\,px$ that are connected via USB 2.0.

Further, evaluation of the algorithms developed within this thesis was performed based on a Kinect v2 system realized together with Beyl [11]. It consists of four Kinect v2 cameras that offer a lateral resolution of $512 \times 424\,px$ and are connected via USB 3.0.

Optical tracking system The *ARTtrack2* Optical Tracking System (OTS) by *Advanced Realtime Tracking GmbH*, Germany, is used in a six camera configuration. It offers 6D tracking at $60\,\mathrm{Hz}$ for up to 20 rigid bodies fitted with retro-reflective marker spheres, which are compatible with clinical tracking systems such as the *NDI Polaris* series. Compared to the two-camera Polaris devices, the ARTtrack2 system offers a significantly higher tracking volume. Data provided by the ART tracking system was available through OP:Sense.

High precision measurement arm For obtaining high precision individual measurements, a *FARO platinum* measurement arm by *FARO*, Germany, was used. It can be fitted with either a ball probe or a laserline probe and is certified for a measurement accuracy of up to $50\,\mu m$. As no drivers for Linux are available for the measurement arm, a custom ROS node was implemented in Windows during this thesis. This allows to stream all data acquired by the measurement arm, such as probe position and laserline scanning, to ROS and thereby exposing the data to the OP:Sense system.

4.2.3. Robotic systems

Two *Light Weight Robot (LBR) 4* by *KUKA*, Germany, have been used in this thesis. They feature seven Degrees of Freedom (DoF) and are equipped with torque-sensors in each joint, which allows sensing of external forces. This is leveraged for a so-called *gravity compensation* mode, in which the user can guide the robot by hand, applying only minimal force. This mode will also be referred to as *hands on* mode in the following. Low-level control of the robots was available through OP:Sense.

To allow research into minimally invasive scenarios, OP:Sense offers a surgeon console. Different standard laparoscopic instruments are available that have been motorized and can be attached to the Light Weight Robot (LBR) and be controlled via ROS [64].

As a proof of concept, the proposed system has also been integrated with the MiroSurge system by DLR in the scope of the European research project Patient Safety in Robotic Surgery (SAFROS).

4.2.4. Interaction modalities

GUI-based system control To offer an easy interface for both development and usage of the system for research purposes, a modular Graphical User Interface (GUI) has been developed within this thesis. It interfaces via ROS with various components of both the general OP:Sense system and the proposed systems. It allows to control the supported systems, e.g. start and stop the different camera systems at a simple click, and visualizes their status. The GUI will be referred to as *system gui* from here on. It is implemented as a website based on `rosbridge`. As it is accessed via web browser, it allows full control of the system components on both static computers/touchscreens and mobile devices such as tablets or smartphones.

A 23 inch touch screen, the *HANNS-G HT231HPB* by *Hanns-G*, Taiwan, was installed for stationary usage of the system gui. It is connected to a dedicated Small

Form Factor (SFF) PC that was configured to directly boot into displaying the system gui, thereby enabling a comfortable one-click solution to visualizing system information.

For non-stationary usage of the system gui, a *Galaxy Note 10.1 2014* tablet by *Samsung*, Korea, was available.

Projection system For Spatial Augmented Reality (SAR), a short-throw projector was integrated into the setup, the *TH681* by *Benq*, Taiwan. It offers a resolution of $1\,920 \times 1\,080\,px$ with a light output of $3\,000$ ANSI lumens. The focal length is adjustable between $16.88\,\mathrm{mm}$ and $21.88\,\mathrm{mm}$, which corresponds to a throw ratio, i.e. ratio of distance to projection surface divided by projection width, of 1.15 to 1.5. As this allows for a wide projection area at short distances, the projector can cover the whole operating table from a short distance.

As a deflecting mirror, a custom made front face mirror by *G&P Optoelectronics*, Germany, was commissioned.

4.2.5. Servers

Control of all PMD cameras and processing of all resulting 3D data is performed on a dedicated server. It includes an *AMD Phenom(tm) II X6 1090T* hexa-core processor at $3.2\,\mathrm{GHz}$, $12\,\mathrm{GB}$ of working memory and a *NVIDIA GTX 480* graphics card. The graphics card features $1.5\,\mathrm{GB}$ dedicated memory and supports the parallel computing platform Compute Unified Device Architecture (CUDA) 2.0.

Processing of the raw depth data provided by the Kinect v1 cameras is performed on the so-called *Central Services* server. It features an *Intel Core i7 3770* processor at $3.4\,\mathrm{GHz}$, $8\,\mathrm{GB}$ of working memory and two *NVIDIA GTX Titan* graphics card. Each graphics card features $6\,\mathrm{GB}$ dedicated memory.

For tasks that require little computational power while being restricted to a small footprint, SFF PCs have been employed. The *ZBOX nano AD10* by *Zotac*, Hongkong, has been selected as basis for the Kinect v1 subsystem. At a small footprint of $45 \times 127 \times 127\,\mathrm{mm}^3$, it features an *AMD E-350* dual-core processor at $1.6\,\mathrm{GHz}$ and $4\,\mathrm{GB}$ of working memory. Connectivity is provided by both a USB 2.0 and a USB 3.0 host controller as well as a $1\,\mathrm{Gbit}$ Ethernet port.

The successor of this SFF PC, the *ZBOX nano AD13* by Zotac, has been used for the touch screen and for controlling the surgical instruments. At an even smaller footprint of $37 \times 106 \times 106\,\mathrm{mm}^3$, it features an *AMD E2-1800* dual-core processor at $1.7\,\mathrm{GHz}$ and $4\,\mathrm{GB}$ of working memory.

4.3. System architecture

4.3.1. Design goals

The proposed system has been designed with the following goals:

- *Modularity*: Components are realized as independent modules that focus on a specific logical (sub)task and communicate via defined interfaces.
- *Fault tolerance*: In case of failure of single components, e.g. hardware or infrastructure, the system degrades gracefully.
- *Extensibility*: Additional functionality in terms of hardware capabilities, e.g. new sensors, and software functionality, such as interfaces to surgical robot systems, can easily be added to the existing system.
- *Spatial distribution*: The sensors and processing hardware are spatially separated in order to minimize the system's footprint close to the situs.

4.3.2. Overview

The high level overview of the architecture of the proposed system and its main components is shown in Figure 4.1. Directed data flow between components is represented as solid arrows and annotated with the type of information that is transferred on the according connection. Dashed arrows represent perception of the world by 3D cameras, i.e. PMD cameras for the first level scene model and Kinect v1 for the second level scene model.
Unlabeled connections between the system gui and other components allow for the following interactions: Both supervision subsystems can be switched on or off via the system gui, which also displays their current status. If the robot system provides an according interface, as is the case for OP:Sense, the system gui can also be used to switch between different robot control modes and visualize their status.

4.3.3. Distributed system

The proposed system as well as all subsystems have been realized based on ROS as a communication framework. Therefore, a short introduction into the basic principles of ROS is necessary before further aspects of implementation are discussed.

ROS provides a structured communication layer through which a peer-to-peer topology of multiple so-called *nodes* can be established and dynamically modified. Data is exchanged based on the publish-subscribe pattern via named *topics*, over which *messages* of different types can be sent. While different network protocols

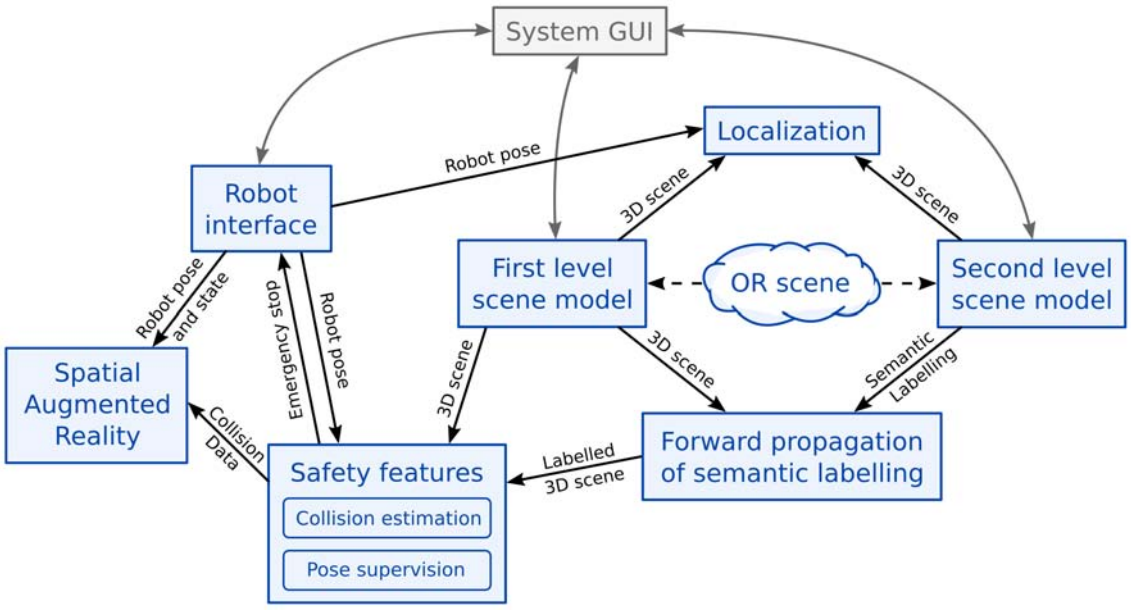


Figure 4.1.: High-level overview of the system architecture and data flow.

are supported, all implementations of this thesis are based on TCP/IP. Connections between different nodes are established based on a central naming service, the ROS *master*, that maintains a list of active nodes and advertised topics. After a connection has been established, data is exchanged directly between the participating nodes without involvement of the master. Therefore, once a system is running, the master does not represent a single point of failure. In ROS, spatial relations, i.e. transformations, between different entities are represented using the so-called *tf* mechanism. Based on pairwise transformations between named entities that can be published by arbitrary nodes, a *tf tree* is maintained that can be queried for arbitrary concatenated transformations.

In the proposed system, some nodes have physical dependencies which limit them to certain hardware. Examples are the node controlling the PMD camera subsystem, which needs to run on a machine that is directly connected to the CamCube via USB and to the S3 cameras via network, and the projection node, which needs to be executed on the SFF PC to which the projector is attached. All other components can be executed on arbitrary machines as their required input data is available via network. In case of hardware failures, this makes it possible to switch affected nodes to a different physical machine without delay.

Further constraints and details of the distributed implementation will be discussed below with the specific subsystems.

4.3.4. Connection to surgical robot systems

As the field of surgical robotics is quickly developing and several new systems are in a state of advanced research or close to commercialization (see section 2.2.1.4), the proposed concept for safe and intuitive usage needs to be applicable to different surgical robot systems.

To construct the safety zone and discern e.g. between correct and faulty robot actions, the exact shape of the robot has to be known at any given moment. Therefore, pose information for each segment needs to be available continuously as well as CAD data for each segment of the robot, from which the mesh models for the inner and outer hull are generated in an offline step. The pose of each segment can be either directly provided by the robot control or calculated by the proposed system based on a known kinematic model of the robot and the current joint angles. The latter allows calculating the robot's pose independently from the robot control, thereby adding another layer of redundancy to the system.

Both methods have been realized during this thesis to connect to two different research surgical robot systems:

- OP:Sense: A kinematic model has been implemented to calculate the pose of each segment of the LBR based on its Denavit-Hartenberg parameters. The joint angles are provided via the ROS network.

- MiroSurge: The full pose of each segment was provided by the robot control over network via Remote Procedure Calls (RPCs).

For merely supervising the safety of the system without triggering a reaction in case of adverse situations, only a unidirectional interface to the robot is required through which the current pose is streamed (see Figure 4.2). For taking full advantage of the safety features, a bidirectional interface is necessary so that information such as the safety state can be provided or an emergency stop can be triggered.

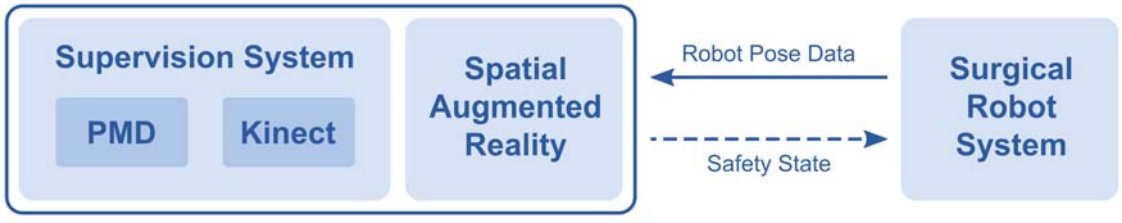


Figure 4.2.: Communication with surgical robot system: Robot pose data is required, sending back the safety state is optional.

4.4. Supervision system

4.4.1. Architecture

The supervision system as described in section 3.1.2 has been realized as a distributed system. Its network topology is shown in Figure 4.3 as an illustration of the architecture.

The PMD subsystem consists of a dedicated server which controls all PMD cameras and processes all acquired data. Six pmd[vision] S3 cameras are connected via Ethernet on dedicated networks whereas the CamCube 2.0 is connected via USB. The usage of dedicated networks prevents potential load problems on the general ROS network from interfering with the safety-critical PMD camera subsystem. In terms of the system design goals, it also increases the fault tolerance against hardware failure and the modularity of the system.

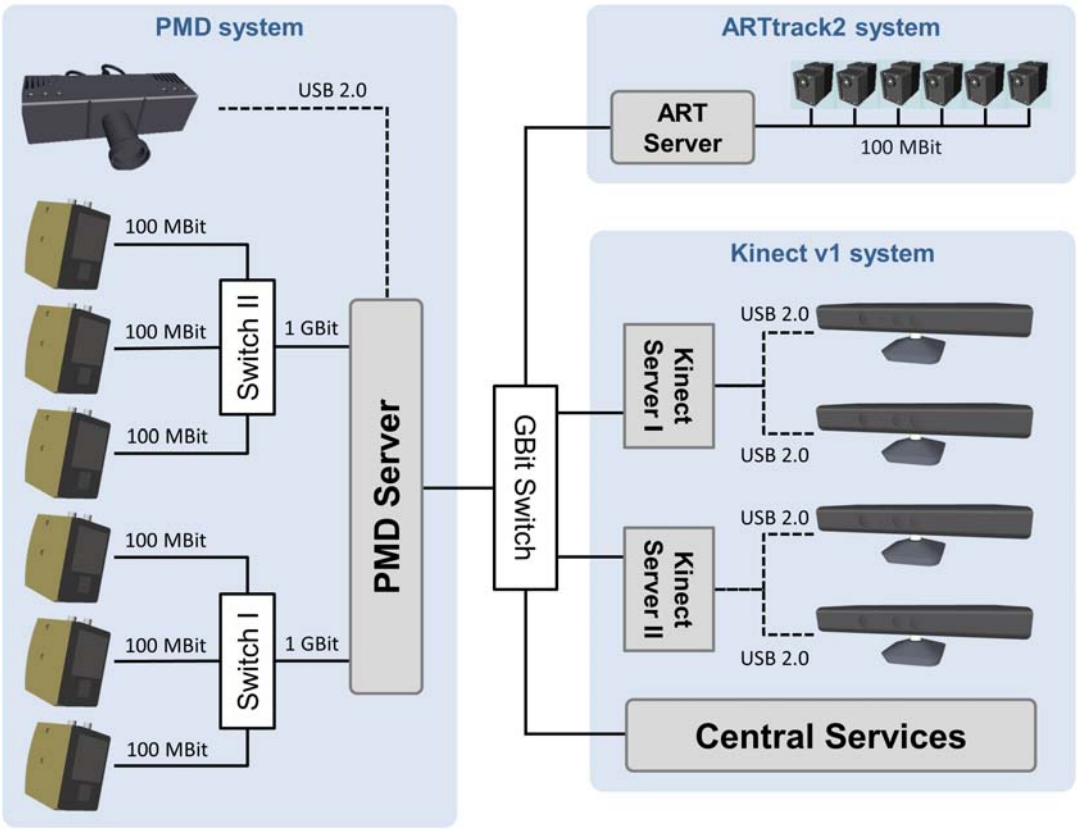


Figure 4.3.: Network topology of the supervision system with the PMD and Kinect v1 camera system as well as the standalone OTS ARTtrack2.

The Kinect v1 subsystem is realized as a distributed system within the proposed system: It is based on two AD10 SFF PCs to which two Kinect v1 are connected each. Processing of the acquired raw data is performed on the Central Services

server. As the second level scene model, which is based on the Kinect v1 camera system, is not safety-critical, latencies or frame drops in the Kinect processing pipeline due to potential network overloads can be tolerated. Usage of the AD10 SFF PCs enables to transfer the acquired data via Ethernet to the processing nodes. This allows for spatial separation between the sensors and the processing server as Ethernet segments feature a higher maximum length and reliability than extended USB cables.

Further, the standalone ARTtrack2 is connected via Ethernet to the main ROS network.

4.4.2. Camera placement

The task of placing sensors to supervise a scene has been studied in various works. Most recently, the PhD thesis of Hänel proposed different algorithms for achieving optimal coverage by freely placing cameras in a 3D environment, taking into account occlusions caused by static and dynamic obstacles. Hänel concludes that compared to a random placement, optimal placement of cameras can double the covered area; in comparison to simple heuristic approaches, optimal placement still achieves better results. However, it is also noted that complex scenes, e.g. including multiple rooms or complex obstacles, profit more from optimized sensor placement than simpler scenes. Additionally, Hänel reports that in both exemplary environments that were analysed in the thesis, for more than $75\,\%$ of the cameras the optimal placement was found to be located at the boundaries of the respective domain [52].

Both based on these findings and due to the spatial constraints of the laboratory in which the OP:Sense system is developed, the camera systems of this work needed to be realized as a ceiling mounted camera system where the cameras are located at the boundaries of the ROI around the OR table.

For ceiling-mounted cameras, Figure 4.4 illustrates the occlusions that can be caused by either personnel standing between camera and operating table or the OR lamp being in a low position. It is based on a person height range of $1.68\,\mathrm{m}$ – $1.86\,\mathrm{m}$, which corresponds to the majority of male US citizens (10% – 90% percentile) at an age of 20 years [36], and also is in line with the average European male height of $1.77\,\mathrm{m}$ [44]. The operating table is shown in an even position at a height range of $0.66\,\mathrm{m}$ – $1.1\,\mathrm{m}$, with the OR lamp at the *Central Illumination* distance of $1\,\mathrm{m}$.

For actual interventions, there are countless more variables that influence potential occlusions: number, position and pose of persons around the operating table, the shape of the operating table itself (which can be raised, lowered and angled in segments), the position of the OR lamp(s), etc. However, the simple illustrations of Figure 4.4 serve to show that there is an equilibrium of camera mounting height: Mounting the cameras in a higher position can decrease occlusions by persons, but

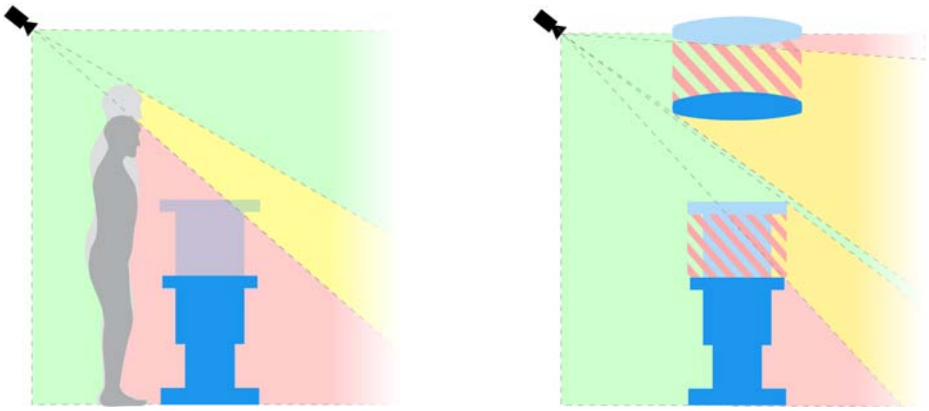


Figure 4.4.: Occlusions caused by personnel and OR lamp, illustrated for one camera. Green areas have clear Line of Sight (LoS), yellow areas might be occluded, red areas are not visible for the camera. Striped areas indicate volume where objects might be located, e.g. operating table height might be raised or lowered. *Left:* Occlusions caused by person standing between camera and OR table; *right:* Occlusions caused by OR lamp and operating table.

Figure 4.5.: Visualization of the camera poses relative to the OR table. Poses of PMD cameras are depicted in orange, Kinect v1 cameras in red and Kinect v2 cameras in gray. The coordinate system between both robots represents the origin of the optical tracking system ARTtrack2. Shapes of the camera pose markers represent the FoV of the corresponding camera.

increases the possibility of occlusions by the OR lamp(s) and vice versa. Therefore, cameras have been mounted as depicted in Figure 4.4 at a height of $2.15\,\mathrm{m} - 2.25\,\mathrm{m}$ which leaves enough space to walk comfortably below each camera and offers a high coverage of the scene.

Figure 4.5 shows the poses of all cameras of the realized camera systems as well as those of the Kinect v2 cameras and the origin for the marker-based tracking system.

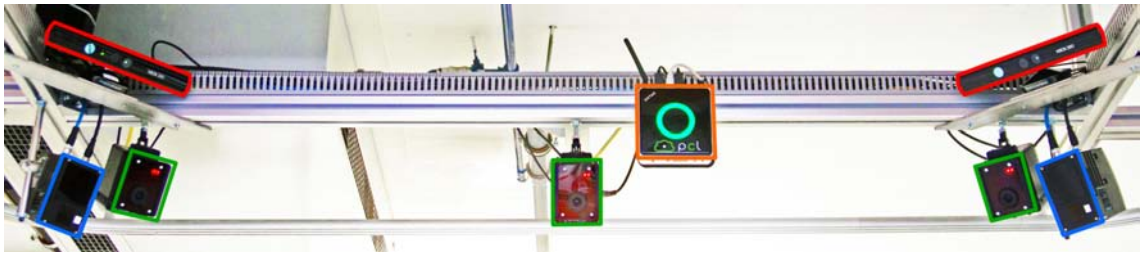


Figure 4.6.: Partial view of the realized supervision system with first level 3D cameras ([pmd]vision S3, blue), second level 3D cameras (Kinect v1, red, with Kinect server, orange) and the optical tracking system (ARTtrack2, green).

Figure 4.6 shows a part of the realized supervision system. It illustrates the physical size of an AD10 SFF PC (bordered orange), to which two Kinect v1 cameras are connected.

4.4.3. ToF subsystem

All cameras are controlled by a dedicated server that triggers each camera and performs light preprocessing on the raw data. It publishes the resulting point cloud as well as amplitude and depth images to the ROS network for further processing by high-level nodes.

4.4.3.1. Trigger modes

The pmd[vision] S3 cameras support different modes for image acquisition: *free-run*, *hardware trigger* and *software trigger*. In free-run mode, the camera acquires range images at maximum speed which can then be polled by the processing software. This is not feasible for systems where multiple cameras need to operate in the same volume, both for reasons of crosstalk (see section 2.2.2.3) and synchronization.

To enable optimal synchronization between multiple [pmd]vision S3 cameras, a hardware trigger is supported. This allows to physically connect the *Ready pin* of one camera, which then servers as the master camera, to the *TriggerIN* pins

of one or multiple slave cameras. Triggering the master camera (via software) automatically triggers the connected slave cameras, either at the same time as the master (trigger on negative edge) or consecutively (trigger on positive edge). While this optimizes the synchronization, it has two major drawbacks: First and most severe, a component failure affecting the master camera also affects the slave cameras. Both a failure of the master camera itself and a failure of the network connection, via which the master camera is triggered, would result in complete loss of information of the master camera itself and all connected slave cameras, as they would not be triggered any more. Second, both triggering cables and Ethernet cables are now necessary for each camera. This results in a mesh network topology that is a mixture between the original star topology and a ring topology. It decreases the flexibility of the physical system setup and increases the probability of error, as failure of one connection results in the loss of at least one camera.

Due to the drawbacks of both free-run mode and hardware trigger, the final system was realized using software trigger. This allows for a clean and flexible star topology of the network, reduces the amount of connections and thereby the potential points of failure, and provides full flexibility in triggering. In a dedicated network where all traffic is controlled by the PMD server, the loss of timing accuracy as compared to using a hardware trigger is minimal.

4.4.3.2. Time and frequency multiplexing

ToF cameras are prone to interferences due to their sensing principle as described in section 2.2.2.3. Therefore, a time and frequency multiplexing scheme has been devised to prevent crosstalk between all PMD cameras.

As the [pmd]vision S3 camera only provides three different modulation frequencies, the six [pmd]vision S3 used in the supervision system are split into two logical groups. In each group, all three cameras are configured to different modulation frequencies and triggered at the same time. This results in three entities that need to be time-multiplexed, namely two S3 camera groups and the CamCube. Using a fixed configuration of the cameras, the total time for acquiring each frame was measured and analyzed to determine the initial time span during which the camera actively emits light. Based on these measurements, different time multiplexing schemes have been implemented throughout the course of this thesis as depicted in Figure 4.7:

- *Simple alternating*: Both S3 groups are triggered alternately with the CamCube triggered in between. The CamCube can be configured to arbitrary modulation frequencies.
- *Synchronized alternating*: Both camera groups are triggered alternately with the CamCube synchronized to both groups. The CamCube needs to be set to a modulation frequency which is different from the modulation frequencies used by the pmd[vision] S3 cameras.

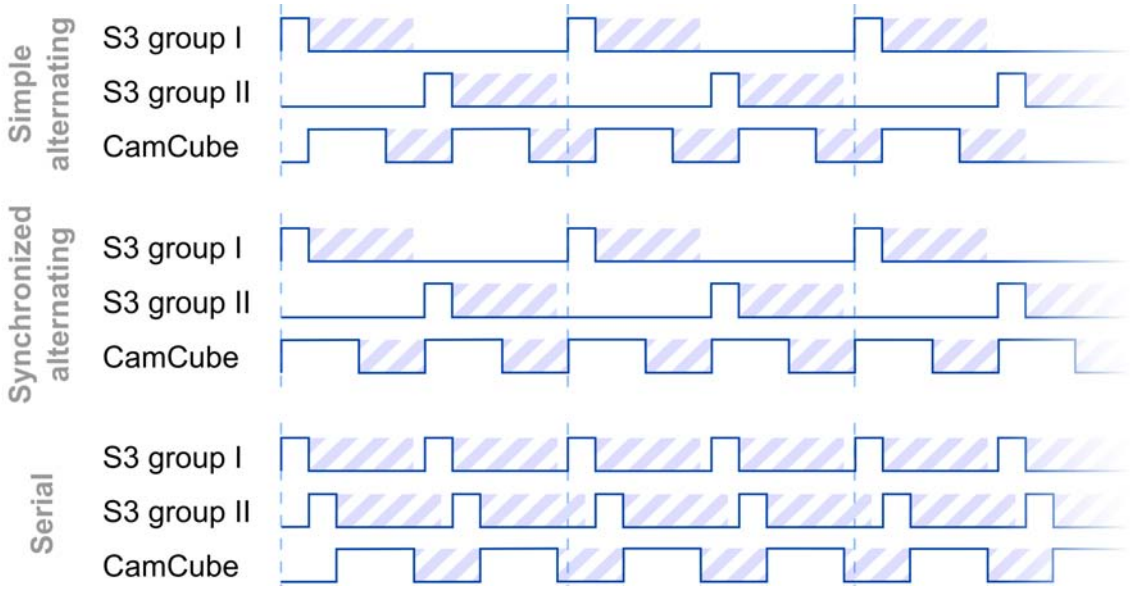


Figure 4.7.: Different triggering schemes for PMD camera subsystem. High edges represent the time that a camera is actively emitting light, stroked appendices illustrate the remaining time for transfer and processing.

- *Serial*: The camera groups and the CamCube are triggered serially. The CamCube can be configured to arbitrary modulation frequencies.

The serial triggering scheme was selected as best candidate as there is the least amount of time where no camera emits any light, i.e. it offers a maximum amount of information acquired in a minimal amount of time.

4.4.3.3. Low-level preprocessing of raw data

ToF cameras exhibit a special characteristic of noise, the flying pixels that occur at boundaries between foreground and background (see section 2.2.2.3). Different methods have been proposed to correct the distance information at the concerned pixels, ranging from simple median filtering on neighboring pixels to more involved filtering pipelines that model the different return paths [104]. However, these methods come at a computational cost and are mostly required if only one camera is available as single data source and therefore a maximum amount of information needs to be extracted out of each measurement. As the PMD camera system consists of multiple cameras monitoring the scene from different points of view, this thesis implements an approach to detect and remove such outliers in the raw data of each camera.

In general, it is not possible to detect flying pixels solely based on their amplitude. Instead, an edge detection filter is employed for identifying flying pixels as they occur at the boundaries of objects by definition. This thesis uses a combination of Sobel operators that calculate the gradient of the image in different directions as

depicted. For visualization, an example based on a depth image acquired by the the CamCube is shown in Figure 4.8.

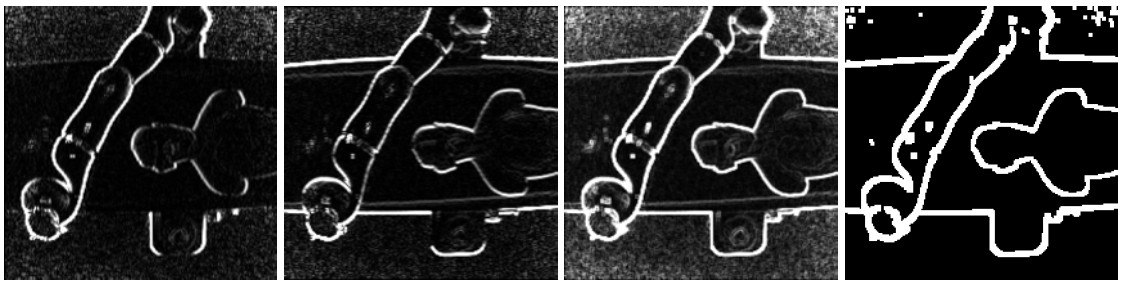


Figure 4.8.: To detect flying pixels, Sobel operators are applied to the depth image in different directions *(1st, 2nd image)*. Results are combined *(3rd image)*, binarized and dilated, resulting in a mask *(4th image)* which is then applied to the original data.

Further, a sliding window with a size of two frames is employed to temporally smoothen the data acquired by the PMD cameras. While pixels that are flagged as invalid by the PMD driver are removed, no other filtering, e.g. based on amplitude or spatial coherence, is performed by default during the low level preprocessing in order to retain the according data for higher level applications.

4.4.3.4. Operating modes

Due to the nature of camera systems in general, short exposure times result in a lower SNR than longer exposure times, but are more suited for capturing non-static scenes. This holds especially true for ToF cameras, where each range image is calculated based on multiple consecutive sub-images and differences between the sub-images result in incorrect reconstruction of the 3D data.

To enable high-level nodes to choose between a higher frame rate and a higher quality while encapsulating the actual implementation, two different operation modes of the PMD subsystem have been implemented that can be switched at runtime:

- *Performance mode:* The cameras are configured to shorter integration times, i.e. $750\,\mu s$ for the pmd[vision] S3 and $1\,500\,\mu s$ for the CamCube 2.0, and are triggered using time- and frequency multiplexing. Extended modes such double frequency acquisition are disabled.
- *Quality mode:* The cameras are configured to longer integration times of $4\,000\,\mu s$ for each camera and are triggered serially. Double sampling mode for increasing the SNR and double frequency mode for increasing the unambiguity range are enabled for the pmd[vision] S3 cameras. Both modes are not supported by the CamCube 2.0.

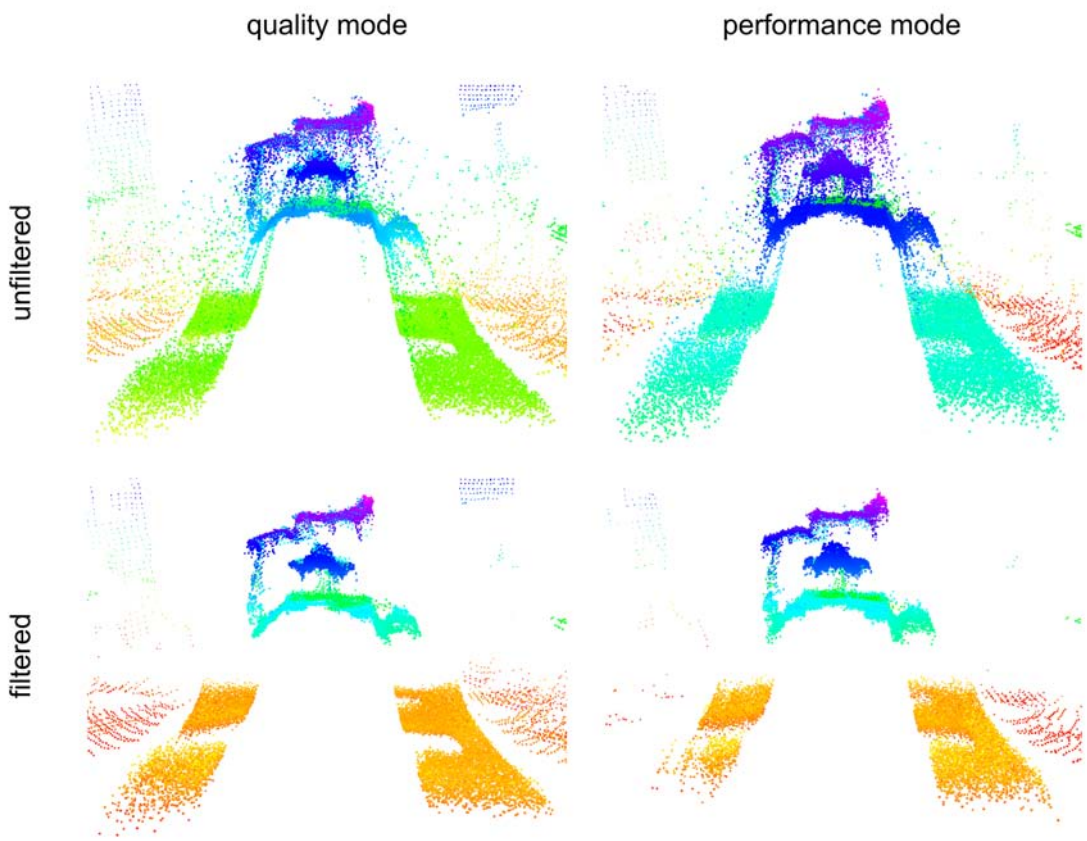


Figure 4.9.: Visual comparison of a scene acquired in quality mode *(left)* and performance mode *(right)* with removal of flying pixels disabled *(top)* and enabled *(bottom)*.

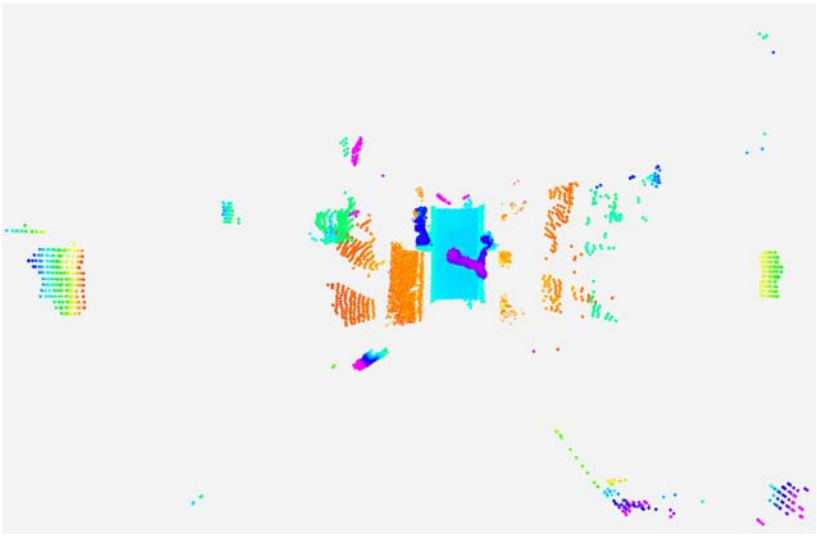


Figure 4.10.: Top down view of a virtual representation of the medical robotics laboratory at IAR-IPR acquired by PMD camera system.

As a visual comparison of the result of the different operating modes and the effect of filtering for flying pixels, Figure 4.9 shows a side view of a scene captured by the full PMD subsystem. The scene consists of an OR table on which a phantom is placed and to which an LBR is attached. Figure 4.10 shows the resulting virtual scene acquired by all cameras of the PMD camera system.

4.4.4. Kinect v1 subsystem

The Kinect v1 subsystem consists of four Kinect v1 cameras. Contrary to the industrial grade PMD cameras, the Kinect v1 offer no configuration options. Instead, they output a high-resolution range image with color information at their maximum frame rate. Figure 4.11 shows the resulting virtual scene acquired by all cameras of the Kinect v1 camera system.

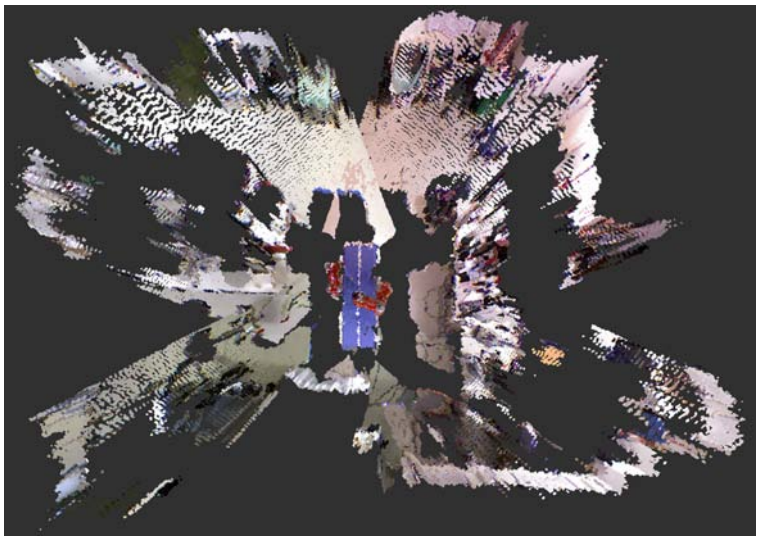


Figure 4.11.: Top down view of a virtual representation of the medical robotics laboratory at IAR-IPR acquired by Kinect v1 camera system.

4.4.4.1. System design

The Kinect v1 provides raw data depth and color streams at a resolution of $640 \times 480\,px$ and a frame rate of $30\,\mathrm{fps}$ via USB 2.0. As the distance measurement for each pixel is encoded as an $11\,\mathrm{bit}$ value, the raw depth stream requires a theoretical bandwidth of $640 \times 480 \times 11\,\mathrm{bit} \times 30\,\mathrm{Hz} = 101\,376\,000\,\mathrm{bit/s} \approx 96.7\,\mathrm{Mbit/s}$. In the raw color stream from the Bayer sensor, each pixel is encoded as an $8\,\mathrm{bit}$ value, which results in a bandwidth requirement for the color stream of $640 \times 480 \times 8\,\mathrm{bit} \times 30\,\mathrm{Hz} = 73\,728\,000\,\mathrm{bit/s} \approx 70.3\,\mathrm{Mbit/s}$. In total, the required bandwidth for the raw data streams is $96.7\,\mathrm{Mbit/s} + 70.3\,\mathrm{Mbit/s} = 167\,\mathrm{Mbit/s}$ plus communication overhead.

After raw data from a Bayer sensor is transformed into a full `RGB` image, each pixel contains the information of three color channels and therefore requires $3 \times 8\,\mathrm{bit} = 24\,\mathrm{bit}$ of memory per pixel. Each point of a `XYZRGB` pointcloud therefore requires $3 \times 32\,\mathrm{bit} = 96\,\mathrm{bit}$ for three `float` values representing the point position, plus $24\,\mathrm{bit}$ for the color information. For a pointcloud at the full Kinect v1 resolution, this amounts to a bandwidth requirement of $640 \times 480 \times (24\,\mathrm{bit} + 96\,\mathrm{bit}) \times 30\,\mathrm{Hz} = 1\,105\,920\,000\,\mathrm{bit/s} \approx 1\,054.7\,\mathrm{Mbit/s}$.

Even without acknowledging protocol overhead, this bandwidth requirement exceeds the available bandwidth of $1\,024\,\mathrm{Mbit/s}$ provided by a standard $\mathrm{Gbit/s}$ Ethernet connection. To enable transferring the full Kinect v1 data without decreasing the frame rate or the resolution, a streaming setup was realized which transfers the raw data of the Kinect v1 over Ethernet using ROS. The calculation of the resulting pointcloud per camera is thereby shifted from the SFF PC, to which the camera is directly connected, to the Central Services server on the network. This lowers the computational load of the SFF PCs and nowadays would allow to use fanless PCs, which complies to the requirement of not disturbing the laminar air flow over the OR table (see section 3.1.1.1).

Due to the protocol overhead of USB and reserved bandwidth for the operating system, the Kinect v1 requires over 50% of the practically available USB 2.0 High Speed bandwidth. This means that each Kinect v1 has to be connected to a dedicated USB 2.0 host controller. However, even with the small size of the selected SFF PCs, it is desirable to keep their number to a minimum in order to prevent unnecessary clutter. For this reason, the AD10 SFF PCs were deliberately selected based on both their form factor and their connectivity: With one USB 2.0 and one USB 3.0 controller, they allow to connect two Kinect v1 cameras simultaneously per USB and stream the raw data over the $1\,\mathrm{Gbit/s}$ Ethernet port into the network.

4.4.4.2. Human Tracking

For human tracking, the approach of Beyl [9, 10] was employed in combination with the Kinect v1 camera system. Human tracking information is streamed into the ROS network as joint positions and `PointCloud2` messages containing the full body pointcloud of each tracked user. For forward propagation of semantic labelling (see section 4.4.7), the human tracking serves as ground truth.

4.4.5. Kinect v2 subsystem

Like the Kinect v1, the Kinect v2 is officially supported only with the Windows operating system. For both cameras, unofficial drivers for Linux have been available shortly after the commercial availability of the camera; these are used in the Kinect v1 camera system described above. However, the original body tracking algorithms developed by Microsoft are not available when using open drivers

and open source implementations of body tracking at non-frontal views have only recently reached maturity.

For the Kinect v2, it was therefore desired to enable access to the full functionality of the official Kinect v2 SDK, i.e. color and depth data as well as body tracking based on the original algorithms trained by Microsoft, for further processing in ROS. A custom, windows-based ROS node was implemented in this work that accesses the camera using the official SDK, processes the data and makes it available in ROS as native messages.

This serves as the basis for a Kinect v2 camera system with four cameras that was realized by Beyl [9]. In this thesis, the Kinect v2 camera system was used in addition to the PMD camera system and the Kinect v1 camera system to evaluate proposed algorithms and to serve as a comparison system. It is however not applicable to the scenario of this work, as the Kinect v2 cameras can exhibit severe interferences in the presence of two or more cameras targeted at the same region due to a lack of synchronization [9]. This effect is shown in Figure 4.12.

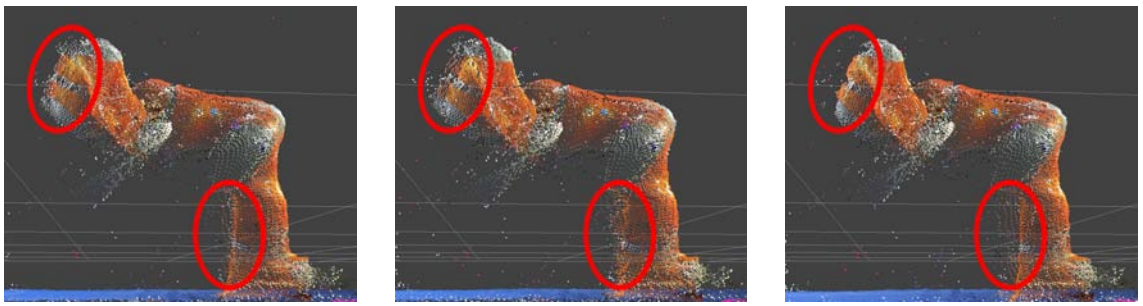


Figure 4.12.: High-frequency interferences regularly observed by operating four Kinect v2 cameras in the same volume. Distortions are visible at the marked parts of the robot.

For the depth measurements acquired by the uppermost and lowermost pixel rows of each Kinect v2, there is no color information available due to the different aspect ratios of the color sensor and depth sensor. These depth measurements have been colored pink. Figure 4.13 shows the resulting virtual scene acquired by all cameras of the Kinect v2 camera system.

4.4.6. Projection-based registration

Both in the proposed system as well as in current operating rooms, various devices are employed for acquiring 3D information and/or displaying information based on 3D information: 3D cameras (PMD, Kinect), projectors, navigation systems (ARTTrack2) and potentially other devices (such as the FARO measurement arm). In order to exchange geometric data, all devices need to be registered to a common reference frame.

Figure 4.13.: Top down view of a virtual representation of the medical robotics laboratory at IAR-IPR acquired by Kinect v2 camera system.

A typical registration process for multi-camera systems contains the following general steps:

1. *Feature detection*: Features in the scene are detected based on the image acquired by each camera.
2. *Correspondence estimation*: Correspondences are estimated between the feature detections of multiple cameras.
3. *Camera registration*: Based on the estimated correspondences, the pose of each camera is estimated either w.r.t. each other or to world coordinates.

For the registration of 2D cameras, feature detection is often performed by introducing artificial features such as a calibration object with known geometric properties. The known geometric properties enable to estimate the 3D pose of the features w.r.t. the camera. Similar methods have also been applied to RGB-D camera systems by combining the texture-based feature detection of a checkerboard pattern with the depth-based distance measurements of each detected feature [10, 90]. However, registration methods for which a registration object has to be manually placed in different fixed positions rely heavily on human involvement and are often cumbersome processes. Approaches for RGBD camera registration have recently been proposed that allow registration based on a checkerboard which is dynamically moved through the scene [129] or that completely eliminate the need for calibration objects and only require unstructured motion in the scene [122].

However, these registration methods are only targeting camera registration and generally do not allow for registration of other devices, such as projectors or, in the case of registration methods without fixed positions of the calibration target, optical tracking systems. Concerning the specific task of projector-camera

calibration, e.g. for structured light based scanning systems, different methods have been proposed, e.g. by Moreno et al. who estimate the intrinsic and extrinsic parameters of both camera and projector based on projecting structured light patterns onto a checkerboard [127]. However, this also requires careful manual positioning of the checkerboard as well as the usage of a high resolution camera to accurately reconstruct the intrinsic parameters of the projector.

For this reason, a projector-based registration method is proposed that (i) replaces the need for calibration targets such as a checkerboard, (ii) can perform automatic registration of multiple cameras and (iii) allows for additional manual registration of e.g. optial tracking systems or components such as the FARO measurement arm. As features are projected serially, this methods also eliminates the correspondence estimation step.

4.4.6.1. Registration workflow

The registration method consists of three main stages as depicted in Figure 4.14. After initialization, the scene needs to be arranged so that there is a non-cluttered surface which is at least partially visible from all cameras, lies inside the projection frustum of the projector and is at least partially covered by any marker-based tracking system which need to be registered. In the clinical scenario with the proposed setup, the OR table can be used as projection surface as it fits all these requirements. As last part of the scene arrangement, the user interactively determines the desired projection area as a subset of the full area covered by the projector. Moving lines are projected into the scene that allow the user to iteratively set the boundaries of the desired projection surface as depicted in Figure 4.15.

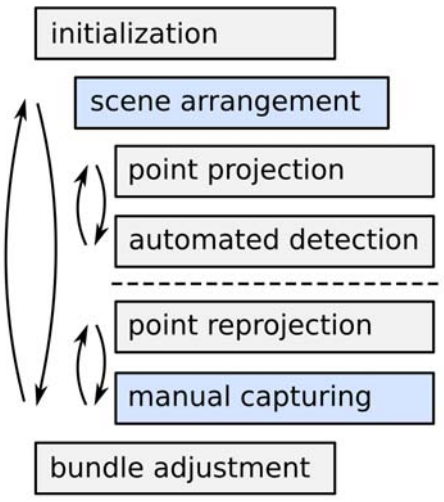


Figure 4.14.: Logical flow of the registration procedure. Steps highlighted in blue require manual actions of the user.

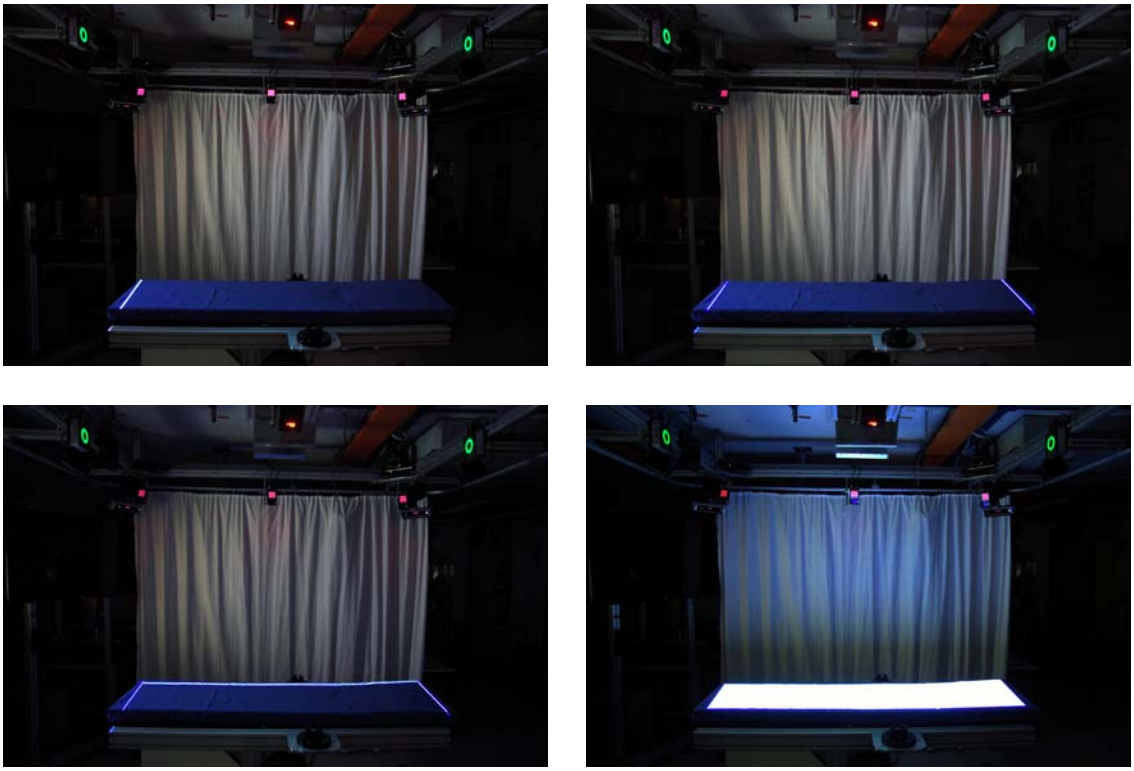


Figure 4.15.: Exemplary steps of the realized registration procedure. The user interactively determines the projection area by setting its boundaries via lines projected into the scene, as visible on the OR table. After all boundaries are set, the resulting projection area is shown for confirmation.

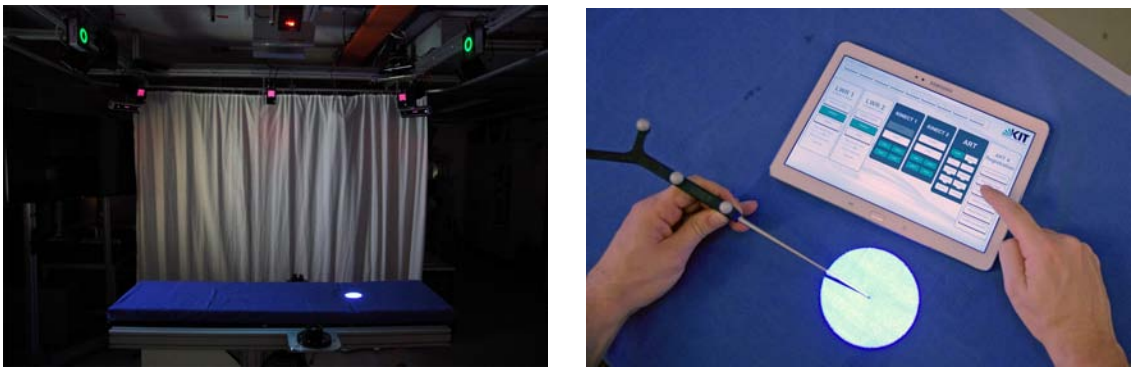


Figure 4.16.: Features are projected into the scene inside the user-determined projection area *(left)*. For non-camera devices such as an OTS, features can be annotated manually *(right)*.

After the scene is arranged, all cameras capture the scene without illumination by the projector for several seconds. For each camera, an averaged scene representation is calculated over all measurements both in 3D as a *reference point cloud* and as a 2D *reference image*. Afterwards, features are projected as filled circles one after another. Again, the scene is captured and averaged by each camera, resulting in a 3D point cloud and a *feature image* with the projected circle. To localize the feature position, the differences between reference image and feature images are calculated and thresholded, resulting in a mask of the projected circle in 2D camera pixel coordinates. The feature location can then be calculated as the center of the circle based on the coordinates of the points of the reference point cloud that correspond to the 2D circle mask.

While the camera feature acquisition described above does not require input by the user, the registration of other systems needs to be performed manually. The features are projected to the same locations as before and can be manually annotated with a modality depending on the specific device. In this thesis, annotations with the FARO arm were performed by positioning the measurement tip at the feature position in the scene and confirming by pressing a button on the device. The ARTtrack2 OTS was registered by positioning an NDI pointer to the feature location and capturing the position via button press in the system gui displayed on a mobile device (see Figure 4.16).

This procedure can be iterated multiple times with different scene arrangements, i.e. by adjusting the OR table height, to increase the number of annotated features. After the last iteration, the pose of each camera or device as well as the intrinsic parameters of the projector are calculated using *bundle adjustment*.

4.4.6.2. Bundle adjustment

Triggs et al. define *bundle adjustment* as "the problem of refining a visual reconstruction to produce jointly optimal 3D structure and viewing parameter (camera pose and/or calibration) estimates." [178]. This optimization problem can be formulated as a *non-linear least squares* problem, using the squared Euclidean norm of the reprojection error of each feature in each camera as the optimization criterion.

In this thesis, the devices which need to be registered to a common reference frame can be split into two different categories:

- Devices for which a static intrinsic calibration can be assumed, such as cameras with a fixed focal length, and navigation systems as the ARTtrack2 or NDI Polaris devices.

- Devices with intrinsic parameters that depend on the current device configuration, such as cameras or projectors with a variable focal length that can e.g. be adjusted by the user via a zoom ring.

Projection models In order to perform bundle adjustment, a model for the reprojection error, the *residual*, needs to be established for each device. In case of devices with static intrinsic calibration, the component-wise residual $\vec{e}$ can simply be modelled as

$$\vec{e} = (R \cdot \vec{p} + \vec{t}) - \vec{o}, \tag{4.1}$$

where p denotes the predicted point in the camera coordinate system, R the camera rotation matrix, $\vec{t}$ the camera translation vector and $\vec{o}$ the observed point in world coordinates. Therefore, only the extrinsic parameters of the according devices need to be optimized.

For devices of the second category, both intrinsic and extrinsic parameters need to be optimized. Especially in a setting where multiple persons interact with a system on a daily basis, it cannot be asserted that intrinsic parameters were not changed since a previous calibration. For this reason, an extended model is required for calculating the residuals.

The wide-angle projector is the only device with variable intrinsic parameters used in this thesis. Similar to cameras, projectors consist of an optical system that establishes correspondences between coordinates of 3D points in a scene and pixel coordinates on a chip. Therefore, the standard *pinhole camera* model with additional distortion coefficients as given in Appendix B can also be used to model the optical properties of a projector.

Due to the inverted nature of a projector compared to a camera system, residuals are calculated in pixel space as the difference between pixel coordinates that were projected into the scene and the reprojected pixel coordinates that correspond to the 3D coordinates of the projection in the scene, based on the intrinsic and extrinsic parameters of the projector.

The registration problem is then modeled as a bundle adjustment problem, i.e. a minimal least squares problem. The residual, i.e. the reprojection error, for all observations by cameras, the ARTtrack2 and FARO, is calculated as given in Equation 4.1, whereas the residual for the projector is calculated as $\vec{e} = (u, v)^\top - (p_x, p_y)^\top$ with $(u, v)^\top$ calculated as given in Equation B.6 and $(p_x, p_y)^\top$ denoting the pixel coordinates of the originally projected point. The ceres solver is then applied to solve the minimal least squares problem.

Outlier handling For optimizing the registration result, it is desirable to automatically detect and exclude outliers which can either result from the automatic feature extraction or from erroneous manual annotation, i.e. with the FARO arm or the ARTtrack2 system. There are multiple possibilities for outlier handling:

- *Prior removal*: Before the bundle adjustment problem is modelled, outliers are detected and rejected based on the known geometrical properties of the projected features. If features were e.g. projected in a grid arrangement on a flat surface such as the OR table, the detected features can be mapped onto a plane and checked for consistency with the projected grid.

- *Weight adjustment*: All features are used to construct the bundle adjustment problem. The influence of outliers is mitigated by an according cost function.
- *Statistical removal*: After solving the bundle adjustment problem based on all detected features, features whose residuals fall outside of a certain statistical variation are removed and the bundle adjustment problem is solved again with the smaller feature set.

Contrary to the latter options, prior removal of outliers includes assumptions about the underlying spatial distribution of the detected features. To avoid this, weight adjustment and iterative removal of outliers have been combined, using one standard deviation of the mean as threshold for outlier removal in combination with an absolute threshold of $50\,\mathrm{mm}$.

4.4.6.3. Implementation

The projection based registration has been implemented as a modular system in which a central registration controller node controls an arbitrary amount of camera nodes as depicted in Figure 4.17. Control messages are distributed via a shared control topic and include both control commands and a specific identifier for each projected point.

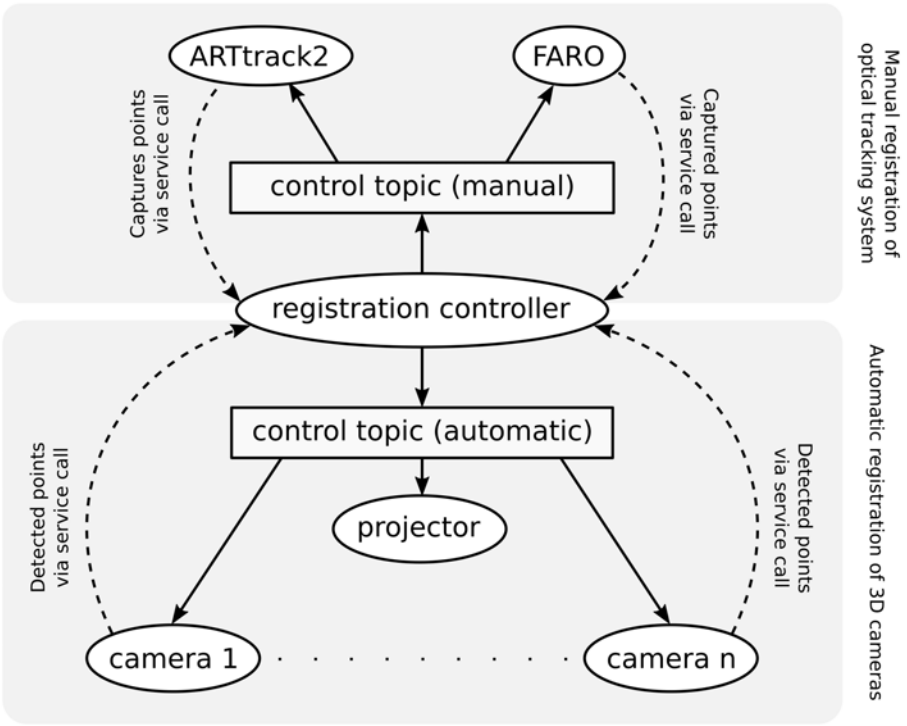


Figure 4.17.: Realization of the registration procedure based on ROS.

Based on this information, each camera node performs feature detection and sends back each detected feature's coordinates to the registration controller via `service calls`. The manual device registration is performed similarly with the difference that the user triggers the service call when the according position is acquired by the device.

Using ROS, the camera nodes are distributed to different servers based on the network topology, e.g. the nodes feature detection for the Kinect v1 cameras are run on the Central Services server in order to keep network load low.

Due to the modular design, no prior knowledge about the amount of cameras or devices to be registered is required; rather, bundle adjustment is performed for all cameras and devices that reported detections.

4.4.7. Forward propagation of semantic labelling

As motivated in section 3.1.4, it is desirable to annotate the first level scene model with semantic information obtained by the second level scene model. This problem can be generalized as follows: Given two independent data streams with a known mapping between each other, where one data stream contains a known *ground truth* in the form of semantic information and the other data stream has desirable characteristics such as a lower latency, higher frame rate or higher robustness, establish a forward propagation of the semantic information. The proposed algorithm can therefore be regarded as a model-free tracking algorithm which is based on a delayed ground truth.[1]

While the proposed algorithm will be detailed with and is based on the supervision scenario given in this thesis, with the PMD camera system acting as the low-latency data stream and the Kinect v1 system providing the labelling i.e. of human tracking, it is not tailored to this application implicitly or explicitly. Rather, as a model-free algorithm that processes an external ground truth, it is designed to be adaptable to applications with different combinations of tracking tasks and modalities with the only requirement that a mapping between the modalities is known.

To stay consistent with the supervision scenario and allow for easier reading, in the following the data source for the ground truth will be named Kinect camera, the source for the low-latency data stream will be named ToF camera and the tracking application will be named human tracking.

[1] This section as well as section 5.2 is based on a paper published in the scope of this thesis, titled *Continuous Pre-Calculation of Human Tracking with Time-delayed Ground-truth* [134], which was presented at the *12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. After selection as one of the best conference papers, it was invited for publication in the *Lecture Notes on Electrical Engineering* by Springer, where a revised and extended version is due to be published with the title *Model-free (Human) Tracking Based on Ground Truth with Time Delay* [135]. Parts are quoted verbatim.

A similar approach to model-free tracking is proposed by Teichman et al. [175] who study model-free tracking with RGBD sensors. Their work focuses on tracking of deformable objects based on initial segmentation provided by the user, with the goal of simplifying the collection of large training data sets for object classification by removing the need for manual annotations on each frame. Contrary to this thesis, their approach makes use of color information for segmentation and forward propagation, which is not applicable to ToF cameras, and does not allow for correction of tracking by a delayed ground truth.

4.4.7.1. Use cases

Based on the general definition of the algorithm given above, different use cases are defined:

- *Latency minimization*: Provide semantic labelling with a lower latency than that of the original data stream.
- *Optimization of tracking robustness*: Continuously provide semantic labelling for each frame, even if the ground truth is intermittently lost.

Both use cases have been realized in the scope of this thesis. For the latency minimization scenario, human tracking provided by one or more Kinect v1 cameras from the second level supervision system is used as ground truth and different pmd[vision] S3 cameras of the first level camera system provide the low latency data stream. Optimization of tracking robustness was evaluated using human tracking provided by the full second level Kinect v1 camera system as intermittent ground truth and the full first level PMD camera system as robust data stream.

A further, third use case is frame rate optimization, where a high-speed ToF camera is used to perform forward calculation of less frequently available ground truth to enable semantic labelling with a higher frame rate. While a realized prototype using a Kinect v2 and an Argos^{3D} P100 showed positive results [135], this was not expanded further.

4.4.7.2. Processing pipelines

In the proposed algorithm, two different processing pipelines are executed in parallel (see Figure 4.18): The *precalculation pipeline* performs processing of the data stream provided by the ToF camera as well as forward propagation of ground truth, which is regularly updated by the second processing pipeline. Therefore, for each incoming frame, a tracking estimation is directly calculated. Human tracking information provided by the Kinect camera is processed in the *ground truth processing pipeline* which updates both the ToF tracking state and a background model based on the ground truth.

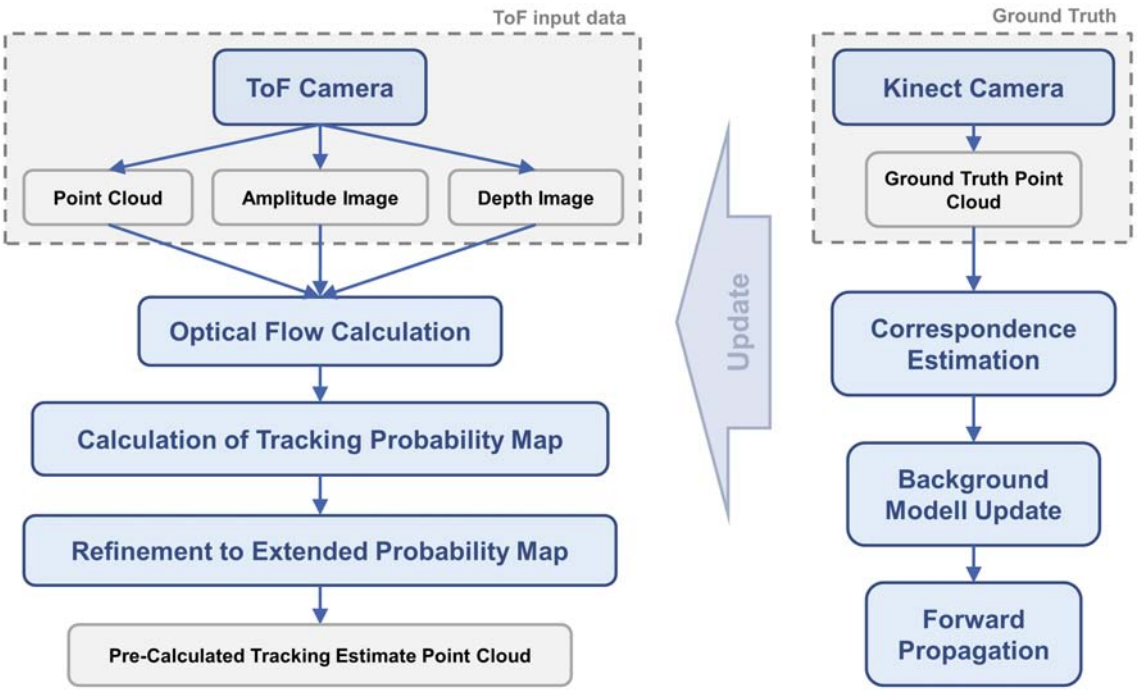


Figure 4.18.: High-level structure of the two processing pipelines: The precalculation pipeline *(left)* processes all ToF data and precalculates a tracking estimation based on the latest available updated ground truth, which is injected by the ground truth processing pipeline *(right)*.

ToF Processing Different types of data are associated with the ToF data in each time step: The source data acquired by the camera, i.e. the 3D point cloud, the amplitude image, the depth image and the acquisition time, as well as information calculated based on source data, such as a flow field, a tracking probability map and geometric information about tracked targets. In the following, all this data will be referred to as *ToF frame*. During processing of each ToF frame, it is necessary to preserve the pixel-to-point correspondences between the 2D image domain and 3D space, e.g. between amplitude image and the point cloud. Therefore, the structure of the point cloud needs to be preserved and no filtering can be applied that alters the original point cloud.

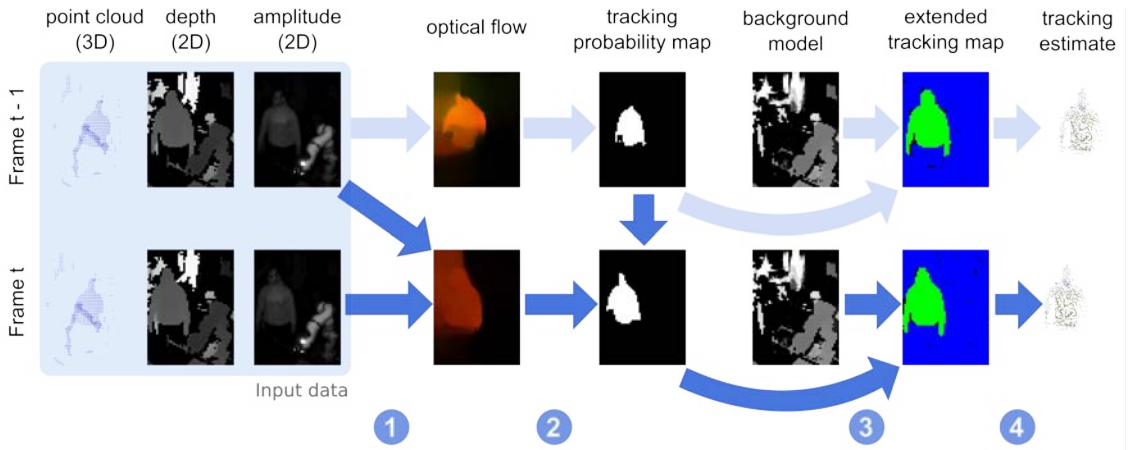


Figure 4.19.: Structure of the precalculation pipeline that processes ToF data to calculate a tracking estimate. *Step 1*: The flow field between the current and previous amplitude map is calculated by 2D optical flow. *Step 2*: Based on the flow field and the previous tracking probability map, a current tracking probability map is calculated. *Step 3*: The tracking probability map is refined using information from the background model as well as geometric and semantic information, resulting in the extended tracking map that contains the estimated semantic labelling. *Step 4*: The extended tracking map can be applied to the original point cloud to segment a point cloud of the user.

The precalculation pipeline for processing ToF data is visualized in Figure 4.19. After transforming the received point cloud into a shared coordinate system and calculating the flow field between the previous and current amplitude map, the preprocessed ToF data is stored into a ring buffer. A first estimate of the current semantic labelling, e.g. tracked target(s), is calculated in image space based on the previous probability map and the flow field. It is then refined using the background model and spatial information encoded in the depth map in order to filter false positives and prevent false negatives. The resulting extended tracking map can then be applied to the original point cloud to segment a point cloud representing the user's body.

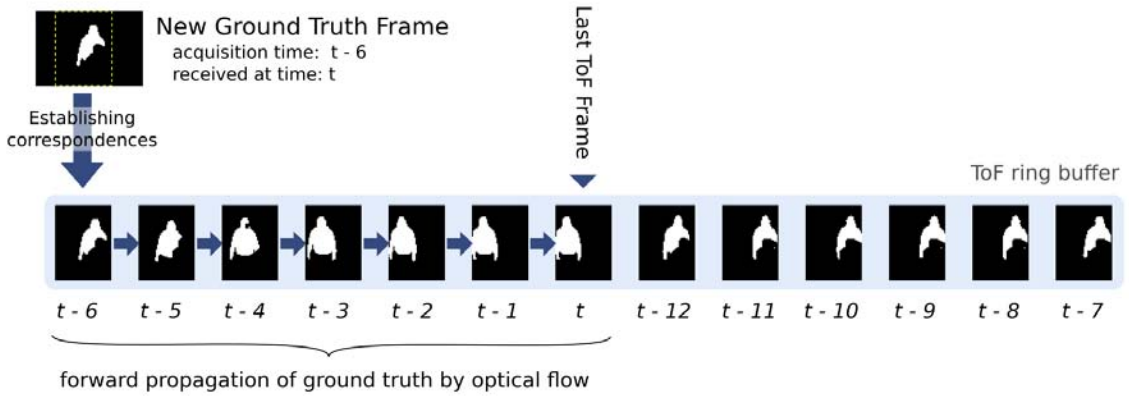


Figure 4.20.: Forward propagation of ground truth in the ToF frame ring buffer. Upon arrival of a new ground truth frame, the corresponding ToF frame is identified and correspondences are calculated. The updated ground truth is then propagated forward to the last received ToF frame based on the stored flow fields.

Ground Truth Processing Like ToF data, incoming point clouds that correspond to tracked humans are first transformed into the shared coordinate system. As depicted in Figure 4.20, the closest matching ToF frame is identified in the ring buffer based on the acquisition time. The previously calculated tracking probability map of the ToF frame is then updated with the known ground truth, based on the correspondences between the ground truth point cloud and the point cloud of the ToF frame. The ToF frame is marked as a *key frame*, as the probability map is now based on the known ground truth for this exact frame, in contrast to being calculated from the probability map of a previous ToF frame. Using the accurate tracking probability map and the corresponding depth data, an update of the background model is then performed. The new information, which has now been introduced in the form of an updated tracking probability map, needs to be propagated forward through the ring buffer in order to be taken into account on the arrival of the next ToF frame. This forward-propagation is iteratively performed based on the flow fields stored in each ToF frame.

4.4.7.3. Background modelling

While most information is stored and processed on a frame-by-frame basis, some information needs to be modelled as global components that represent the state of the scene. One of them is the background model, which is updated whenever new ground truth is processed.

The standard OpenCV implementation of the background model, based on the works of Zivkovic and van der Heijden [199], has been used and extended for taking advantage of the specific data flow of this algorithm. A masking capability has been implemented that allows to restrict the update of the background model

to certain parts of the scene. The updating of the background model was split into two different steps to decouple the update of the background model from the actual background subtraction step.

The extended background model is then used as follows:

- *Model update*: Upon arrival of a new ground truth frame, the background model is updated based on the depth map of the corresponding ToF frame, using the tracking probability map as a mask to prevent the tracked user from being incorporated into the background model. Thereby, the common problem of slowly incorporating non-moving persons or entities into the background model is eliminated.

- *Foreground detection*: Forward propagation of the tracking probability mask based on flow fields often introduces errors, such as false positives. Therefore, a foreground mask is calculated for each new arriving ToF frame by applying background subtraction onto its depth map. During the calculation of the extended tracking map (as depicted in step 3 of Figure 4.19), the foreground mask is then used to filter potential false positives from the tracking probability map.

4.4.7.4. Processing steps

Optical Flow Calculation The original TV-L1 algorithm proposed by Sánchez Pérez et al. [161] was used to calculate optical flow between the amplitude maps of two consecutive ToF frames. In contrast to calculating optical flow based on RGB images, texture of objects does not influence the optical flow calculation based on amplitude images. While this is a disadvantage in applications where an accurate tracing of pixel trajectories is desired, it is of benefit in this algorithm. For example, a uniform object that rotates around its own axis will be detected as non-moving in the amplitude-based flow field, as the previously occluded back of the object consists of the same material as the front and exhibits the same reflectivity. Therefore, the corresponding areas of the probability map stay associated with the whole object, resulting in correct classification.

Tracking Probability Propagation In each ToF frame, the probability that a certain pixel belongs to a tracked human is stored in a 2D probability map. On the arrival of each ToF frame at time t, its probability map m_t is calculated based on the current flow field applied to the previous probability map m_{t-1}. Each pixel p_i in m_{t-1} with a positive probability value is thereby projected onto a new location p'_i in m_t, with the probability value being split onto multiple target pixels based on their L2 distance to p'_i in case p'_i has non-integer coordinates. Further, the total number of tracked targets is stored as part of the global tracking state.

4.4.7.5. Tracking Estimation

While the forward propagation of the tracking probability map by flow fields is an efficient way of estimating the tracking probability map for new ToF frames, it can also become a source of errors as the optical flow calculation parameters need to be balanced between accuracy and speed requirements. Working with low resolution ToF cameras, false negatives have regularly been observed concerning human extremities, such as arms and head, which were lost during forward propagation. On the other hand, false positives have been observed if e.g. a person closely interacts with an object, leading to the object being marked as tracked in the tracking probability map.

Therefore, specific steps have been introduced to first refine the tracking estimate and then reject outliers:

Tracking Refinement Stage The correction of false negatives, such as non-detected extremities, is performed based on the depth map of the ToF frame. Connected probable tracking regions r_i are segmented out of the binarized tracking probability map and their center of mass m_i is calculated. For all m_i, a flood fill operation is then performed on the depth map to connect regions with local continuity in 3D space. Thereby, an extended tracking region r_i' is obtained for each connected region r_i.

Outlier Rejection Stage Both the previous tracking refinement stage and the general forward propagation may have introduced false positive detections. To reject these outliers, the total number of potential tracked regions is checked against the number of tracked objects. If there is a discrepancy, pairwise similarity comparisons are performed between each tracked region of the previous frame and potential tracked regions of the current tracking probability map. Using both 2D and 3D similarity metrics, the best fitting tracked regions are confirmed as tracking estimates. As the last refinement step, background subtraction is performed as described above to filter remaining outliers. The resulting extended tracking map can then be applied to the original point cloud of the ToF scene, resulting in the estimated full body point cloud of the users in the scene.

4.4.7.6. Application to supervision system

The proposed algorithm is designed to operate on data streams from single cameras as it relies on 2D calculations. For a multi-camera system, multiple instances are be executed in parallel. Figure 4.21 shows the system architecture realized for performing full forward propagation of the human tracking information with six PMD cameras: The ground truth, which is obtained from the human tracking system (see section 4.4.4.2), is accessed by six independent instances of the forward propagation algorithm. Each instance processes data from a different ToF camera

and is executed as a standalone ROS node. Each node independently performs forward propagation to estimate a tracking probability map as described above, which can either be used to segment the full body point cloud of the tracked human(s) in the scene or to semantically label the PMD point clouds.

As the result, the fused multi-camera ground truth that represents the user body point cloud perceived by the second level camera system is propagated forward to a fused multi-camera tracking estimation of the user body point cloud based on the first level camera system. In the context of this thesis, this is employed to minimize the latency of tracking and improve the tracking robustness.

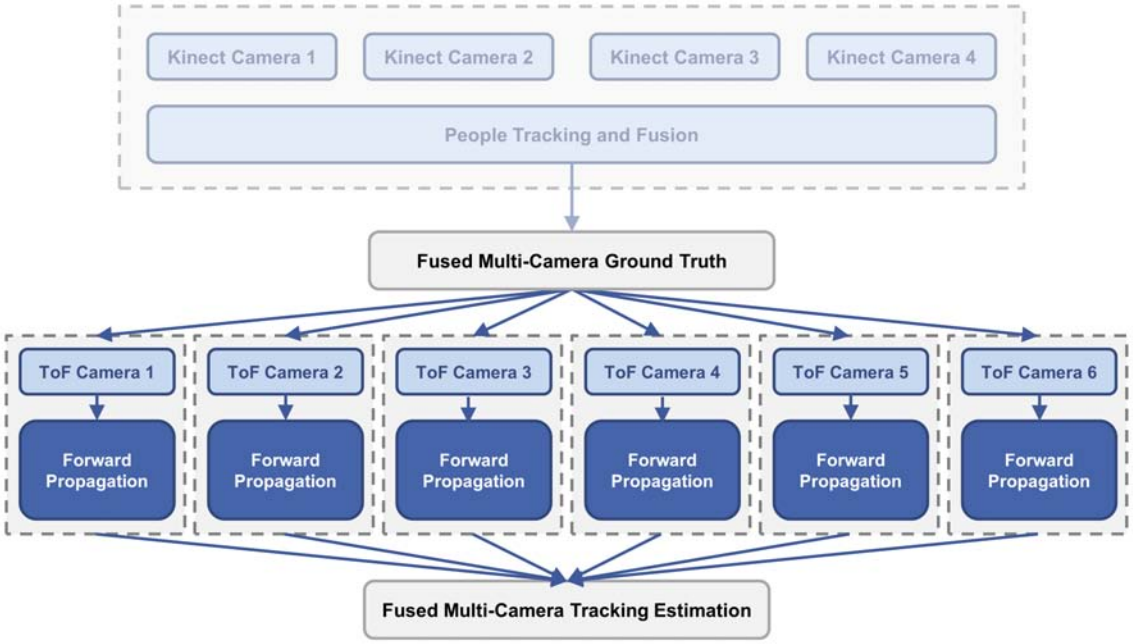


Figure 4.21.: System architecture of the forward propagation for semantic labelling applied to the supervision system.

4.5. Safety concept

4.5.1. Shape Cropping

As the safety features proposed in this thesis are all based on Shape Cropping, the Shape Cropping algorithm needs to support arbitrary shapes, i.e. both convex and concave hulls. Therefore, an efficient implementation is required for calculating multiple instances of Shape Cropping in real time on multiple robot arms.

As basis for the implementation, the `cropHull` class of the Point Cloud Library (PCL) was used which supports testing points for inclusion in arbitrary meshes. It is mathematically based on the ray-triangle intersection method of Sunday [169].

Given a point p and a mesh as a structure of triangles, three rays originating in p with arbitrary directions are checked for intersection with the triangles of the mesh. A majority vote is then taken on the detected odd-numbered ray-triangle-intersections to determine if the point is inside or outside the given mesh: If more than one ray has crossed a triangle an odd number of times, the point lies inside the mesh.

The problem of testing points for inclusion in meshes is a highly parallelizable problem, as each point can be checked at the same time without dependencies between each other. Therefore, the underlying ray-triangle-intersection code was ported to CUDA for parallel execution on the graphics card to enable real-time processing. As the bandwidth between host memory and graphics card memory is a bottleneck for many General-Purpose Computing on Graphics Processing Unit (GPGPU) tasks, transfers between both were minimized. The meshes that represent the hulls used for Shape Cropping are transferred to GPU memory only once, at the initialization of the algorithm, and later transformed in memory based on the current robot pose. As the point cloud to which Shape Cropping shall be applied changes between each execution of the algorithm in most scenarios, it needs to be transferred to GPU memory in each iteration. The result is transferred back to host memory as the number of ray crossings per point.

4.5.2. Robot localization

To determine the location of a robot arm in the scene, first its position is estimated coarsely using either active or passive initial localization (see Figure 3.4). The detected pose is then iteratively optimized to yield the final detection.

4.5.2.1. Initial localization

Passive Passive robot localization does not require any motion sequences to be performed by the robot arm for the initial localization. Rather, it is based on landmarks in the scene that are detected first and then serve to reduce the size of the search space. In case of robots that are mounted to the OR table, such as the OP:Sense system, the MiroSurge system or other systems presented in section 2.2.1.4, the OR table itself can be used as such a landmark. In the following, general knowledge about the camera positions with respect to the OR table is assumed, such as the fact that cameras are ceiling-mounted around the OR table as in the realized supervision system.

For real applications, it is expected that an OR table with a stationary column is used and that information about the table configuration, such as height, rotation, longitudinal and transversal shift and angle of segments, can be accessed via according interfaces. In this case, the search space can be restricted based on accurate knowledge which is already available. For the laboratory experiments and evaluation, in which a mobile OR table was used that does not offer access

to said parameters, a custom detection of the OR table pose based on its surface geometry was implemented as follows:

1. The table surface plane is detected by Random Sample Consensus (RANSAC).
2. By performing Principal Component Analysis (PCA) on all inliers of the detected plane, the orientation of the OR table surface is calculated (with an ambiguity of $180°$ in the rotation around the plane's normal due to the symmetry of the OR table).
3. All inliers are projected onto the plane and its outer contour is detected by fitting a minimum-sized rectangle in 2D space.
4. Based on the surface plane parameters and the table boundaries, the pose of the OR table is calculated.

For robots mounted at the sides of the OR table, this defines two box-shaped ROIs at both sides of the OR table to which the search space for the robot base position can be restricted. To detect potential robot base positions, the ROIs are extracted from the scene point cloud and Euclidean Clustering based on Kd-Trees is performed. Detected clusters are projected to 2D space, where they are analyzed for correlation with the cross-section of the outline of the robot base link.

As the robot arms are rigidly attached to the OR table, their relative orientation R_{rel} w.r.t. to the OR table can be assumed static and known. Therefore, the orientation of each robot base R_{robot} can be calculated based on the table orientation R_{table} as $R_{robot} = R_{rel} \cdot R_{table}$, but is still ambiguous to a $180°$ rotation around the normal of the OR table surface plane.

If multiple robot arms are attached to an OR table, only one iteration of passive localization is necessary to detect all robot bases. However, the cross-section-based detection does not provide any information as to which detected robot base corresponds to which robot arm.

Active It cannot be assumed that robot localization based on landmarks is possible in all situations. Therefore, an active localization approach has been developed where each robot is detected based on a short motion sequence it performs. Concerning the system architecture (see section 4.3.4), this requires an interface to the robot through which either the robot can be controlled on joint-level or a pre-defined motion sequence can be triggered.

Active robot localization consists of the following main steps for each robot arm:

1. A pre-defined motion sequence is performed by the robot arm. For the LBRs used in this work, this motion is defined as an oscillating motion in the second joint of the robot, starting from an upright pose. In regular intervals, an octree representation of the scene is calculated and compared to the previous one. The detected differences are added to a point cloud p.

2. PCA is performed on p to calculate its two major components. The first component directly corresponds to the upward axis of the robot. The robot position can then be estimated based on the shape and position of p together with the known geometric properties of the robot and the known joint orientations for which p was detected.

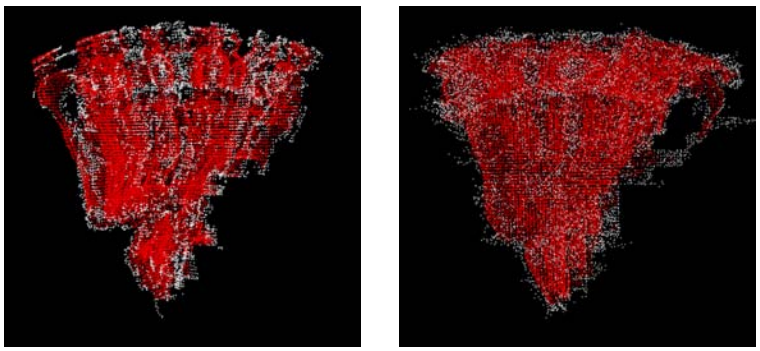


Figure 4.22.: Spatial change point cloud showing the accumulated point clouds of a LBR acquired during a continuous sweeping motion. In the point cloud acquired by the Kinect v1 camera system *(left)*, the shape of the LBR is clearly recognizable, whereas it appears fully blurred as captured by the Kinect v2 system *(right)*.

Figure 4.22 shows a visualization of the spatial change clouds obtained with the Kinect v1 and Kinect v2 camera systems.

Contrary to passive localization, active localization can be performed in multiple passes in order to unambiguously detect each robot arm. Alternatively, it is also possible to detect multiple robot bases in one pass, which again results in an ambiguity concerning the mapping of detected robot bases to the corresponding robot arms.

4.5.2.2. Localization optimization

Both methods for initial localization provide an estimation of the robot position which is based on partial information only, i.e. the cross-section of the base segment or the "blurred" spatial change information. Therefore, localization optimization is performed which leverages the full shape of the robot in a given pose for refining the localization, recovering the ambiguity w.r.t. its orientation and matching the detected robot bases to the corresponding robots.

Based on information provided by Shape Cropping, i.e. the number of inliers and outliers as defined in section 3.2.3, different criteria are proposed for assessing the quality of a pose estimation in a given scene. The total number of inliers $\#in$ and outliers $\#out$, both accumulated over all segments of the robot, are employed as a measure for quantifying the overall validity of a detection. If a detection is valid,

e.g. $\#in$ is higher than a specified threshold, the inlier to outlier ratio r_{io} is used as a qualitative criterion:

$$r_{io} = \frac{\#in}{\#out + 1}. \tag{4.2}$$

Estimation refinement In order to refine an estimated robot base pose $p_{initial}$, a grid of positions in a spherical neighborhood of $p_{initial}$ is sampled by applying shape cropping to each position. This yields up to three positions that each optimize one of the criteria, i.e. maximize $\#in$ or r_{io} or minimize $\#out$. If all poses are distinct and fulfill additional constraints, such as surpassing a threshold of $\#in$, geometrical averaging of the remaining positions is then used to determine a final position.

This sampling process is repeated multiple times with decreasing search radii and inter-position distances, resulting in a refined pose estimate.

Outlier-based correction If the initial localization results in a high amount of outliers, this can be the effect of a slight misdetection. If, for example, the robot pose is estimated too far away from the camera, parts of the surface of the robot will register as outliers among all segments. Therefore, an analysis of the spatial distribution of all outliers is performed per segment. Using Euclidean clustering of the outliers, a displacement vector between the center of the according segment and the center of the outliers is calculated. The estimated robot pose is then shifted by the calculated displacement vector. Both resulting positions are rated by Shape Cropping and compared to the initial one. The best rated position is kept as the new estimation.

Single camera optimization As the estimation refinement step detailed above results in poses that maximize $\#in$ and minimize $\#out$, problems can arise in scenes acquired from a single viewpoint. As only the "front" of the robot is represented in the scene due to the 2.5D nature of the sensor, it is possible that the whole surface of the robot is placed in the center of the inner hull without incurring the penalty of high $\#out$, as the back side is not visible. Therefore, positions that are estimated too close to the camera may be rated better than the correct position if Shape Cropping is performed on a point cloud acquired by a single camera.

This can be corrected by minimizing the distances between the inner hull of the robot and all inlier and outlier points. As the misdetection results from an incorrect distance between camera and robot, a vector between the camera and the current robot position is established. The position of the robot is then optimized along this vector, using the sum of distances as the optimization criterion.

Sequence of optimization Based on the specific optimization steps presented above, the localization optimization can be performed either globally, based on a full scene model that contains the fused point clouds of all cameras, or locally on a per camera basis. Local optimization has the benefit of providing a robot base pose estimation per camera, which can be leveraged for performing consistency checks on the camera registration. If e.g. base poses estimated by one camera show a consistent offset in comparison to the estimations by the other cameras, this is an indication for a miscalibration. The local, per camera optimization is only applicable if it can be expected that multiple segments of the robot are in the field of view of each camera. This requires that the FoV of the cameras is large and overlapping, as is the case with the Kinect cameras in the supervision system. In case of the PMD cameras, both the narrow FoV and the lack of overlap prevent local optimization.

The per camera optimization is performed as follows:

1. For each camera:

   1.1 Perform estimation refinement to calculate r_{io} for the pose estimation provided by initial localization. In case of a low result, apply outlier-based correction, followed by estimation refinement.

   1.2 Rotate estimated robot base pose by $180°$ around its upward axis.

   1.3 Perform estimation refinement to calculate r_{io}. In case of a low result, apply outlier-based correction, followed by estimation refinement.

2. Decide on the correct orientation by taking a majority vote of all cameras, set correct orientation for all further processing.

3. For each camera: Perform single camera optimization.

4. *Optionally:* Set the robot to different joint configurations and repeat step 1 for each configuration.

5. Calculate the final robot base pose based on the results obtained by all cameras in all iterations.

If different robot configurations are evaluated during step 4, multiple iterations of Shape Cropping are performed that each provide a position estimate. The validity and quality of these estimates can vary, i.e. due to the visibility of the robot in a given pose. Therefore, a weighted approach is employed to derive the final position p_f based on n individual position estimates p_i. To take into account both the validity and quality of each estimate, the weight w_i of each estimate is calculated as the product of the respective number of inliers $\#in_i$ and inlier to outlier ratio $r_{io,i}$:

$$w_i = \#in_i \cdot r_{io,i}. \tag{4.3}$$

The final position p_f is then calculated as

$$p_f = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \cdot p_i. \tag{4.4}$$

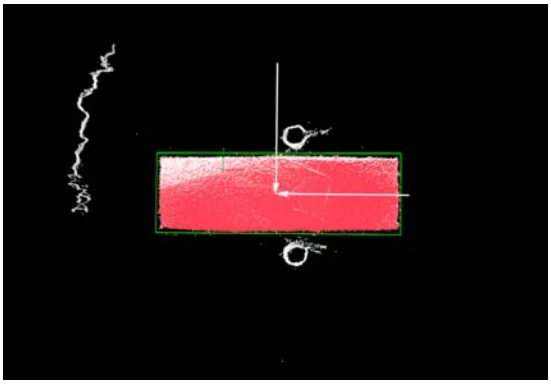
Figure 4.23 illustrates the single steps of a passive localization and subsequent optimization, based on data acquired by the Kinect v2 camera system.

4.5.3. Detection of impending collisions

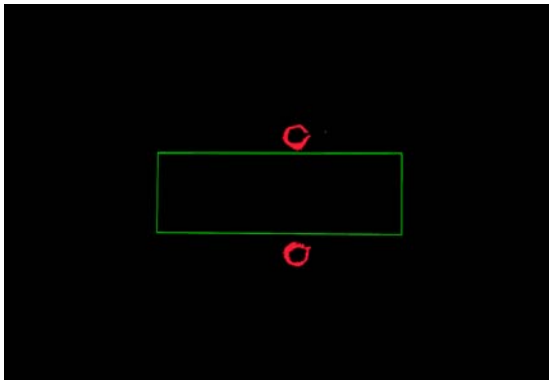
As detailed in section 3.2.4.2, impending collisions are detected based on violations of the safety zone by outside objects. One obvious criterion for detecting such violations is again the ratio r_{io} of inliers to outliers as given in Equation 4.2, which decreases as soon as the safety zone is violated. However, r_{io} is only viable as a single criterion in situations where only discrete measurements need to be evaluated and visibility of the robot for all cameras remains the same. If a temporal analysis is performed and the robot is moving, r_{io} can be affected e.g. when a segment of the robot moves out of the FoV of one camera, as the number of inliers $\#in$ drops, even if the number of outliers $\#out$ stays equal.

Therefore, the absolute number of outliers per segment is employed as the main criterion for detection of impending collisions. $\#out$ is monitored and a moving average is calculated over multiple frames. If the number of outliers abruptly increases by over $20\,\%$ as compared to the moving average, a potential collision is flagged for further checking. Since small sample sizes are more prone to variations caused by e.g. noise or other interferences, the number of outliers in the current frame needs to surpass a fixed threshold. If this is the case, all detected outliers are classified using Euclidean clustering of the spatial neighbourhood of the robot. If they belong to an outside object, an impending collision is detected and an appropriate reaction can be triggered.

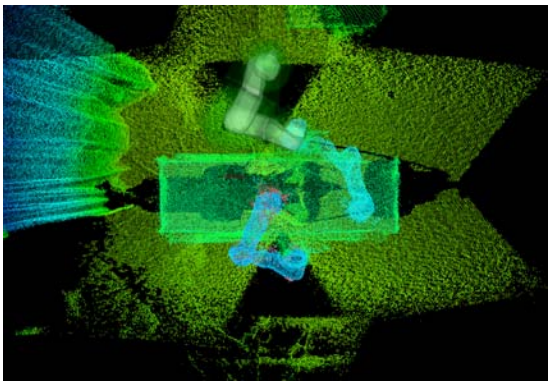
While this method is directly applicable to static robot poses, there is a major drawback when applied directly to a moving robot: If the robot is moving and the latency of the camera system is different from the latency with which robot pose data is received, the virtual inner and outer hull are not synchronized to the robot pose as perceived by the camera system. This is due to the fact that two sources of information about the scene from different times are being mixed into one representation: The safety zone is constructed based on the current robot pose, which is usually available with a low latency in the range of $\leq 1\,\mathrm{ms}$, whereas the scene, which is segmented by Shape Cropping, has a higher latency, depending on the concrete embodiment of the camera system. This leads to the occurrence of "phantom collisions": A part of the points that belong to the robot are not segmented into the inner hull of the safety zone, but register as outliers, while objects directly in front of the robot would be classified incorrectly as inliers.



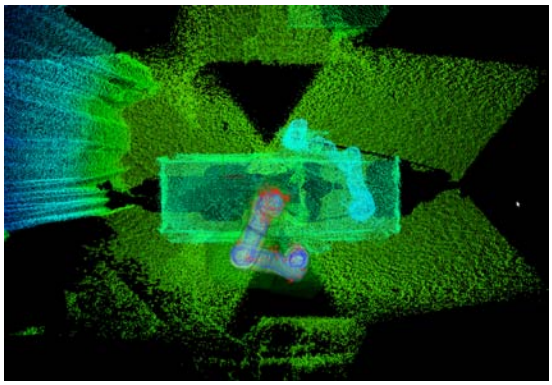
(a) Detected surface and orientation of OR table.



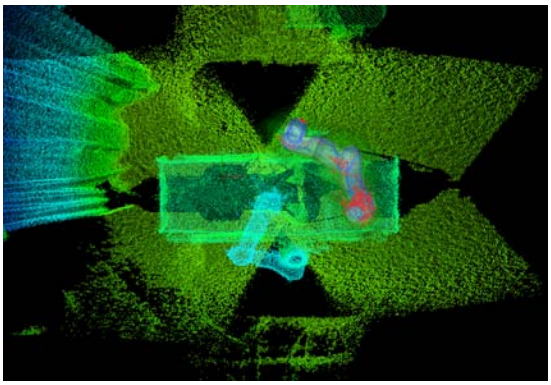
(b) Cross section of detected robot bases.



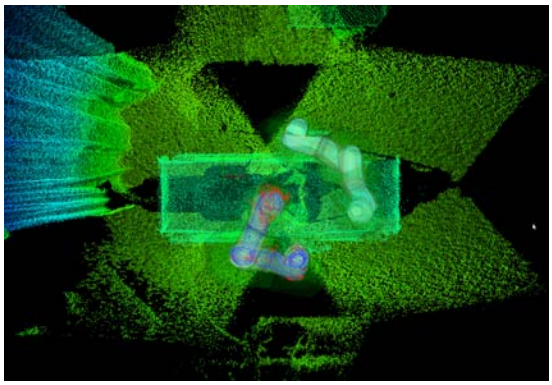
(c) Incorrect matching of detected base to lower robot arm.



(d) Correct matching of detected base to lower robot arm.



(e) Pose optimization for upper robot arm.



(f) Final result with both robot arms correctly matched and optimized.

Figure 4.23.: Steps of passive localization performed with Kinect v2 camera system. For image (c) to (f), green points depict the virtual scene, blue points inside the robot hull are inliers of the currently estimated robot pose and red points are outliers.

To solve this, a temporal separation of the inner hull of the safety zone and its outer hull is proposed. Based on the known time delay between reception of robot pose data and availability of the fused point cloud provided by the supervision system, the inner hull is constructed using delayed robot pose data which is synchronized to the analysed point cloud. The outer hull of the safety zone is constructed based on the most recent robot pose data to keep the safety zone centered around the robot's pose in the real scene.

4.5.4. Continuous pose supervision

The daVinci™ system, which is currently the only system for MIRS that is commercially available and approved by the FDA, features internal redundancy for safeguarding the correct operation of the robot arms. Specifically, the angular position of each joint is determined by two different sensors, the encoder and a potentiometer. A safety warning is triggered if a differential change is detected between both sensors [184]. However, the Manufacturer and User Facility Device Experience (MAUDE) database lists at least one report of a case where the surgeon appears to have misinterpreted the safety warning and carried on with the intervention multiple times [180]. MAUDE further contains multiple reports of cases in which patients have been injured by involuntary motion of the robot arm without a prior warning of the daVinci™ safety system [182, 183], sometimes even described as a "stabbing motion" [181].

It cannot be known if these instances were preceded by faulty joint actuation of the robot arm and/or if the motions deviated from the control commands that were sent to the robot. If these were the case, the issue could have been recognized early and therefore prevented by continuous supervision of the robot's pose, which would allow the supervision system to detect such incorrect motions and issue e.g. an emergency stop. Either way, the continuous supervision of the correct pose of each robot arm complements the internal sensors of the robot and adds an independent, additional safety layer.

In the proposed system, continuous pose supervision is performed the same way as detection of impending collisions: By applying shape cropping to the current scene, safety zone violations are detected. If Euclidean clustering shows that a violation is caused by the robot itself, i.e. all outliers are connected to the same spatial region as the robot, this results in the detection of an incorrect robot pose. As this is technically identical to the detection of impending collisions, it is realized in the same code path, i.e. transfer from the current scene to GPU memory and Shape cropping only need to be applied once. Therefore, no performance penalty is incurred for continuous pose supervision.

4.6. Feedback to OR personnel

4.6.1. Physical setup

Within the spatial constraints of both the OP:Sense setup and real ORs, realization of spatial augmented reality on a projection area that covers the full OR table requires either a wide-angle projector or a system consisting of multiple projectors with a standard projection ratio. The use of multiple projectors has advantages in terms of preventing occlusions and shadows caused e.g. by OR personnel or robot arms, but the technical feasibility of actual installation in an OR is doubtful at best due to the space requirements and obstructions caused for ceiling-mounted equipment such as the OR lamp. Using a single short-throw projector considerably reduces the space requirements while allowing for a wide projection surface.

Due to spatial constraints in the OP:Sense setup, the projector had to be set up in a sideways configuration and combined with a deflection mirror at a $45°$ angle. As the mirror surface is planar, this only inverts the y-axis of the projection and does not alter the optical properties of the projection system. A special front reflecting mirror was selected to avoid double projections that are inherent to projection systems that employ standard mirrors.

4.6.2. Software implementation

As the proposed system is realized as a distributed system, the projection system needs to allow for multiple sources to perform reactive projection onto the scene, while at the same time not requiring excessive network bandwidth. Conceptually, this requires to split projections into a high-level part, i.e. description of abstract graphical aspects of a projection that signals information to the user, and a low-level part, i.e. the concrete rendering of the output to the projector based on the high-level information.

This was realized by implementing a central projection node that can take input by arbitrary ROS nodes in the OP:Sense system. The projection node is executed on one of the SFF PCs of the Kinect v1 camera system to which the projector is connected due to the spatial proximity. The projection node is based on openFrameworks for display purposes and fully integrated into the ROS system in order to access current information about the state of the system as shown in Figure 4.24. It offers two distinct methods by which other nodes can provide information to be projected into the scene:

- *SVG-based*: Client nodes send a string containing SVG markup to the projection node. This is rendered and projected onto the scene until new SVG markup is received by the projection node. To allow for dynamic projection without requiring continuous updates of the SVG data by the client node, an extension to SVG markup has been implemented that allows to directly

use a `frame_id` of a named entity in ROS as coordinates in SVG. The SVG markup is then continuously evaluated by the projection node by querying the according transformations from the tf-tree and creating and rendering the resulting graphics.

- *Cloud-based*: Client nodes provide point clouds whose outlines shall be projected onto the scene. Point clouds are sent as native `PointCloud2` messages whose `frame_id`s are extended by additional, custom projection markup that specifies the graphical properties of the projection. This includes e.g. color, animations like fading and blinking as well as the possibility to define a fixed time for the projection after which it is to be removed. Within the projection node, all received clouds are stored, evaluated and rendered in each frame based on their graphical properties. Client nodes can update the point clouds without breaking the animation cycle.

For both projection methods, rendering refers to generating a geometrically correct image for the projection that takes into account the specified information (e.g. SVG markup) and the intrinsic and extrinsic parameters of the projector.

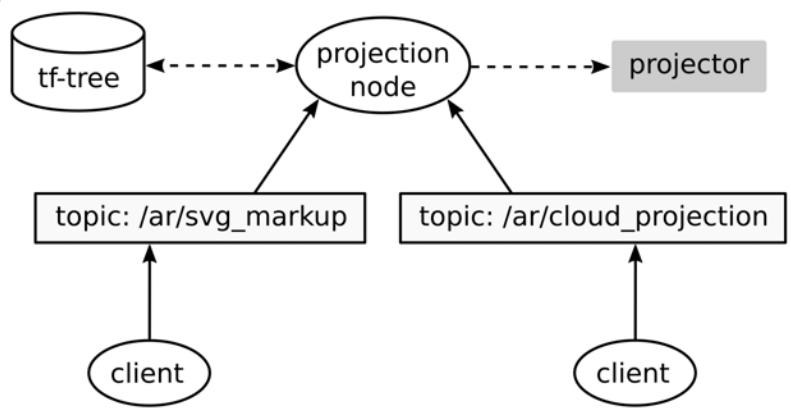


Figure 4.24.: ROS-based implementation of the projection system: Multiple client nodes can publish high-level descriptions of the projection, which are interpreted and rendered by the projection node based on spatial information provided on the tf-tree.

A major consequence of this design is that it allows for a modular system setup with light-weight client nodes, as each client node only needs to provide a high-level description of the desired projection, based on information available within the node. This description only needs to be sent once to the projection node. Therefore, it is possible, but not required for clients to store an internal state of the projection and regularly update it.

4.6.3. Vertical surface mapping

When a projection is performed onto an unknown scene, different kinds of distortion can occur which are related to the geometric properties and to the color distribution of the scene as well as to the viewpoint of the observer. Geometric distortions are caused by projecting onto any surface which is non-planar and not perpendicular to the projection system. They can be eliminated by pre-warping the projected image based on the geometry of the projection surface. For example, in case of planar projection surfaces, a simple homography can be calculated between the projection surface and the projector system. For arbitrarily-shaped surfaces, more involved correction functions have been proposed in literature [13, 141].

However, an observer-independent projection onto arbitrarily shaped surfaces is generally not possible because the perceived geometric distortion directly depends on the viewpoint of the observer. Figure 4.25 illustrates this based on a minimally invasive scenario: The direct lines of sight of each observer to the instrument tips intersect the patient's body surface at different, observer-dependent points. For direct projection, each intersection again differs from the position to which the instrument tips are projected on the patients body.

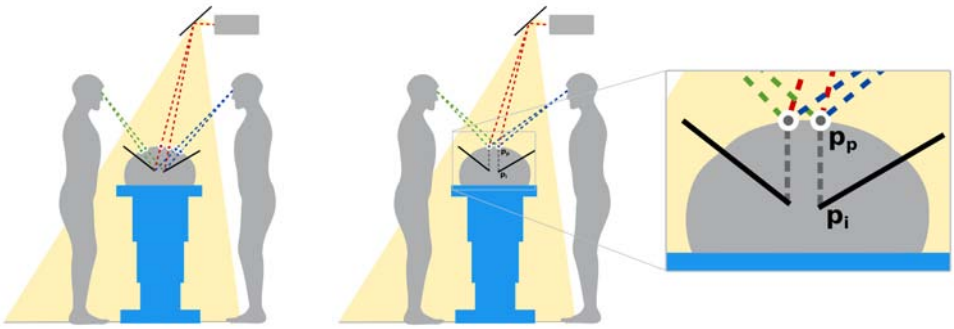


Figure 4.25.: Projection of laparoscopic instruments onto the patient's body surface. *Left*: With direct projection, the projected positions (red) differ from the perspectively correct positions for both observers (blue and green); *right*: Vertical mapping of the instrument tips to the body surface before projection results in a consistent representation which is visible from different viewpoints.

As one of the goals for employing SAR within this thesis is to facilitate a shared situation awareness, multiple observers, such as the OR personnel, need to be able to perceive the projection in an intuitive way. Therefore, a vertical mapping of the instrument tips onto the patient's body is proposed that facilitates an intuitive, shared understanding of the projected positions. Based on the known 3D surface geometry of the patient's body as acquired by the camera system, which can be seen as a 2D height function $z_s(x, y)$, all points of interest $p_i = (x_i, y_i, z_i)$ inside of the patient's body are mapped vertically to corresponding projection points

$p_p = (x_p, y_p, z_p)$ on the surface of the patient's body by correcting their height:

$$z_p = \begin{cases} z_s, & \text{if } z_i \leq z_s(x_i, y_i) \\ z_i, & \text{otherwise} \end{cases}$$

This results in a faithful representation of the spatial relations which is independent of the projector position, consistent among all observers and easy to interpret. For different types of interventions or angles of the OR table, the same principle can be applied by adapting the projection directions, e.g. using the normal of the surface of the OR table.

4.6.4. Attention direction

In case of adverse events, the system needs to be able to capture and direct the attention of the OR personnel to important locations, e.g. such as the location of an impending collision between robot and patient. Factors that influence the ability of a system to attract attention of humans have been studied in several works, of which Green gives an extensive account [45]. Two main concepts of attention are *visibility* and *conspicuity*:

- *Visibility* states if a viewer is in principle able to detect a sensation, but it does not imply whether they will notice it.
- *Conspicuity* is defined as "the likelihood that a viewer will notice and perceive visible information" [45].

The main potential sensory conspicuity cues are (i) color and (ii) flicker and motion. While red has long been regarded as the most conspicuous color, different studies have shown that yellow and yellow green are more conspicuous in different scenarios. Regarding temporal aspects, which are directly applicable to projected optical cues, the main variables are the properties of the modulation. This includes the modulation amplitude, i.e. the difference between between highest and lowest brightness, the modulation frequency and the modulation waveform. Based ono different studies, Green estimates that a combination of a maximal modulation with a frequency of $3\,\mathrm{Hz}$ - $5\,\mathrm{Hz}$ and a sharp offset of the waveform will exhibit the highest conspicuity.

4.6.5. Features

In subsection 3.3.3, three main applications for providing feedback to the OR personnel have been identified: (i) visualizing the robot state, (ii) visualizing the instrument positions in MIRS and (iii) drawing attention to adverse events.

These are implemented as follows:

- *Visualization of robot state*: The different states of the robots are visualized based on the ISO 22324 guidelines for color-coded alerts, which correspond to a common interpretation of colors in safety-related scenarios. Specifically, the following combinations of colors and temporal cues have been used:
  - Green: The robot is in hands-on mode. It can be approached and manually moved without risk.
  - Yellow: The robot is controlled remotely by the surgeon, i.e. the clutch is pressed. The robot can move at any time.
  - Red: The robot is in autonomous mode and is or will be performing a motion immediately. It is unsafe to approach the robot.
  - Gray: The robot arm is in a static mode and cannot move. It is safe to approach the robot.
- *Augmentation of surgical instruments in MIRS*: The poses of laparoscopic instruments and the endoscopic camera are dynamically projected onto the patient body. Additionally, the frustum of the endoscopic camera is projected to allow for intuitive understanding of the position of the instrument tips relative to the camera. This can be further assisted by optionally superimposing the instrument tip with a circle that dynamically adjusts its size depending on the distance between instrument tip and abdominal wall. The distance to the optical axis of the endoscopic camera can be visualized by color-coding for instrument insertion.
- *Feedback in case of adverse events*: If an adverse event is detected, the attention of the user needs to be attracted and directed to a specific location. This is performed by projecting a shape that corresponds e.g. to the outline of a potentially colliding object. An evaluation of the effectiveness of attracting the users attention to arbitrary positions using visual cues is presented in subsection 4.6.4.

5. Results

5.1. Supervision system

5.1.1. Interference analysis

To analyze interferences between different types of 3D cameras operating in the same volume, i.e. [pmd]vision S3, [pmd]vision CamCube 2.0, Kinect v1 and ARTtrack2 OTS, a test bed was set up as shown in Figure 5.1. A [pmd]vision S3, a [pmd]vision CamCube 2.0 and a Kinect v1 were mounted to a rigid support frame on a measurement table. Further cameras, i.e. the ARTtrack2 system and five additional [pmd]vision S3 cameras, were ceiling-mounted above the test bed. A white board was mounted as a planar surface to an LBR and mechanically coupled to the FARO measurement arm. With the distance between the board and the support frame kept static, the cameras were switched on in different combinations and the distance to the board as measured at the center pixel was recorded with each camera for 100 iterations. The distance between the surface of the camera lens and the white board was measured and annotated.

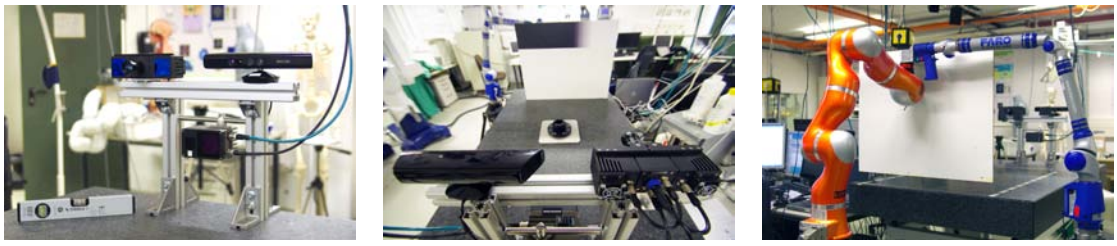


Figure 5.1.: Test bed for analyzing interferences between [pmd]vision S3, [pmd]vision CamCube 2.0, Kinect v1 and ARTtrack2.

Table 5.1 shows the numerical results for the conducted measurements. Plots for the according data are given in Appendix A. With exception of the [pmd]vision S3, it can be seen that the measured variances per camera only slightly differ between measurements with different combinations of camera systems turned on. For the [pmd]vision S3, the magnitude of variance increases by two orders of magnitude when other [pmd]vision S3 cameras are operated in the same volume at the same modulation frequency. Applying the time- and frequency-multiplexing as detailed in section 4.4.3.2 remedies the interferences.

	S3	CamCube	Kinect v1
standalone	$1.6\,\mathrm{mm}^2$	$10.0\,\mathrm{mm}^2$	$2.1\,\mathrm{mm}^2$
with ARTtrack2	$1.7\,\mathrm{mm}^2$	$7.6\,\mathrm{mm}^2$	$0.2\,\mathrm{mm}^2$
with S3	–	$9.4\,\mathrm{mm}^2$	$2.8\,\mathrm{mm}^2$
with CamCube	$1.6\,\mathrm{mm}^2$	–	$0.2\,\mathrm{mm}^2$
with Kinect v1	$1.8\,\mathrm{mm}^2$	$11.3\,\mathrm{mm}^2$	–
with unmultiplexed S3	$377.6\,\mathrm{mm}^2$	$9.2\,\mathrm{mm}^2$	$0.8\,\mathrm{mm}^2$
with multiplexed S3	$2.7\,\mathrm{mm}^2$	$9.8\,\mathrm{mm}^2$	$3.7\,\mathrm{mm}^2$
with all cameras	$3.1\,\mathrm{mm}^2$	$11.3\,\mathrm{mm}^2$	$1.5\,\mathrm{mm}^2$

Table 5.1.: Variance over 100 distance measurements at central camera pixel, carried out with [pmd]vision S3, [pmd]vision CamCube 2.0 and Kinect v1 with different combinations of other IR-emitting camera systems active in the same woking volume.

As result, it can be seen that while different combinations of cameras influence the variance, the magnitude of the difference is less or equal than the depth resolution of the corresponding camera. Therefore, it is viable to operate the different subsystems of the proposed supervision system in a common working volume. The multiplexing scheme detailed in section 4.4.3.2 is shown to eliminate crosstalk between the PMD cameras.

Concerning potential interferences between multiple kinect v1 cameras and potential side-effects caused by e.g. usb-cable length, which is not applicable to the proposed setup, extensive evaluations have since been performed by Lemkens et al. [105].

5.1.2. Projection-based registration

To enable fusion of scene information acquired by different devices, such as the 3D cameras and the OTS, all devices need to be registered to a common coordinate frame. This thesis proposes a projection-based registration method where features are projected by visible light into the scene, onto the surface of an OR table, and detected by the different camera systems as discussed and depicted in section 4.4.6. In the following, the concrete evaluation procedure is described as well as the obtained results.

5.1.2.1. Evaluation procedure

The projection-based registration was carried out for all camera systems and the ARTtrack2. The FARO measurement arm with a certified accuracy of $0.026\,\mathrm{mm}$ and a standard deviation of $2\sigma = 0.010\,0\,\mathrm{mm}$ after calibration of the measurement probe was used as ground truth. Five iterations were performed with 27 features projected in each iteration. To obtain results that are realistic for a clinical setting,

an OR table covered with table cloth was used as projection surface without any further modifications. Iterations were performed at table heights of $0.73\,\mathrm{m}$, $0.88\,\mathrm{m}$, $1.03\,\mathrm{m}$, $1.18\,\mathrm{m}$ and $1.33\,\mathrm{m}$.

The PMD cameras used in this thesis feature a so-called Suppression of Backlight Illumination (SBI) which cannot be disabled. By default, only an amplitude image can be obtained, which contains pixelwise information about the strength of the reflected IR signal sent out by the camera. However, it contains no information about the ambient light present in the scene. To overcome this restriction, the four phase images were extracted from the raw data of the [pmd]vision S3 cameras and analyzed for susceptibility to ambient light. However, it has been found that the hardware-based SBI prevents ambient light information from being included even in the raw data. Detection of projected features was therefore not feasible with the PMD cameras.
As an alternative solution, detection was instead performed manually by simulating the projected features based on the amplitude map. In the manual acquisition stage, a non-reflecting circle of cloth with the same diameter as the projected features was positioned in the location of each feature and the detection of the according feature with the PMD cameras was triggered manually.

All systems were switched on for the full registration procedure, except for the Kinect v2 and PMD camera systems which were switched on alternately to prevent crosstalk. The registration procedure consisted of the following steps for each iteration:

1. Start Kinect v2 camera system, stop triggering PMD camera system.
2. Interactively determine projection area.
3. Start automatic feature detection by Kinect camera systems.
4. Enable PMD camera system in quality mode, stop Kinect v2 camera system.
5. Manually annotate projected points by ARTtrack2, FARO and PMD camera system.
6. Adjust OR table height for next iteration.

For evaluating the accuracy of the obtained registration, two different measures have been employed: the registration error based on the feature positions that are reconstructed by the bundle adjustment algorithm and the registration error w.r.t. the ground truth feature locations obtained with the FARO measurement arm. To prevent measurements from the FARO arm from influencing the bundle adjustment, bundle adjustment was first performed without inclusion of data obtained by the FARO arm. The FARO was then registered to the reconstructed feature positions in a separate step, providing a transformation between the original registration results and ground truth captured by the FARO arm. The reconstructed camera and feature positions are shown in Figure 5.2.

Figure 5.2.: Reconstructed origins of the coordinate system of different devices and locations of features that were projected on the surface of an OR table at five different heights, visualized from two different perspectives.

Evaluation set	PMD	Kinect v1	Kinect v2	ART	FARO	Projector
P	×					×
K_1		×				×
K_2			×			×
PA	×			×		×
PK_1	×	×				×
PK_1A	×	×		×		×
PK_1K_2	×	×	×			×
PK_1K_2A	×	×	×	×		×
PK_1K_2FA	×	×	×	×	×	×

Table 5.2.: Overview of evaluation sets for registration accuracy evaluation. The leftmost column lists the name of the evaluation set, crosses represent the inclusion of a camera system in the respective evaluation set.

Registration accuracy was evaluated separately for different evaluation sets that each consist of a subset of the available cameras, as listed in Table 5.2. Per evaluation set, three evaluations have been performed: The *initial* and *local* registration error per camera are calculated as Euclidean distances between the detected features and the features reconstructed by bundle adjustment, using all detected features (*initial*) or the features remaining after two iterations of outlier removal (*local*) as presented in section 4.4.6.2. The *global* registration error is also calculated based on the results of double outlier removal, but w.r.t. the ground truth obtained by FARO. Table 5.3 lists the different combinations used for evaluation.

	initial	local	global
All features	×		
2 iterations of outlier removal		×	×
W.r.t. reconstructed features	×	×	
W.r.t. ground truth			×

Table 5.3.: Naming convention for evaluated combinations.

5.1.2.2. Registration results

For all following boxplots, the red line represents the median and the box borders represent the lower and upper quartile, between which 50% of all measurements are located. The whiskers are drawn in default MATLAB style that corresponds to the definition given by Tukey, i.e. whiskers are extended to the next value outside the Interquartile Range (IQR) that is still inside a maximum distance of $1.5 * IQR$ to the lower or upper quartile. Therefore, the area within the whiskers covers about $99.3\,\%$ of all measured values. Outliers are marked as red crosses. To maintain readibility, all plots are cut off at a limit of $70\,\mathrm{mm}$, if applicable. This is represented with a cut-off line on which outliers over that threshold are marked.

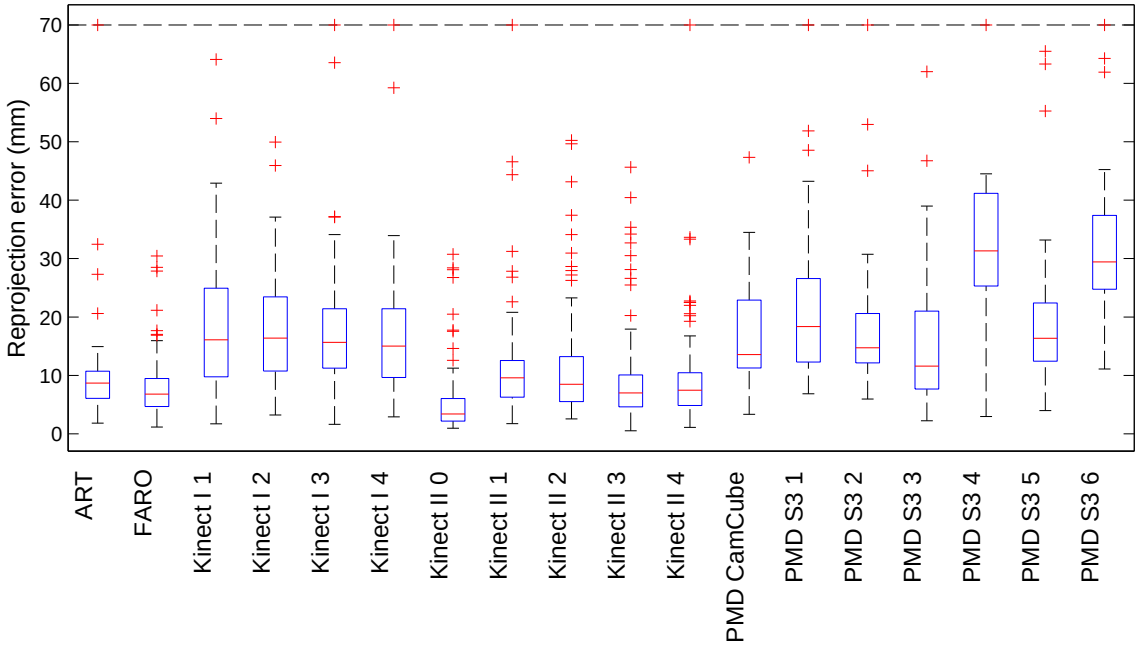


Figure 5.3.: Initial registration error of all cameras for evaluation set PK_1K_2FA.

Figure 5.3 shows the initial registration errors for all cameras in evaluation set PK_1K_2FA. The different types of cameras exhibit clearly different outlier distributions: The optical tracking system and FARO measurement arm show a small median error of $8.7\,\mathrm{mm}$ and $6.8\,\mathrm{mm}$ with a narrow IQR, based on manual annotation of $97.0\,\%$ and $91.9\,\%$ of the projected features. Single outliers occur due to inaccurate manual annotation or technical reasons up to a maximum of $1.025\,\mathrm{m}$ (ARTtrack2).

Kinect v1 cameras with a wide angle and resolution of $640 \times 480\,px$ observed $60.7\,\%$ to $70.4\,\%$ of the projected features with a median registration error between $15\,\mathrm{mm}$ and $16.4\,\mathrm{mm}$. The Kinect v2 camera 0 needs to be regarded separately, as due to its wide FoV and top-down perspective centered above the OR table, it is the only device that detected all but one of the projected features. This results in a low median error of $3.4\,\mathrm{mm}$. Other Kinect v2 cameras that still observe a high

percentage of the projected features (79.3 % – 84.4 %) consistently feature a median error between 7 mm and 9.6 mm.

PMD cameras (excluding camera 4 and camera 5) have a median error of 11.6 mm to 18.3 mm based on a detection rate of 29.6 % to 41.5 % due to their low FoV. Camera 4 and 6 are located sideways to the OR table and feature a high amount of outliers in the raw data, leading to a high median of 31.1 mm and 29.4 mm. The high initial number of outliers for these cameras is caused both by the small coverage area of the projection surface and by unintended side-effects of the manual parts of the registration procedure. The latter include shifting positions of the FARO arm after each measurement as well as other changes outside the ROI which were detected as "features" due to amplitude changes. This is not expected in a real-world scenario where all cameras are registered based on visible light without manual intervention.

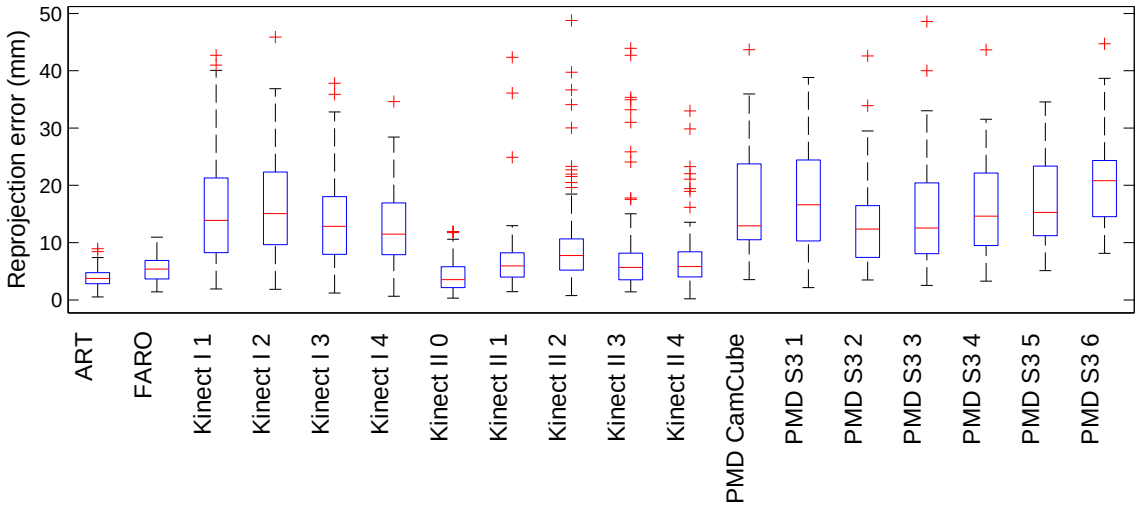


Figure 5.4.: Local registration error of all cameras for evaluation set PK_1K_2FA.

The registration errors after two iterations of statistical outlier removal are shown in Figure 5.4. Compared to the initial registration, all cameras and devices exhibit a smaller median registration error, both due to the modification of the underlying sample and due to the better registration result of the iterative bundle adjustment. This is especially obvious for PMD cameras 4 and 6, for which many incorrect feature detections were present in the raw data set.

The global registration result w.r.t. the ground truth acquired by the FARO measurement arm is shown in Figure 5.5. It shows an increase of median error of 21.4 % as averaged over all cameras.

Figure 5.6 shows the initial, local and global registration error for evaluation set PK_1A, i.e. the supervision system as realized in this thesis. For all systems, the registration error necessarily decreases by outlier filtering. The registration accuracy of both the Kinect v1 system and the ARTtrack2 system also decrease from local to global evaluation, which means that the features detected by the

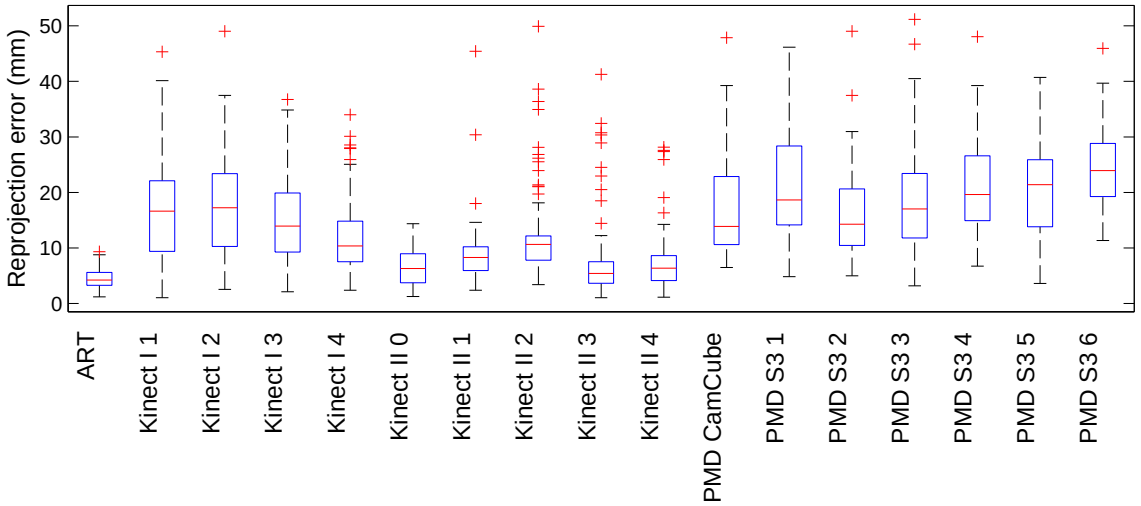


Figure 5.5.: Global registration error of all cameras for evaluation set PK_1K_2FA.

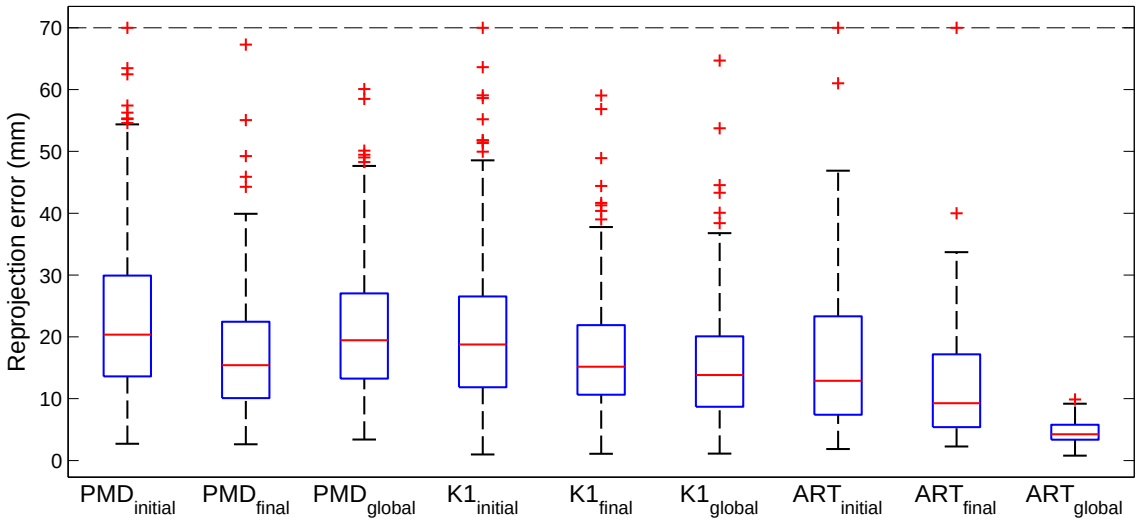


Figure 5.6.: Initial, local and global registration error for cameras of the proposed supervision system, grouped by camera type, for evaluation set PK_1A.

Kinect v1 cameras are more accurate than the reconstructed features. This also explains the increase in registration error for the PMD cameras from local to global evaluation. As expected, the ARTtrack2 system features the highest global accuracy.

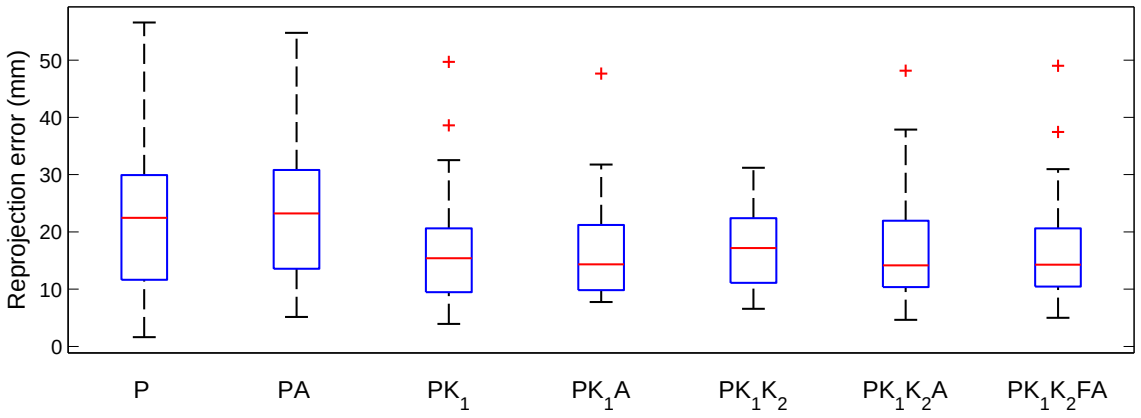


Figure 5.7.: Global registration error of [pmd]vision S3 camera 1 in registration results of different evaluation sets.

Figure 5.7 shows the registration error of the first [pmd]vision S3 camera in different evaluation sets. The highest registration errors are present in the evaluation sets P and PA where only the PMD camera system was used as 3D camera system: Median error and standard deviation are $22.5\,\mathrm{mm}$ and $11.9\,\mathrm{mm}$ for set P and $23.2\,\mathrm{mm}$ and $11.4\,\mathrm{mm}$ for set PA. The error decreases for evaluation sets that include features detected by other camera systems, e.g. from the Kinect v1 (median: $15.4\,\mathrm{mm}$, standard deviation: $10.1\,\mathrm{mm}$) or from the Kinect v2 (median: $17.2\,\mathrm{mm}$, standard deviation: $7.4\,\mathrm{mm}$). These results are in line with those obtained from analyzing evaluation set PK_1K_2FA, where the PMD cameras also exhibited the highest registration error. There is no significant difference between the inclusion or exclusion of a non-camera-based measurement system, such as the ARTtrack2 OTS or the FARO measurement arm.

The registration accuracy for each single camera system is shown for the according evaluation sets P, K_1 and K_2 in Figures 5.8, 5.9 and 5.10. Again, the PMD cameras show the highest registration error with a median of $19.7\,\mathrm{mm}$ over all cameras, due to their low lateral resolution and the resulting inaccuracy of detection of the features. The Kinect v1 camera system performs better with a median registration error of $15.7\,\mathrm{mm}$ over all cameras. Best registration results are achieved for the Kinect v2 system with a median registration error of $7.3\,\mathrm{mm}$ over all cameras.

As result, the proposed projection-based registration process allows for successful registration of multiple camera systems with minimal user interaction. While the PMD camera system that was realized in this work needed to be registered semi-manually, all current ToF cameras acquire both depth and intensity information and can therefore be registered without these manual steps. The results obtained by registering the Kinect v2 camera system clearly show that a higher FoV and

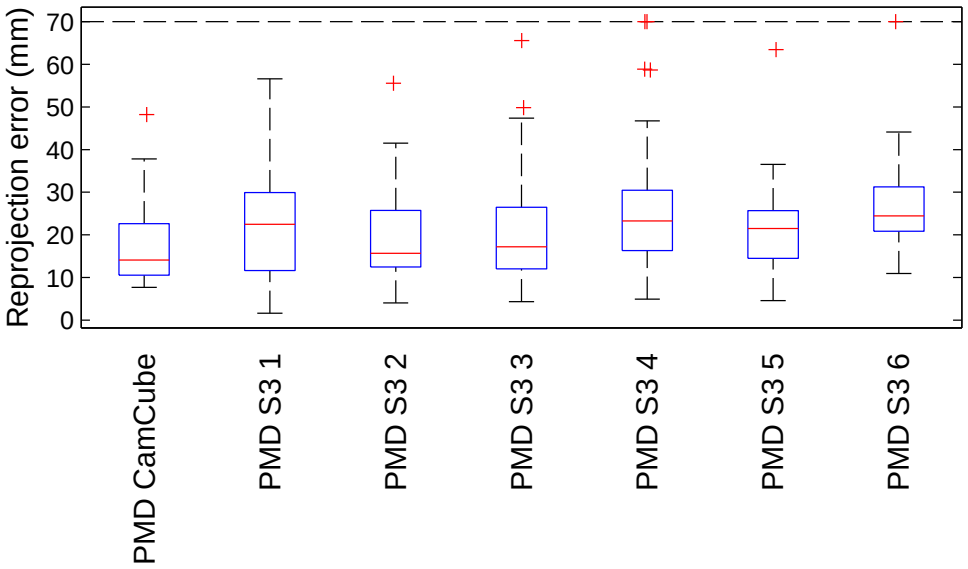


Figure 5.8.: Global registration error of all PMD cameras for evaluation set P.

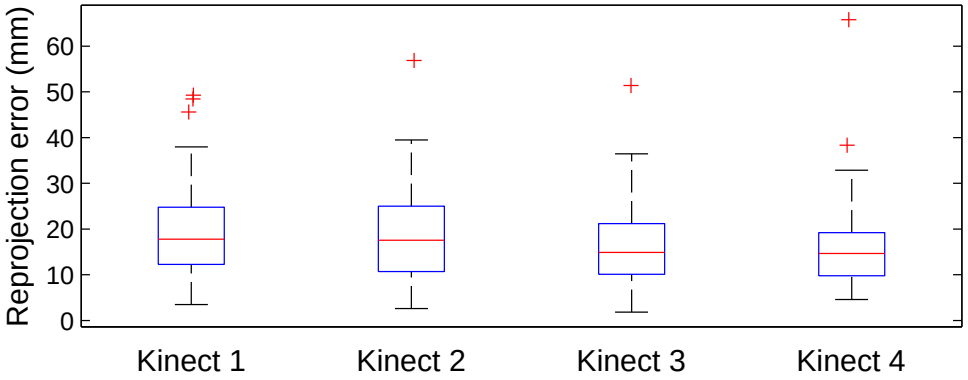


Figure 5.9.: Global registration error of all Kinect v1 cameras for evaluation set K_1.

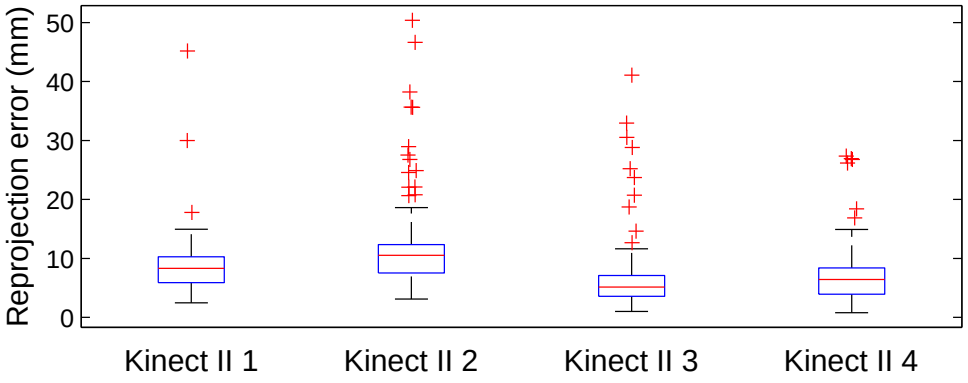


Figure 5.10.: Global registration error of all Kinect v2 cameras for evaluation set K_2.

better lateral and depth resolution increase the achievable registration accuracy, as more features can be detected by each camera.

5.1.3. Frame rate and latency

Both the PMD and the Kinect v1 camera subsystem were evaluated for their performance in terms of frame rate and latency. Table 5.4 shows the results of the PMD camera subsystem, split into performance and quality mode. The given timing statistics for the [pmd]vision S3 and the [pmd]vision CamCube 2.0 correspond to the time between triggering image acquisition by the respective camera and triggering the next camera. For performance mode, this is the active illumination time, i.e. the next camera (group) is triggered immediately after one camera has stopped flashing. For quality mode, this includes illumination, processing and publishing to the ROS network. The *cycle update rate* denotes the update rate of the whole subsystem, i.e. the number of times per second that all cameras have acquired new data. However, new information is available at a higher rate than the cycle update rate because the different cameras of the PMD camera subsystem are triggered at different times (see section 4.4.3.2). The resulting rate at which new information is available in the scene is denoted as the *hybrid update rate*.

Compared to the [pmd]vision CamCube 2.0, the [pmd]vision S3 shows a significantly higher increase of time spent on each camera when enabling quality mode, as can be seen in Table 5.4. The reason is that for [pmd]vision S3 cameras, the integration time is increased by a higher factor and manufacturer-specific implementations of quality enhancement techniques, i.e. Double Sampling and Double Frequency, are enabled in quality mode (see section 4.4.3.4). These modes are not available for the [pmd]vision CamCube. The total time for each update cycle of the performance mode is $79.3\,\mathrm{ms}$. As the region around the OR table is covered by only one [pmd]vision S3 camera each, the cycle update rate is the actually achievable update rate for this region.

For evaluation of the latency, the real environment and its camera-based representation on a screen were filmed simultaneously with a high-speed camera at $240\,\mathrm{fps}$. By annotating frames in the recording where an event happened in the real scene and on the screen, latency could be calculated based on the known frame rate

	Performance mode	**Quality mode**
S3	$16.5\,\mathrm{ms}$	$145.2\,\mathrm{ms}$
CamCube	$43.2\,\mathrm{ms}$	$86.5\,\mathrm{ms}$
Cycle update rate	$12.6\,\mathrm{fps}$	$1.04\,\mathrm{fps}$
Hybrid update rate	$37.9\,\mathrm{fps}$	$7.3\,\mathrm{fps}$

Table 5.4.: Frame times and frame rate of PMD camera system in performance and quality mode

	PMD	**Kinect v1**	***Kinect v2***
1-reliable coverage	77.0 %	98.5 %	100.0 %
2-reliable coverage	40.9 %	62.7 %	86.9 %
3-reliable coverage	25.1 %	24.3 %	59.3 %
4-reliable coverage	8.8 %	12.6 %	42.6 %

Table 5.5.: k-reliable scene coverage by different camera systems.

of the recording. The latency of the ToF camera system was evaluated using a *Hero4* by *GoPro*, USA, which is able to record the flashes of the IR illumination of the PMD cameras. In the footage, the start of image acquisition as well as the earliest visibility of the corresponding point cloud on screen were annotated. For the Kinect v1 camera system, a *SpeedCam MacroVis* by *High Speed Vision*, Germany was used to record the scene. Pre-defined events were annotated in the real scene and on the recorded screen.

The resulting latency per PMD camera is $126\,\mathrm{ms}$ from start of image acquisition to display on screen. This includes an uncertainty of up to $16.7\,\mathrm{ms}$ due to display latency of the $60\,\mathrm{Hz}$ monitor. Based on timings given in Table 5.4, an event in the scene will therefore be picked up by the first PMD cameras after $126\,\mathrm{ms}$ and after $186\,\mathrm{ms}$ by the latest PMD cameras. The latency of the Kinect v1 camera system was evaluated to $966\,\mathrm{ms}$.

5.1.4. Coverage analysis

Hänel introduces the notion of *k-reliable coverage* which denotes the coverage by k different cameras [52]. For the realized camera systems, the k-reliable coverage has been calculated based on a working volume of $2.5\,\mathrm{m} \times 2.5\,\mathrm{m} \times 1.2\,\mathrm{m}$, positioned symmetrically around the OR table at a base height of $0.8\,\mathrm{m}$. This volume was sampled by equidistant points at an interval of $2.5\,\mathrm{cm}$. For each grid point, the visibility for all cameras was calculated based on the field of view of the camera and the camera pose relative to the ARTtrack2, as obtained by the projection-based registration. Further, the coverage of points at a distance of no more than $1.5\,\mathrm{m}$ was calculated. Table 5.5 shows the resulting coverage for both realized camera subsystems. The Kinect v2 camera system is listed for reference, as Kinect v2 cameras offer a significantly wider FoV than the PMD and Kinect v1 cameras.

For visualization of the resulting coverage volumes for the different camera subsystems, they have been rendered with the OR table and one LBR for reference. Renderings are based on a sampling grid with a point interval of $10\,\mathrm{cm}$ and the camera poses obtained by the registration procedure. The color of grid points is consistent to the number of cameras for which this point is visible, ranging from dark blue (visible for one camera only) to dark red (visible for seven cameras), which is only achievable by the PMD camera system.

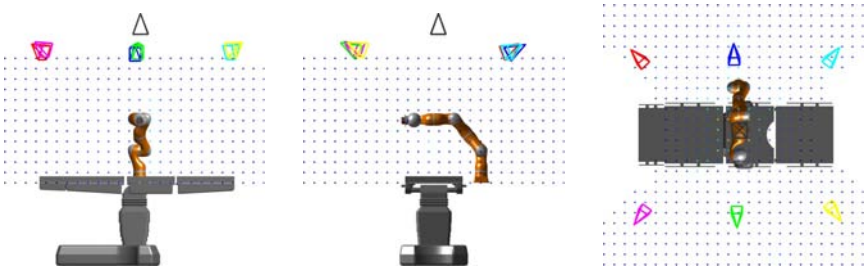


Figure 5.11.: Visualization of volume covered by at least one PMD camera from front, side and top perspective.

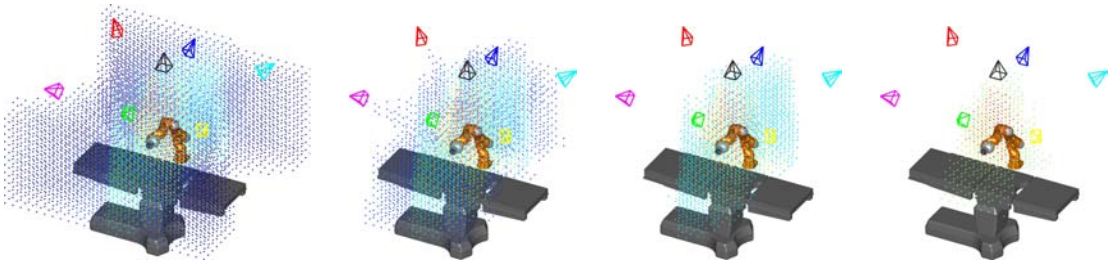


Figure 5.12.: Visualization of k-reliable coverage volume for $k = 1, 2, 3, 4$ *(left to right)* of the PMD camera subsystem.

Figure 5.11 shows the shape of the working volume for the PMD camera system. It is visible from the top-down perspective that the coverage volume of the realized system is hourglass-shaped, i.e. there is a lack of coverage at the both ends of the OR table. In simulation, a better 1-reliable coverage of $96\,\%$ was achieved by angling the cameras in the corners farther outwards in the direction of their diagonal counterparts. However, this would decrease the 2-reliable coverage from $40.9\,\%$ to $34.5\,\%$ and the 3-reliable coverage from $25.1\,\%$ to $14.2\,\%$. Therefore, camera poses as shown in Figure 5.11 were chosen to increase coverage in the region of the robot, which was mounted to the side of the OR table for all experiments. Figure 5.12 shows the k-reliable coverage for the PMD camera system.

Figure 5.13 shows the coverage of the Kinect v1 camera subsystem. While it consists only of four cameras, its coverage is higher due to the wider FoV as compared to the PMD camera subsystem as can be seen in Figure 5.14.

For comparison with the camera system realized within this thesis, coverage of the Kinect v2 camera system was analysed as well. As was expected and is visible in Figure 5.15 and Figure 5.16, the high FoV of the Kinect v2 increases all k-reliable coverages significantly, i.e. by a factor of 2 for the 2-, 3-, and 4-reliable coverage when compared to the PMD camera system.

Figure 5.13.: Visualization of volume covered by at least one Kinect v1 camera from front, side and top perspective.

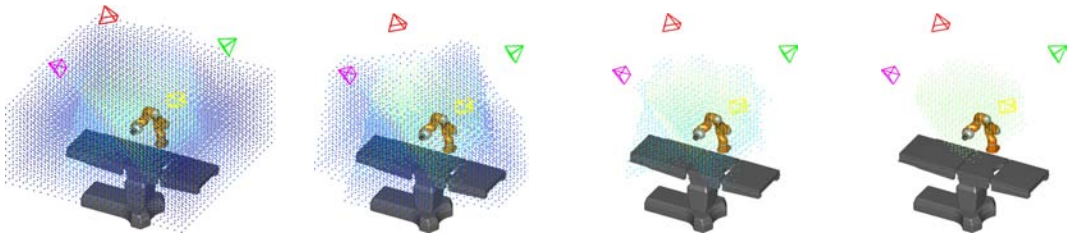


Figure 5.14.: Visualization of k-reliable coverage volume for $k = 1, 2, 3, 4$ *(left to right)* of the Kinect v1 camera subsystem.

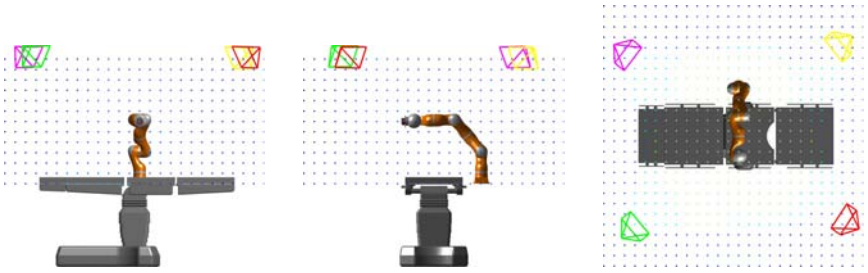


Figure 5.15.: Visualization of volume covered by at least one Kinect v2 camera from front, side and top perspective.

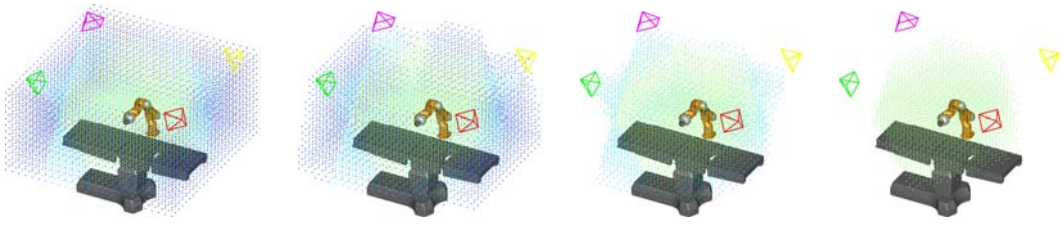


Figure 5.16.: Visualization of k-reliable coverage volume for $k = 1, 2, 3, 4$ *(left to right)* of the Kinect v2 camera subsystem.

Figure 5.17.: Scene representations acquired by PMD *(left)*, Kinect v1 *(center)* and Kinect v2 *(right)* camera system illustrate the achieved coverage of the respective systems.

Figure 5.17 depicts a configuration of the OP:Sense system with two robots mounted to the OR table, acquired by the three different camera systems. As evidenced by coverage analysis, the PMD camera system provides least information of the scene with a high resolution only available in the FoV of the [pmd]vision CamCube. The Kinect v1 camera system provides a more complete scene representation that covers most of the OR table. However, parts of the OR table are not represented as they are shadowed from all covering cameras by the robot system. The Kinect v2 camera system provides the most complete scene representation, as can be seen from the fact that the OR table is fully represented and most parts of the floor are acquired.

5.1.5. Effects of sterile draping

During an intervention, a sterile field has to be maintained around the patient to prevent SSIs. Each robot arm is therefore enclosed in special, transparent surgical drapes as shown in Figure 5.18. As these potentially affect the quality of measurements of 3D cameras with active measurement principles due to additional reflections and scattering of IR light, experiments were carried out to assess the actual effect on the 3D cameras employed in this thesis.

5.1.5.1. Evaluation of influence on measurements

All camera systems were tested individually using the following procedure: Two LBR robot arms were mounted at the sides of an OR table and kept in a static configuration. 100 measurements were acquired per camera, out of which a box-shaped ROI above the OR table surface was extracted, resulting in a subset of the scene that only contains the robot arms. Both robots were then draped using official draping for the daVinci™ S and S_i, the *Instrument Arm Drape* by *Microtek Medical*, Netherlands. With the draped robot arms, a second set of 100 measurements was taken with subsequent extraction of the same ROI as before.

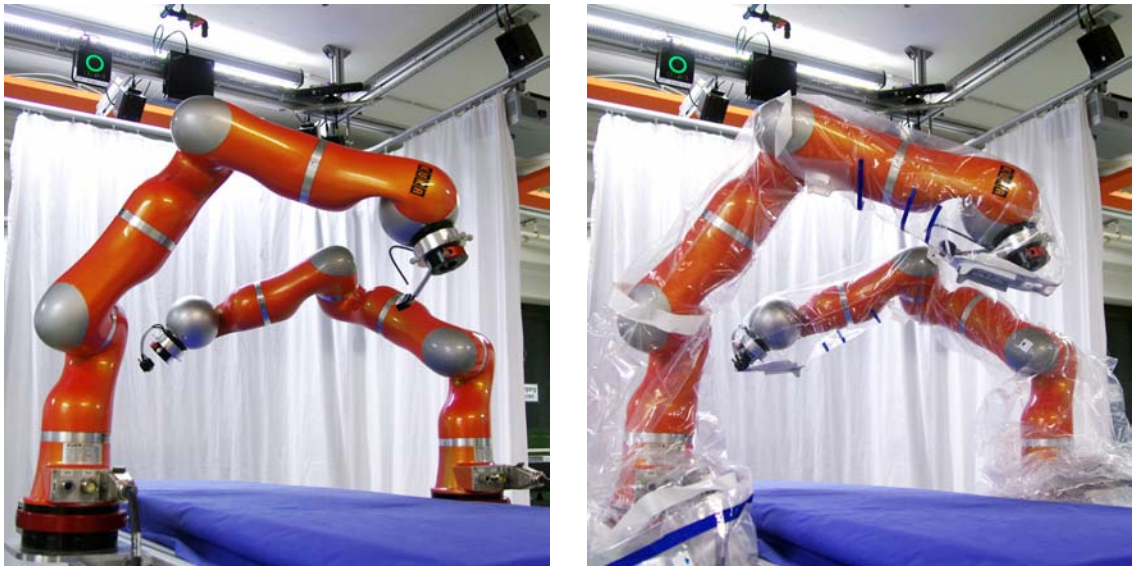


Figure 5.18.: Two LBRs mounted at an OR table without surgical draping *(left)* and with surgical draping *(right)*.

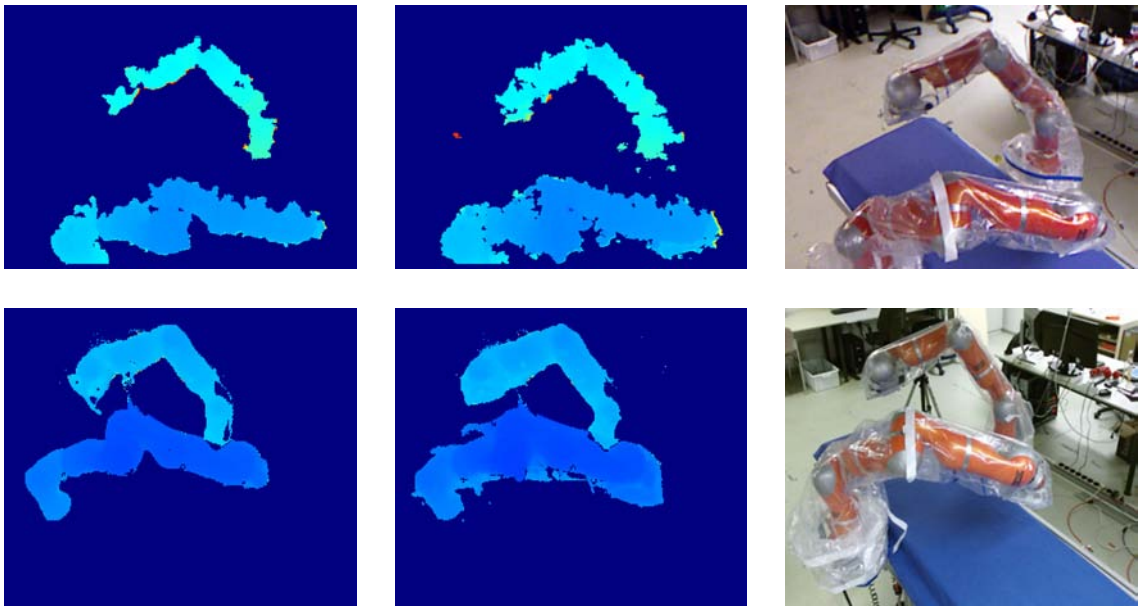


Figure 5.19.: Undraped and draped robot arms as perceived by Kinect cameras. *Top*: Data acquired by Kinect v1 camera, *bottom*: data acquired by Kinect v2 camera. *Left to right*: Depth data of undraped robots, depth data of draped robots, color image of draped robots. All images are crops from the original resolution.

For all measurements obtained with open (non-draped) and draped robot arms, the following metrics were calculated for each pixel where a valid measurement was obtained in at least one frame:

- *Mean distance*: The mean distance of all valid measurements.
- *Standard deviation*: Standard deviation of all valid measurements.
- *Visibility*: The percentage of valid measurements to the total number of frames.

The PMD camera subsystem and the Kinect v1 subsystem were used in their full configuration, i.e. all cameras were enabled for performing the tests. The experiment with Kinect v2 was carried out with only one camera enabled to isolate the effects of the surgical drape and prevent interferences between multiple cameras from influencing the measurements. As the measurements were executed consecutively and are evaluated for the distinct perspective of each camera, the absolute values of the measured distances cannot be compared between the different camera types. Figure 5.19 shows exemplary visualizations of the depth data acquired by Kinect v1 and Kinect v2 cameras for undraped and draped robot arms.

From the results shown in Table 5.6, it can be seen that draping only slightly affects the distance measurements, both absolute and in terms of standard deviation. For all ToF-based cameras, draping unexpectedly improved the visibility, i.e. the percentage of valid measurements per pixel. While the actual improvement is most likely an effect of the glossy surface of the LBR and would need to be tested for differently coated robots separately, it is apparent that the drape at least does not negatively influence the visibility. The Kinect v1 cameras show a distinct drop in visibility, which is consistent among all four cameras with an individual decrease from $12.5\,\%$ to $18.7\,\%$.

	Mean distance		**Standard deviation**		**Visibility**	
	open	draped	open	draped	open	draped
S3 (P)	1.329 m	1.335 m	5.0 mm	5.0 mm	97.0 %	99.0 %
CamCube (P)	1.042 m	1.047 m	9.8 mm	9.8 mm	90.7 %	95.1 %
S3 (Q)	1.378 m	1.383 m	2.8 mm	2.1 mm	95.5 %	99.1 %
CamCube (Q)	1.046 m	1.049 m	5.0 mm	4.7 mm	94.2 %	98.3 %
Kinect v1	1.531 m	1.525 m	16.4 mm	16.6 mm	91.5 %	75.3 %
Kinect v2	1.705 m	1.721 m	14.6 mm	11.7 mm	92.3 %	93.0 %

Table 5.6.: Influence of draping of the robot arms on distance measurement of different camera types. All given results are averaged over all pixels per camera and over all cameras of the same type. (P) and (Q) denote the PMD camera system operating in performance or quality mode.

5.1.5.2. Filtering of draping by PMD camera system

While the draping does not affect the measurement of the PMD cameras, it is important to reliably detect measurements that only correspond to draping in order to isolate the underlying shape of the robot. For PMD cameras, the amplitude value per pixel represents the signal strength of the measurements. As it can be expected that an emitted IR signal that passes through drape and then is reflected by the robot has a higher strength than a signal which is only reflected by drape, the amplitude differences of the received signals should be detectable.

To evaluate the feasibility of filtering draping based on amplitude values, the amplitude of the obtained measurements was analyzed for all PMD cameras. For the measurements performed with draped robot arms, amplitude values were analyzed separately for pixels that correspond to the robot itself and for pixels that returned no valid measurements in the first iteration, i.e. pixels that correspond to draping only. The resulting amplitude values are exemplarily visualized in Figure 5.20. For best visibility, for each image of Figure 5.20 the values have been spread over the full color scale independently. However, the images therefore do not allow direct comparison of absolute values.

Table 5.7 shows the resulting amplitude characteristics. The comparison of amplitude values between open and draped robot arms shows a slight decrease of $3\,\%$ – $13\,\%$ per camera. The amplitude of measurements that correspond to draping is on average $40.4\,\%$ lower than that of measurements that correspond to the draped robot. Therefore, the amplitude is a usable indicator for filtering measurements that correspond to draping.

A frame-based amplitude-based filtering approach was therefore implemented to remove distance measurements that correspond to draping only. In each iteration of the Shape Cropping algorithm, the mean m_I of the amplitude of all inliers in all segments of the robot arm is calculated. A threshold t is then calculated as $t = rf \cdot m_I$, where rf is a custom *outlier rejection factor*. All outliers with an amplitude below t are removed.

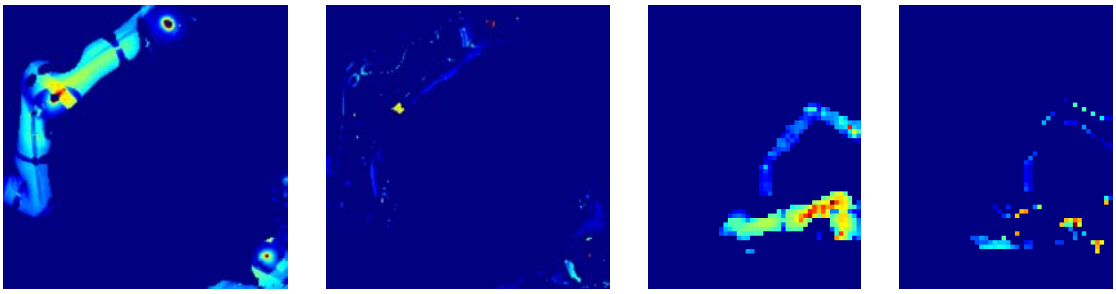


Figure 5.20.: Visualization of amplitude measurements for draped robot. *Left*: Amplitude values of draped robot and amplitude values of drape acquired by [pmd]vision CamCube, *right*: Amplitude values of draped robot and amplitude values of drape acquired by [pmd]vision S3.

	Mean amplitude			Standard deviation		
	open	draped	draping	open	draped	draping
S3 (P)	3.99	3.69	2.22	0.0280	0.0266	0.0170
CamCube (P)	1 550	1 420	838	12.7	15.7	10.3
S3 (Q)	2.95	2.86	1.97	0.0157	0.0197	0.0153
CamCube (Q)	4 550	4 310	2 640	17.4	35.6	48.5

Table 5.7.: Comparison of amplitude of measurements that correspond to the open robot, to the draped robot and to the draping itself. All given results are averaged over all pixels per camera and over all cameras of the same type. (P) and (Q) denote operating in performance or quality mode.

5.2. Forward propagation of semantic labelling

The proposed algorithm has been evaluated for the use cases presented in section 4.4.7.1. All evaluations have been performed based on multiple data sets that were recorded using the `rosbag` mechanism. Each data set contains the time-stamped data received from the respective Kinect v1 camera(s), time-stamped ToF frames of one or multiple PMD cameras as well as the according transformations between all involved cameras. To evaluate the algorithm with different latencies, recorded data sets have been played back with simulated delays for all ground truth data. A list of metrics used for evaluation is given in Table 5.8.

Metric	Definition
True positives tp	Pixels correctly classified as part of the tracked human
True negatives tn	Pixels correctly classified as not part of the tracked human
False positives fp	Pixels incorrectly classified as part of the tracked human
False negatives fn	Pixels incorrectly classified as not part of the tracked human
Precision	$\frac{tp}{tp+fp}$
Recall	$\frac{tp}{tp+fn}$
ToF frame processing time	Time required for processing a single ToF frame (ms)
Ground truth processing time	Time required for forward propagation of the ground truth of a single Kinect frame (ms)
Tracking loss	Percentage of frames with complete loss of tracking

Table 5.8.: Metrics for evaluation of forward propagation of semantic labelling

5.2.1. Latency minimization

Within the proposed algorithm of forward propagation of semantic labelling, the extended tracking map is immediately calculated whenever a new ToF frame arrives. It contains the estimated labelling for the new frame, based on the forward propagated ground truth data (see section 4.4.7.2).

For latency minimization, the extended tracking map was compared to the corresponding ground truth, which is only available multiple ToF frames later. In case of missing ground truth for one or multiple frames, no evaluation was performed for these frames. However, this method of evaluation penalizes longer latencies, because the number of leading ToF frames without any ground truth information frames increases linearly with the latency and therefore more ToF frames do not provide a tracking estimate. To isolate this influence of the delay of the ground truth, each data set S was analyzed in two different subsets S_1 and S_2. The first subset S_1 includes all frames of S, S_2 only includes frames where a tracking estimation was provided by the ToF camera(s), i.e. recall and precision were positive.

The first data set A has a length of $53.5\,\mathrm{s}$ and consists of 317 ToF frames and 265 ground truth frames, acquired by a Kinect v1 and a [pmd]vision S3 with the same angle of view, set apart by a distance of $31.2\,\mathrm{cm}$ (see Figure 4.6). It was evaluated in subsets A_1 and A_2 as detailed above with an artificially induced time delay of ground truth between $1\,\mathrm{s}$ and $10\,\mathrm{s}$. In the data set, a person walks in and out of the scene two times, so the initial delay until the first ground truth is available factors in twice in subset A_1. All results have been averaged over all frames of the according subset.

Processing time per ToF frame was $39\,\mathrm{ms}$, independently of the introduced delay. Figure 5.21 shows the processing time for ground truth frames. It consists of an initial base processing time of about $45\,\mathrm{ms}$, spent on point cloud transformation and correspondence calculation, and a proportional processing time of about $1.7\,\mathrm{ms}$ per second of delay for forward calculation. Figure 5.22 and Figure 5.23 show the number of false negatives and the recall for subsets A_1 and A_2. As expected, for subset A_1 the number of false negatives correlates to the induced time delay and therefore also affects the recall. The number of false positives is below 0.12 for all delays and both subsets, resulting in a precision of $> 99.7\,\%$.

To evaluate the influence of the spatial position of the Kinect camera providing the ground truth and the ToF camera, data set B was recorded with data from Kinect camera 1 and all six [pmd]vision S3 cameras. Figure 5.24 illustrates the position and numbering scheme of all cameras. It can be seen from the results given in Table 5.9 that recall is significantly lower for subset B_1 compared to subset B_2, especially for cameras 4 and 6. This is caused by the different perspectives between Kinect and the evaluated ToF cameras, which lead to a low overlap in the visible field of view, as can be seen from the higher recall result for subset B_2 with the same cameras. Figure 5.25 shows the results of forward propagation for

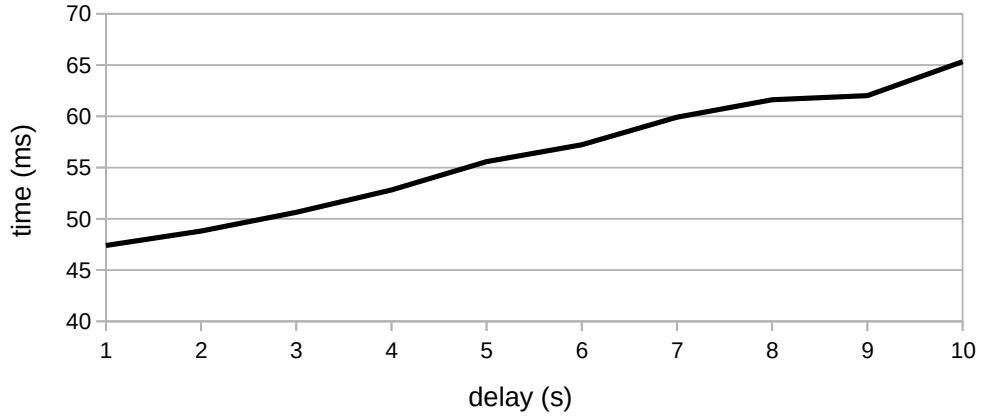


Figure 5.21.: Ground truth processing time for subset A_1 with increasing delay of the ground truth.

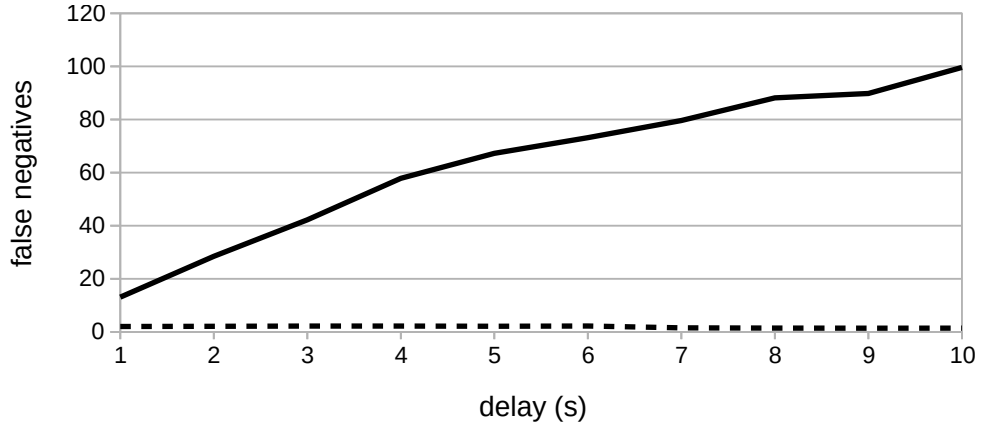


Figure 5.22.: False negative classifications in latency minimization use case for subset A_1 (dotted line) and subset A_2 (continuous line).

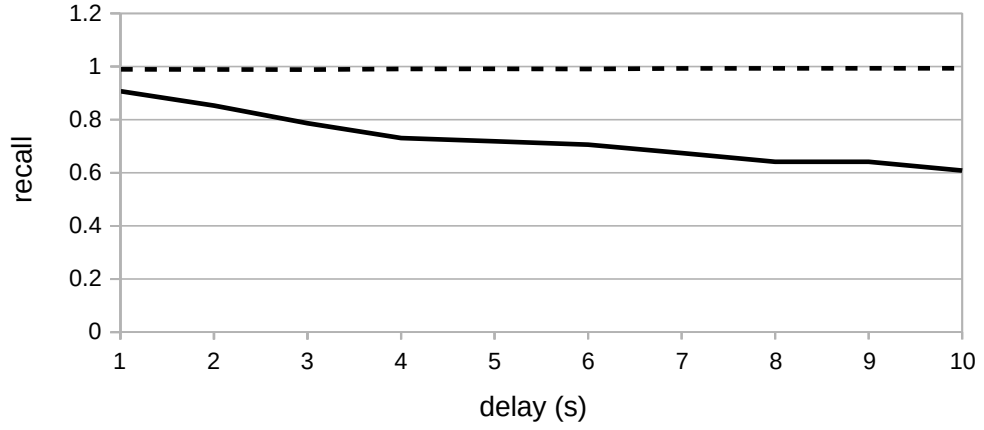


Figure 5.23.: Recall in latency minimization use case for subset A_1 (dotted line) and subset A_2 (continuous line).

	1	2	3	4	5	6
Approx. angle compared to Kinect	0°	90°	90°	180°	45°	135°
Distance to Kinect (cm)	31	163	192	251	92	189
Recall B_1	.71	.71	.80	.66	.80	.64
Precision B_1	.99	.96	.97	.88	.97	.92
Recall B_2	.90	.90	.91	.96	.91	.96
Precision B_2	.99	.90	.97	.88	.97	.96

Table 5.9.: Spatial configuration and accuracy evaluation for six [pmd]vision S3 cameras with different perspectives compared to the Kinect v1 camera and latency of 1 s.

minimizing tracking latency: While the tracked human is perceived in an outdated pose by the Kinect v1 camera (red silhouette), points belonging to the current human pose have been correctly labelled in the ToF camera scene (green points).

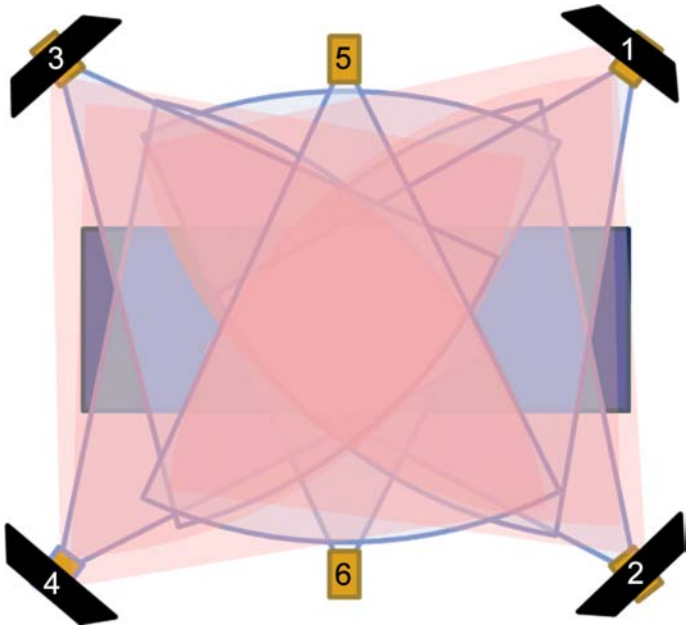


Figure 5.24.: Top down view of spatial camera configuration of the [pmd]vision S3 cameras (brown rectangles) and four Kinect v1 cameras (black trapezes), numbered for referencing.

5.2.2. Optimization of tracking robustness

To evaluate the capability of the forward propagation algorithm to provide continuous tracking estimates during loss of ground truth, a third data set was recorded with a length of 151 s in which one person performs different tasks on both sides of the OR table. The data set contains the fused human tracking data from all cameras of the Kinect v1 camera subsystem as ground truth as well as all ToF frames

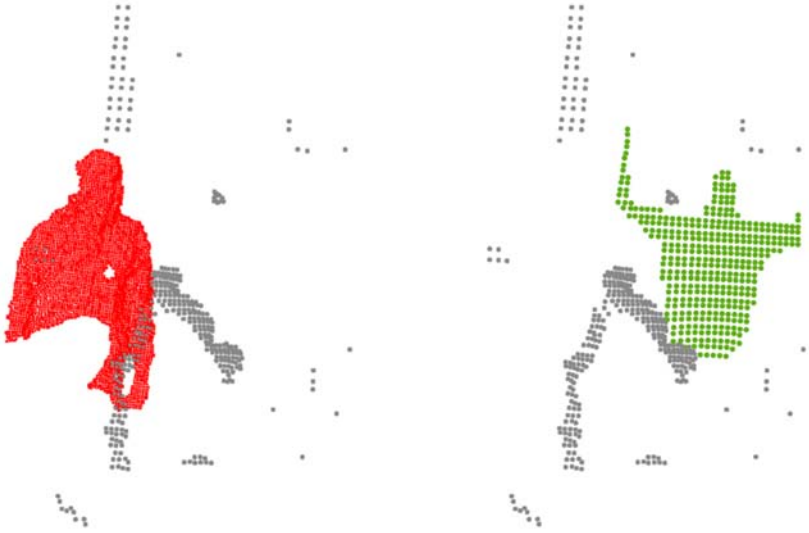


Figure 5.25.: Visualization of forward propagated ground truth for latency minimization. *Left*: Delayed ground truth from Kinect v1 camera shows a human pose from the past (red) in ToF scene (gray); *right*: forward propagation result shows the tracking estimate corresponding to the current human pose (green) in same scene.

acquired by each [pmd]vision S3 camera. This represents the full configuration of the supervision system as shown in Figure 4.21. No artificial delay was induced to the ground truth for evaluation. For each ToF frame, the corresponding ground truth frame was manually annotated using the following classification:

- *Correct*: Ground truth corresponds to the tracked human.
- *Loss*: No ground truth despite the human being visible in the scene.
- *Noise*: Ground truth contains the tracked human, but other parts of the scene are incorrectly also labelled as ground truth.
- *Holes*: Parts of the human were not detected as ground truth.

The data set was analysed for [pmd]vision S3 cameras 1, 2 and 3 whose locations w.r.t. to the OR table are depicted in Figure 5.24. As camera 1 and camera 2 observe different sides of the OR table and camera 3 is positioned without a nearby Kinect v1 camera, these cameras are representative for all different camera positions in the ToF camera system. Loss of ground truth was mostly caused by standing still during scene recording, which is a common reason of failure for human tracking algorithms that are based on detection of motion in their first stage.

To assess the accuracy of the forward propagation algorithm in this configuration, all ToF frames, for which a ground truth was available which was annotated as either "no loss" or "correct", were evaluated For the "no loss" frames, this resulted in a precision of $.97 - .99$ and recall of $.89 - .90$. For frames with a "correct" ground

truth, precision was between $.98 - .99$ and recall in the range of $.87 - .92$. Both precision and recall are similar between the three ToF cameras, showing that the camera position does not affect the accuracy when ground truth from multiple perspectives is available. Compared to the evaluation performed with only one Kinect v1 providing ground truth, recall is on a lower level, whereas precision has significantly increased.

To assess the capabilities of the system to cope with tracking loss of the ground truth, all ToF frames were analysed for data set C, even if no ground truth was present (ground truth marked as "loss"). Results are depicted in Figures 5.26, 5.27 and 5.28 for ToF camera 1 to 3. The Figures show the amount of tracked human pixels over time. Correspondences to ground truth are shown as yellow bars, while the tracking estimate calculated by forward propagation is shown as a curve. The curve is colored green for all frames where ground truth was available and shown as a dotted blue line where no ground truth was available, but tracking still continued.

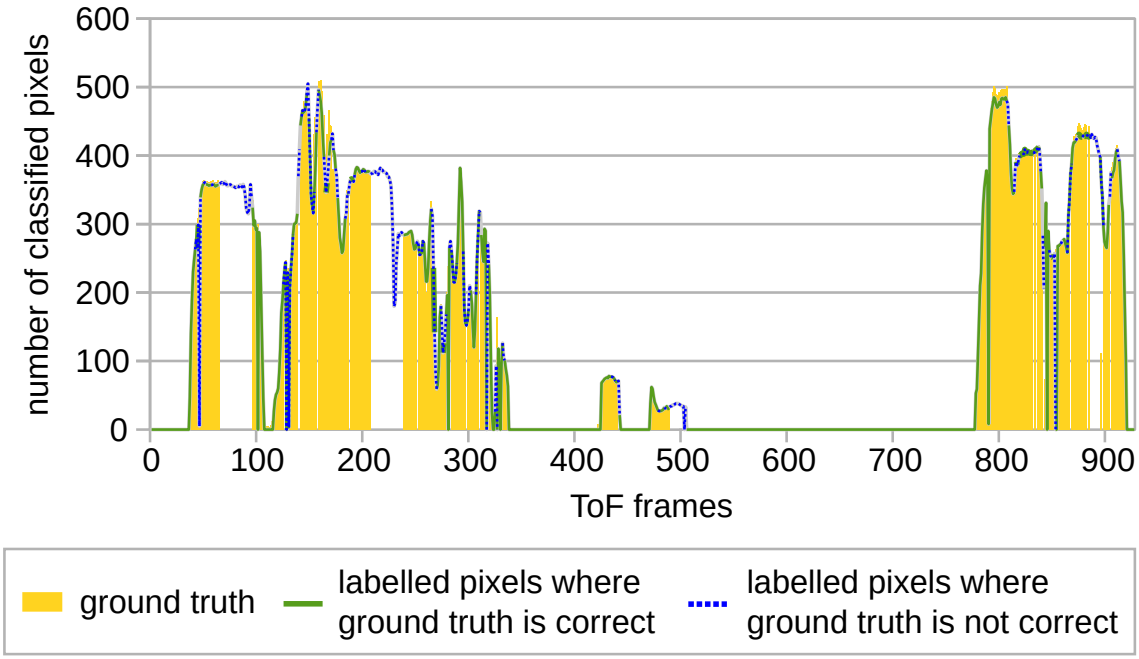


Figure 5.26.: Results of ToF camera 1 for continuous robust tracking based on forward propagation despite multiple losses of ground truth.

As can be seen from Figures 5.26, 5.27 and 5.28, multiple losses of ground truth occur throughout the recorded sequence which are visible as gaps in the yellow bars, e.g. at frames $67 - 97$ or $492 - 594$. The according ToF camera continues tracking until ground truth is recovered, as depicted by the dotted blue line. Continuation of tracking is successful in all instances of tracking loss, even if no ground truth information is available for over 20 seconds as it occurred e.g. with ToF camera 2 around frame 700. Figure 5.29 shows exemplary images from this frame. The tracked person is seen in a pose aimed at re-acquiring tracking by

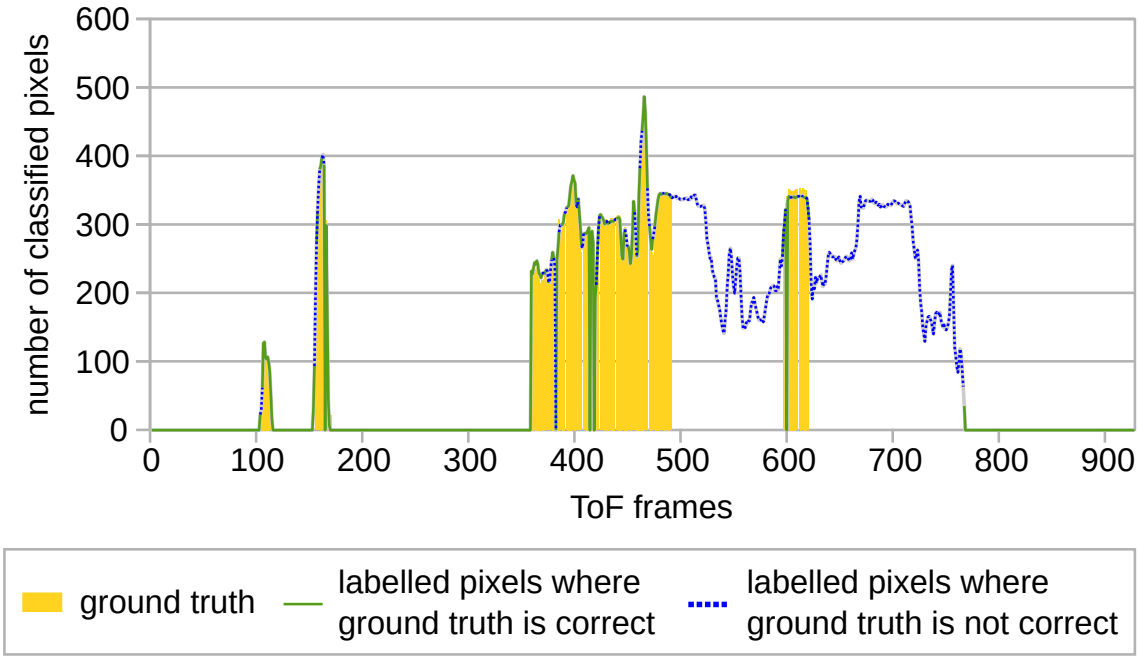


Figure 5.27.: Results of ToF camera 2 for continuous robust tracking based forward propagation despite multiple losses of ground truth.

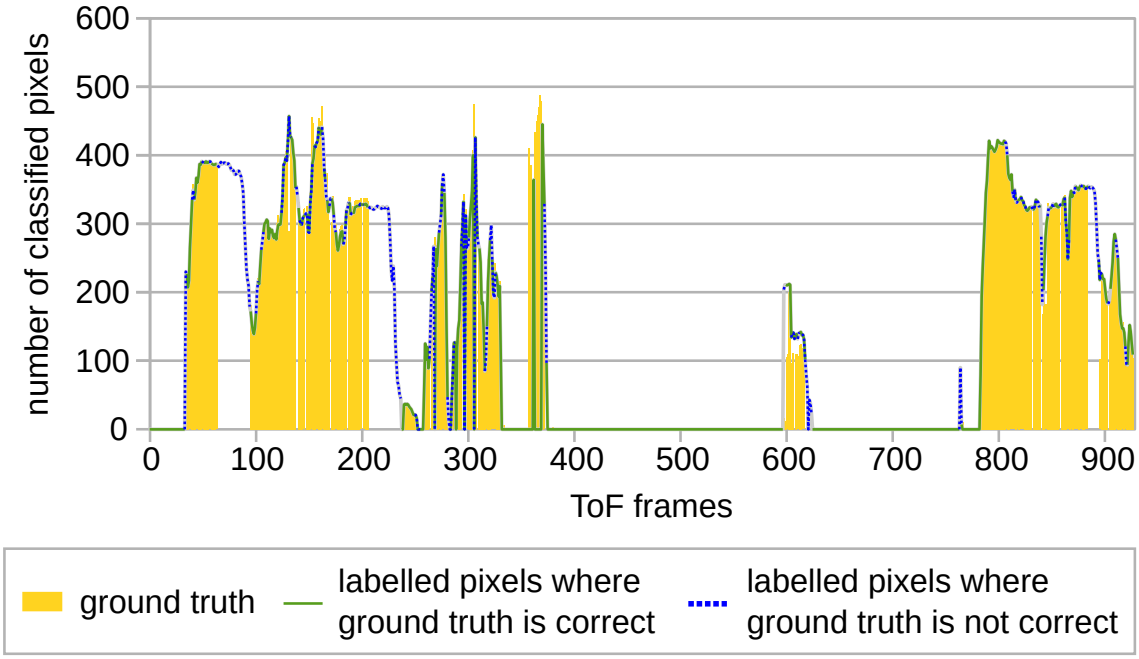


Figure 5.28.: Results of ToF camera 3 for continuous robust tracking based forward propagation despite multiple losses of ground truth.

the Kinect v1 camera system. As no ground truth was available for over $8\,\mathrm{s}$, the forward propagated tracking probabilities are no longer descriptive of the full body of the tracked person, but this information is recovered by the tracking refinement steps described in section 4.4.7.5, as can be seen in the extended tracking map. The removal of boundary pixels, which is clearly visible between the initial tracking map and the extended tracking map, is a consequence of the filtering for flying pixels in the ToF preprocessing (see section 4.4.3.3).

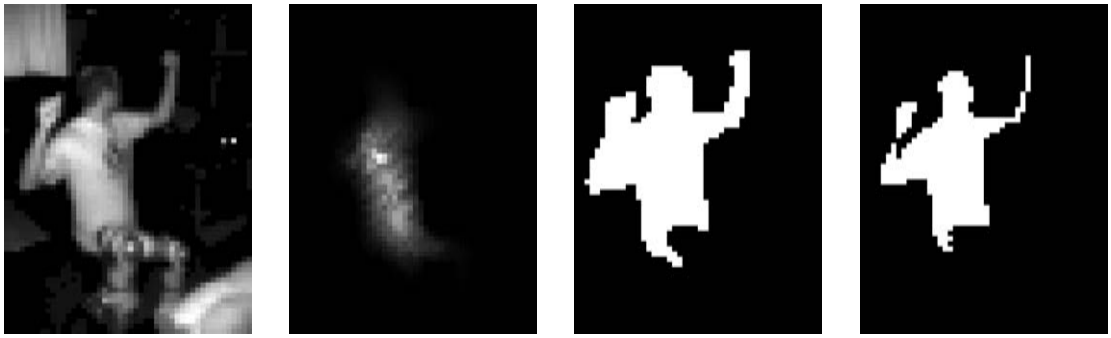


Figure 5.29.: Images depicting the tracking state for [pmd]vision S3 camera in robustness optimization scenario. *Left to right*: Amplitude map, forward propagated tracking probabilities, initial tracking probability map and final extended tracking map.

It can further be seen that the tracking outputs of the different cameras complement each other. While the person is standing on the side of the OR table where the even-numbered cameras are mounted, i.e. outside of their field of view, tracking is performed by camera 1 and 3. Between frame 358 and 766, the tracked person stands on the other side of the table and is then tracked by camera 2 during that time. Afterwards, camera 1 and 3 resume tracking after the person switched sides again. The short gaps between tracking are caused by the lack of coverage at the ends of the OR table in the realized setup, as depicted in Figure 5.11.

In conclusion, forward propagation of semantic labelling and its application to human tracking show very good results. Due to the separated processing pipelines, the processing time for providing the extended tracking map is constant and independent of the delay of ground truth. Taking into account the latency of the first level camera system, the total latency until an extended tracking map is calculated lies between $165\,\mathrm{ms}$ and $225\,\mathrm{ms}$ for a ground truth delay of up to $10\,\mathrm{s}$. This corresponds to an average latency reduction of up to factor 50. Experiments show good precision and recall for scenarios that include ground truth from a single perspective only and better results if a fused ground truth from multiple viewpoints is available. Analysis of the capability to provide continuous tracking during loss of ground truth shows that the system is stable against intermittent tracking and can cover losses of ground truth for as long as $20\,\mathrm{s}$.

5.3. Safety concept

5.3.1. Shape cropping performance

The GPU-based implementation of the shape cropping algorithm was evaluated in a synthetic setting. The virtual robot was positioned inside a 3D grid of equidistant points in a fixed volume and performed an incremental motion in the second joint from -90° to +90° over 1.000 steps. In each step, shape cropping of the robot was performed using the single-threaded CPU implementation as well as the parallelized GPU implementation. The average duration of each operation was calculated over all steps. To take into account the performance penalty of data transfer between host memory and graphics card memory, which is limited by the PCI Express bandwidth, the point cloud of the grid was transferred to the graphics card memory in each iteration and added to the GPU calculation time. The test was repeated for five different grid spacings ($1\,\mathrm{cm}$ – $5\,\mathrm{cm}$). Table 5.10 shows the results.

The test was performed using an NVIDIA GTX 480 GPU and an AMD Athlon Phenom X6 at $3.2\,\mathrm{GHz}$. To estimate the performance on an up to date GPU, the specifications of the GTX 480 were compared to the current Titan X graphics card, also manufactured by NVIDIA (see Table 5.11). Based solely on these specifications and without taking potential efficiency improvements into account, the bandwidth of the Titan X is doubled compared to the GTX 480 and the processing speed is higher by a factor of 9.14. Based on these values, a theoretical performance estimate of the shape cropping algorithm has been calculated which is also shown in Table 5.10.

Grid spacing	1 cm	2 cm	3 cm	4 cm	5 cm
Number of points	2.4×10^6	3.0×10^5	89.780	37.500	19.200
CPU	1.630 ms	195 ms	58.6 ms	25.7 ms	13.1 ms
GPU (GTX 480)	114 ms	18.0 ms	7.96 ms	4.55 ms	4.37 ms
GPU (Titan X, est.)	25.4 ms	3.78 ms	1.45 ms	0.74 ms	0.72 ms

Table 5.10.: Performance evaluation of shape cropping algorithm using a synthetic scene with different numbers of points in a fixed volume.

	GTX 480	Titan X
Host bandwidth	16 GB/s	32 GB/s
CUDA cores	480	3.072
Clock speed	700 MHz	1 000 MHz

Table 5.11.: Main specifications of GTX 480 and Titan X

5.3.2. Robot localization

Evaluation of the proposed passive and active robot localization was performed for all available camera systems. Prior to evaluation, a ground truth was established using the ARTtrack2 OTS: For 15 different joint configurations of the robot arm, the base position of the robot was calculated from the tracked endeffector pose using forward kinematics. The mean of all 15 resulting base positions was then used as ground truth. Based on the extrinsic calibration between each camera and the ARTtrack2 OTS, as obtained by the projection-based registration, the Euclidean distance between the localization result of each camera system and the ground truth was then calculated for each evaluation and interpreted as localization error.

To evaluate the maximal achievable accuracy, localization was performed with robot arms positioned in the central region of the supervised scene, where the k-reliable coverage is maximized for all camera systems. As before, the PMD camera system is denoted with PMD (P) when operated in performance mode and PMD (Q) when operated in quality mode.

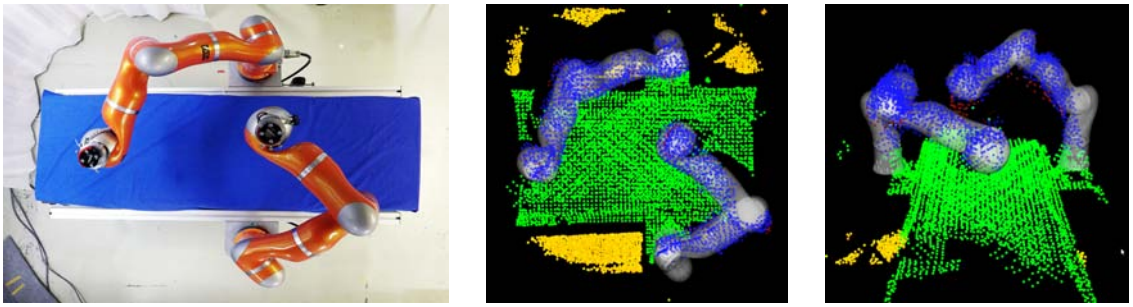


Figure 5.30.: Scene with two LBRs mounted to the sides of an OR table *(left)* and the resulting passive localization with PMD camera system *(center, right)*, shown as CAD models of the robot. Inliers are depicted in blue, outliers in red and neutral scene points in green and yellow.

5.3.2.1. Passive localization

To evaluate the accuracy of passive robot localization, two LBRs were mounted sideways to an OR table (see Figure 5.30, left). Passive localization was performed by using the OR table as a landmark to reduce the size of the search space as described in section 4.5.2.1. Using the base positions detected by passive localization, localization optimization was performed in order to (i) match the detected robot bases with actual robot arms, (ii) correct the orientation of the detected robot poses and (iii) optimize the detected positions (see section 4.5.2.2).

Table 5.12 shows the resulting accuracy of all camera systems for passive localization of undraped robot arms. For the PMD camera system, the initial localization

	Initial error			Final error		
	mean	min.	max.	mean	min.	max.
PMD (P)	36 mm	25 mm	47 mm	36 mm	25 mm	47 mm
PMD (Q)	32 mm	24 mm	40 mm	32 mm	24 mm	40 mm
Kinect v1	113 mm	48 mm	186 mm	53 mm	28 mm	75 mm
Kinect v2	17 mm	9 mm	26 mm	19 mm	12 mm	26 mm

Table 5.12.: Accuracy of passive localization of undraped robot arms with subsequent optimization for all camera systems.

	Initial error			Final error		
	mean	min.	max.	mean	min.	max.
PMD (P)	32 mm	21 mm	42 mm	32 mm	21 mm	42 mm
PMD (Q)	59 mm	16 mm	211 mm	27 mm	16 mm	39 mm
Kinect v1	162 mm	97 mm	233 mm	67 mm	55 mm	79 mm
Kinect v2	97 mm	39 mm	140 mm	32 mm	19 mm	42 mm

Table 5.13.: Accuracy of passive localization of draped robot arms with subsequent optimization for all camera systems.

is already at the same accuracy as the final localization, which means that the optimization steps performed during matching of the robot arms and their orientation already resulted in the optimal position detection. The quality mode of the PMD camera system shows an increase in accuracy of approximately 10 % as compared to the performance mode.
Concerning the Kinect v1 camera system, there is a large initial error of over 11 cm. The optimization steps performed between initial localization and final localization visibly improve the accuracy by 70 %. However, the resulting accuracy is still worse than that of both other camera systems.
For the Kinect v2 camera system, the accuracy of both initial localization and final localization is comparatively high and clearly exceeds the accuracy of both other camera systems.

The same evaluation as above has been performed for draped robot arms. Results are shown in Table 5.13. For the PMD camera system, filtering of outliers was performed with an outlier rejection factor of 1.0, which means that all outliers with an amplitude smaller than the mean amplitude of all inliers were removed. As found in section 5.1.5, the sterile draping actually increases the visibility of the LBR for the PMD cameras, which results in a slightly higher final accuracy than without draping. Again, quality mode shows a slightly better final accuracy than performance mode.
The Kinect v1 camera system repeatedly failed to detect the robot base because of the bad visibility of the draping material around the robot bases. To be able to perform the evaluation, a sheet of paper was attached to the draped robot bases, which increased the visibility. Results of initial localization are therefore partly based on manual intervention and only given for reference. Even with artificially

raised visibility, the accuracy of the Kinect v1 camera system is clearly worse than without draping. For both initial and final localization, it has the largest error by a factor of two.
The Kinect v2 camera system also performs less accurate than in the undraped evaluation, especially in the initial localization, due to the draping that surrounds the robot bases as depicted in Figure 5.31. For final localization, results show a slightly lower accuracy than the PMD camera system in quality mode.

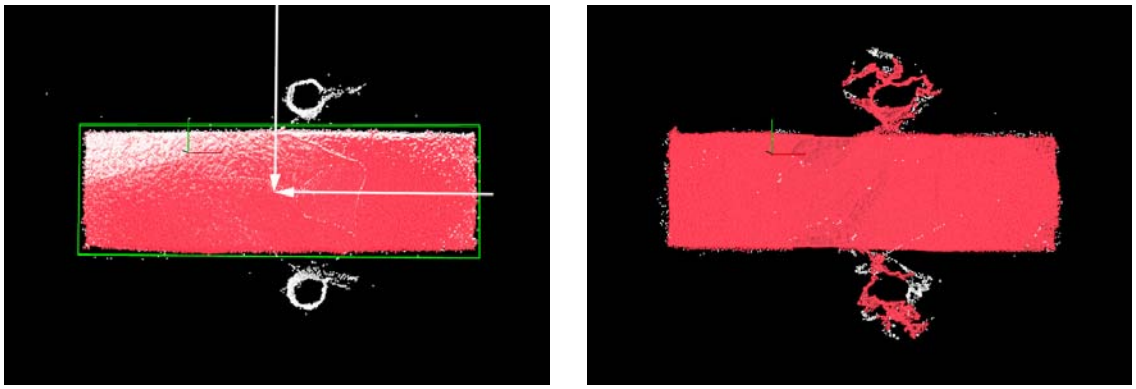


Figure 5.31.: Draping severely affects the perceived robot shape as shown in cross-sections of scene with undraped robot *(left)* and draped robot *(right)*, both acquired by Kinect v2 camera system.

5.3.2.2. Active localization

Active localization has been evaluated using spatial change detection with the subsequent optimization sequence as described in section 4.5.2.1 and section 4.5.2.2. Three different robot configurations were used in step 4 of the optimization sequence. While a per-camera evaluation was performed for both Kinect camera systems, the PMD camera system was evaluated as a whole due to the low FoV of the [pmd]vision S3 cameras.

Table 5.14 shows the results of active localization for undraped robot arms. Compared to the landmark-based initial detection as employed in passive localization, spatial change detection results in a constant, higher initial error for the PMD and Kinect v1 camera system.

Comparison of the different optimizations performed with passive localization (based on only one robot configuration) and active localization (based on multiple different robot configurations) shows an improvement of accuracy by approximately $21\,\%$ for both the PMD camera system and the Kinect v2 camera system. At the same time, the maximum error decreases by $45\,\%$ and $30\,\%$ for the PMD camera system in performance and quality mode, by $23\,\%$ for the Kinect v1 camera system and by $30\,\%$ for the Kinect v2 camera system. Therefore, the range of error also decreases by between $64\,\%$ and $87\,\%$.

	Initial error			Final error		
	mean	min.	max.	mean	min.	max.
PMD (P)	93 mm	92 mm	95 mm	28 mm	25 mm	30 mm
PMD (Q)	40 mm	36 mm	45 mm	25 mm	22 mm	27 mm
Kinect v1	79 mm	68 mm	97 mm	55 mm	52 mm	58 mm
Kinect v2	18 mm	16 mm	19 mm	15 mm	13 mm	18 mm

Table 5.14.: Accuracy of active localization of undraped robot arms with subsequent optimization for all camera systems.

5.3.2.3. Discussion

To assess the feasibility of the obtained results for practical use such as envisioned in this thesis, the actual requirements need to be considered. Specifically, two safety features have been proposed that are based on the robot localization: (i) The verification of the correct setup of the surgical robot system according to a preoperatively determined plan and (ii) the application of Shape Cropping for detecting impending collisions, which requires an initial localization of the robot's position.

The use of robot localization as a prerequisite for detection and avoidance of impending collisions is discussed in section 5.3.3.

The verification of the correct setup of a surgical robot system is necessary to guarantee that it conforms to the pre-planned positions, so that the reachable workspace of the surgical instruments encompasses the full workspace required by the surgeon. For the OP:Sense system, which was used for all evaluations, an analysis of the reachable workspace with pivot restrictions has been performed by Hutzl et al. [63]. As a medical use-case, 20 iterations of a manual rectum resection performed on the OpenHELP-phantom [86] were recorded, annotated and segmented into different phases. The instrument poses of multiple phases were evaluated with different criteria for maximizing the reachable workspace, resulting in a list of ten best-ranked pivot positions relative to the robot base. It was found that these optimal pivot positions vary by $+/- 65\,\mathrm{mm}$ on the robot's x-axis and $+/- 25\,\mathrm{mm}$ on its y-axis, while providing near identical reachability of the workspace among all top-ranked pivot positions ($0.0103 - 0.0141$ on a normalized scale between 0 to 1 with 0 as the optimum).

Comparison of the achieved localization accuracy with these findings shows that the localization accuracy of both the PMD camera system and the Kinect v2 camera system in fact surpasses the practical requirements of the OP:Sense system. This holds true for both passive draped/undraped and for active localization.

5.3.3. Collision avoidance

The performance of the proposed system was evaluated in the OP:Sense setup in different experiments. All experiments were performed using the PMD camera system for acquiring the scene, as it is the only camera system that can reliably filter the sterile draping as evaluated in section 5.1.5. An LBR was mounted to the side of the OR table. Passive localization was performed to obtain the robot base position which is required for employing Shape Cropping for detection of impending collisions. The robot was set to repetitively execute pre-defined trajectories. Based on the acquired scene, the robot base position and the current robot pose, Shape Cropping was carried out for detecting impending collisions as described in section 3.2.4.2.

5.3.3.1. Collision avoidance performance

For evaluating the performance of collision avoidance, the robot performed a swinging motion over the OR table with different joint speeds limit. An obstacle was placed at a fixed location on the OR table, blocking the trajectory of the endeffector of the robot. When the obstacle violated the safety zone and therefore an impending collision was detected by Shape Cropping, a stop of the robot motion was triggered. For assessing the performance, the resulting distance between the end-effector of the robot and the obstacle were annotated (see Figure 5.32) as well as the Cartesian velocity of the endeffector at the time when the robot stop was triggered. The joint velocity of the robot was increased in steps of $0.05\,\mathrm{rad}$. For each velocity, 10 iterations were performed. As the result depends on both the accuracy of the obstacle detection and the braking speed of the robot, which is limited by the acceleration per joint, the experiment was conducted twice with different acceleration limits.

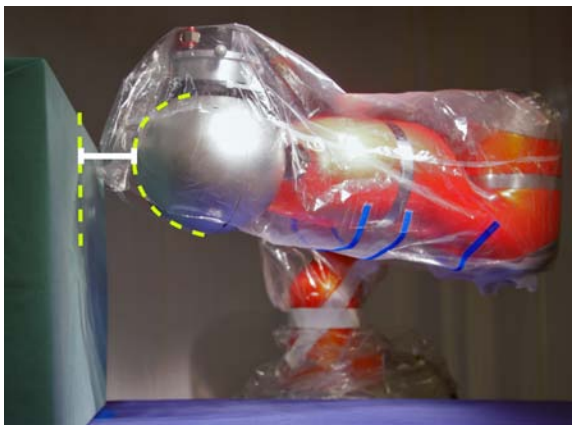


Figure 5.32.: Remaining distance *(white)* between robot endeffector and obstacle *(green)* after an impending collision was avoided by stopping the robot's motion.

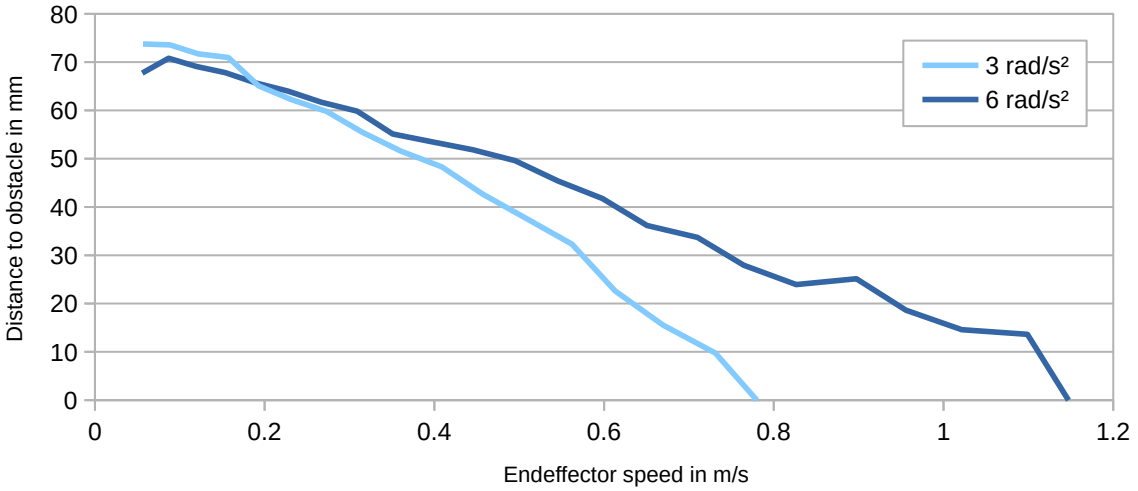


Figure 5.33.: Remaining distances to obstacle at variable robot end effector velocities with different acceleration limits.

Figure 5.33 shows the resulting distances between robot and obstacle over different endeffector velocities. At an endeffector velocity of up to $250\,\mathrm{mm/s}$, which corresponds to the maximum allowed velocity for human-robot interactions specified in ISO 10218-1 and ISO 15066, the robot stopped with a remaining distance of at least $60\,\mathrm{mm}$ to the obstacle. This was achieved independent of the acceleration limits. With increasing velocity of the end effector, the distance to the obstacle decreases. Depending on the acceleration limits, collisions with the obstacle were observed at an end effector velocity of $0.78\,\mathrm{m/s}$ and $1.15\,\mathrm{m/s}$, respectively. In both cases, the collision was consistently narrow, i.e. the robot stopped within the first millimeters of impact.

5.3.3.2. Filtering of sterile draping

In the following, the term *phantom collision* will be used to denote cases where an impeding collision is "detected" despite the fact that no obstacle was close to the robot arm. This especially occurs if the draping is not filtered completely and is therefore classified as an object.

Obstacle avoidance The experiment described in the previous section was repeated with the robot arm covered in sterile draping (see Figure 5.34) in order to assess both the effects of sterile draping on collision detection and the effectiveness of the proposed approach to overcome these effects by amplitude-based filtering (see section 5.1.5). For each trial, the robot performed a fixed trajectory at different velocities with 10 iterations per velocity. Filtering of draping was performed as described above with varying outlier rejection factors between each trial. If a phantom collision was detected, the robot was driven back into its starting

Figure 5.34.: Motion sequence performed by draped robot for evaluation of effects of surgical draping upon detection of impending collisions.

position and the next sub-iteration was started. Therefore, a maximum of 10 phantom collisions per sub-iterations could be perceived. The acceleration limit was set to $6\,\mathrm{rad/s^2}$ for this and the subsequent experiment.

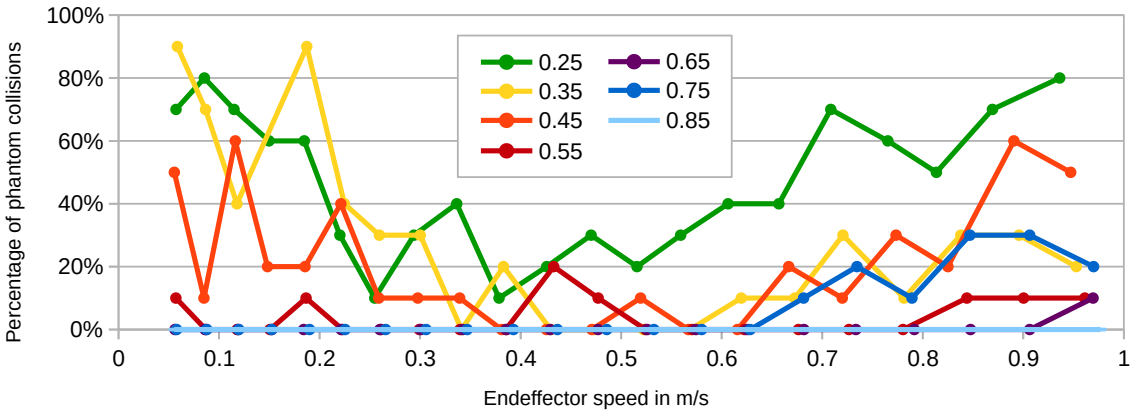


Figure 5.35.: Percentage of phantom collisions at variable robot end effector velocities with varying outlier rejection factors between 0.25 and 0.85.

Figure 5.35 shows the resulting percentage of phantom collisions per iteration for different outlier rejection factors. It is clearly visible that the number of phantom collisions decreases with an increased outlier rejection factor, until no phantom collisions are observed at a factor of 0.85. For all outlier rejection factors, the amount of phantom collisions at an endeffector velocity in the range of approximately $0.35\,\mathrm{m/s}$ to $0.60\,\mathrm{m/s}$ is significantly lower than at other velocities. It is assumed that based on the specific configuration of the PMD camera system and the reflectivity of the surgical draping, draping which moves at this velocity range is more susceptible to being detected as an unreliable measurement by the PMD cameras and therefore eliminated during low level processing (see section 4.4.3.3). However, this phenomenon has not been investigated further as its velocity range is clearly outside the considered velocity range of up to $0.25\,\mathrm{m/s}$.

Further, analysis of all iterations in which no phantom collision was triggered showed that the average remaining distance to the obstacle was slightly smaller compared to the previous experiment with the non-draped robot arm. At a velocity

of about $250\,\mathrm{mm/s}$, the measured remaining distance was in the range of $55\,\mathrm{mm}$ – $59\,\mathrm{mm}$. The first collision with the obstacle occurred at velocities of approximately $1\,\mathrm{m/s}$.

Phantom collision analysis The previous experiment was designed with a focus on evaluating the detection and subsequent avoidance of collisions. As the robot performed a swinging motion towards the obstacle, only the first joint of the robot was moved. Therefore, the surgical draping did not change its shape as it does when all joints of the robot arm are moved.

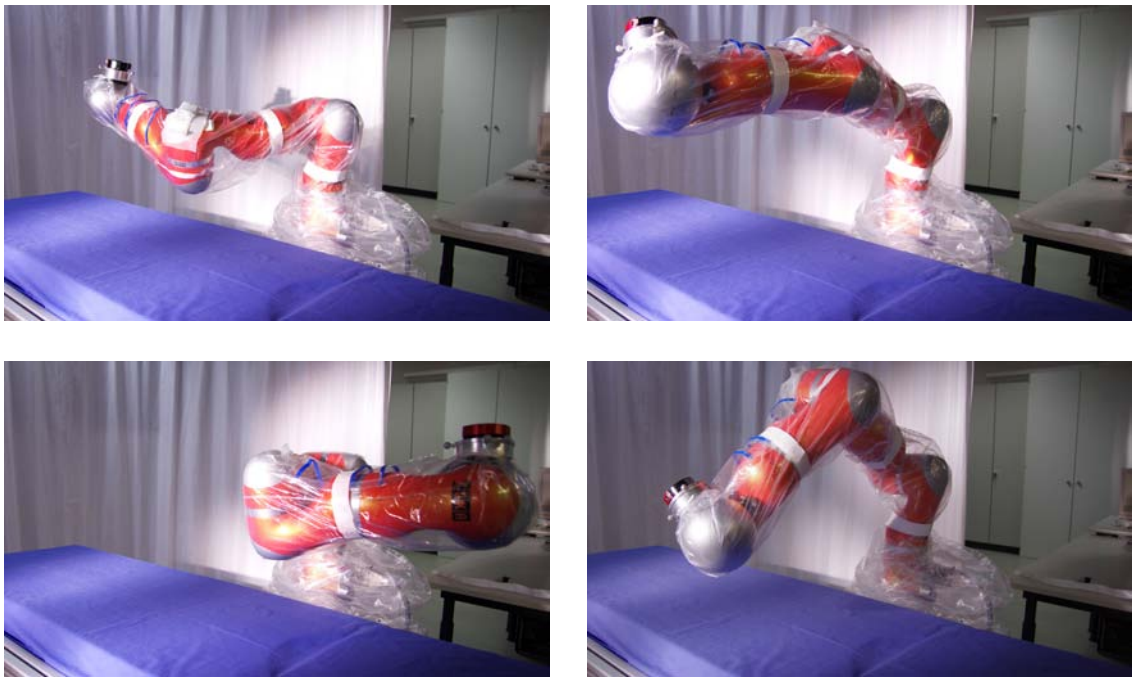


Figure 5.36.: Motion sequence performed by draped robot for evaluation of effectiveness of draping filtering for prevention of phantom collisions.

A second experiment was therefore set up with the goal to evaluate the robustness of the amplitude-based filtering of surgical draping in presence of different draping shapes. The robot iterated between four different poses with completely different joint configurations as seen in Figure 5.36, without any obstacles present on the trajectory. Each phantom collision was annotated and classified as belonging to either the endeffector, elbow or basis of the LBR. Robot motion was performed continuously for a total duration of $127\,\mathrm{s}$ without stopping in case of detected impending collisions. For each evaluated outlier rejection factor, the percentage of Shape Cropping iterations in which a collision was annotated was calculated out of the total number of Shape Cropping iterations.

The resulting percentages of phantom collisions are shown in Figure 5.37. As expected, the percentage of phantom detections decreases with an increased outlier rejection factor. At the elbow, the constant changing of the joint causes the

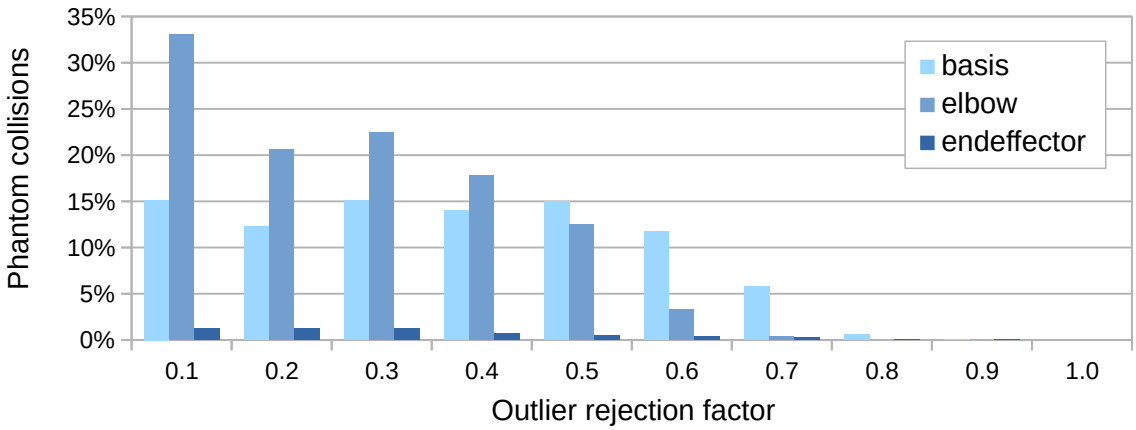


Figure 5.37.: Influence of outlier rejection factor on amount of detected phantom collisions at different segments of the robot.

biggest variations in draping shape, which is reflected in the high percentage of phantom collisions. At the robot basis, the percentage of phantom collisions stays almost constant for outlier rejection factors of up to 0.5. This is due to the fact that at the robot basis, the draping is relatively close to the robot, resulting in a higher amplitude which is removed only for stronger outlier rejection factors.

In accordance with the previous results, an outlier rejection factor of 0.85 and above successfully prevents almost all phantom collisions. As the robot performs exaggeratedly large motions in this experiment compared to real-life scenarios, it is justified to assume that application to real-life scenarios with more restricted motions, e.g. MIRS, will exhibit a similar performance.

5.3.4. Feedback to OR personnel

5.3.4.1. Attention direction

To evaluate the effectiveness of attracting and directing attention, e.g. to potentially hazardous situations, a trial was set up in which the participants performed a given task under different conditions in an interventional scenario.

Trial description The task of the participants was to annotate a series of marked labels on a clinical phantom. To annotate a label, participants had to place the tip of an NDI pointer on the center of a label and press a pedal on a foot switch. The according location was then illuminated by the projector. If necessary, the location could be corrected by removing the last annotation using a second pedal. Throughout the trial, participants stood on the right side of the robot arm depicted to the right in Figure 5.38.

Each participant completed the trial in the following order three times with different goals:

1. *Task focus*: Participants were instructed that their task was to test the combination of foot pedal and NDI pointer for performing accurate annotations. The stated goal for the participants was to annotate 40 numbered circles accurately in ascending order. Each illumination needed to cover the full area of the respective circle. The total time for correctly annotating all labels was visibly recorded and named as the evaluation metric. It was emphasized that accuracy was paramount to discourage fast and sloppy annotations. Further, participants were instructed to verbally communicate if they noticed something unusual with the projection.

2. *Environment focus*: The trial was repeated with the same setup, but participants were instructed to keep an eye on their surroundings and verbally announce if they perceived visual changes.

3. *Environment perception*: Participants were instructed to not perform any actions, but simply watch out for visual changes and announce them as soon as they could detect them.

During each iteration, a visual cue in form of a square of about $7\,\mathrm{cm} \times 7\,\mathrm{cm}$ was projected into the scene in different locations. To maximize its conspicuity, the projection properties were set according to the recommendations of [45] , i.e. lime green projections flickering at a rate of $5\,\mathrm{hz}$. A fixed series of twelve projections of cues was performed during each iteration.

Results The trial was conducted with eight volunteers. During the first iterations with *task focus*, participants were not aware of the projected visual cues and instead focused on the given task. When being asked in a vague way if they had seen anything outside their task, most participants recalled seeing a flickering shape once or twice. However, they were not able to give more detailed specifications. In the *enviromment focus* iteration, participants noticed and verbally communicated the projection of $\approx 70\,\%$ of the projected cues that were visible to them as determined during the *environment perception* trial. When asked about the change of number of projected cues between the three iterations, all participants indicated that based on their perception, "more" or "many more" cues had been projected in latter iterations.

It has to be noted that one cue, which was supposed to be projected onto the edge of the OR table, was instead projected onto the head of the participant in multiple iterations as participants had reached forward to annotate a label. Even with a small sample size, this highlights the susceptibility for occlusion when only one projector is employed. In reverse, it affirms the necessity of supervising the OR field with multiple 3D cameras, as both projector and cameras share the same predicament concerning occlusions.

Generally, the results suggest that it is applicable to attract the attention of OR personnel by projecting visual cues into the scene. However, the personnel needs to be aware that such a safety feature is being used, as intense focus on a specific task prevents perception of projected cues at their peripheral vision, even in confined spaces such as an OR table.

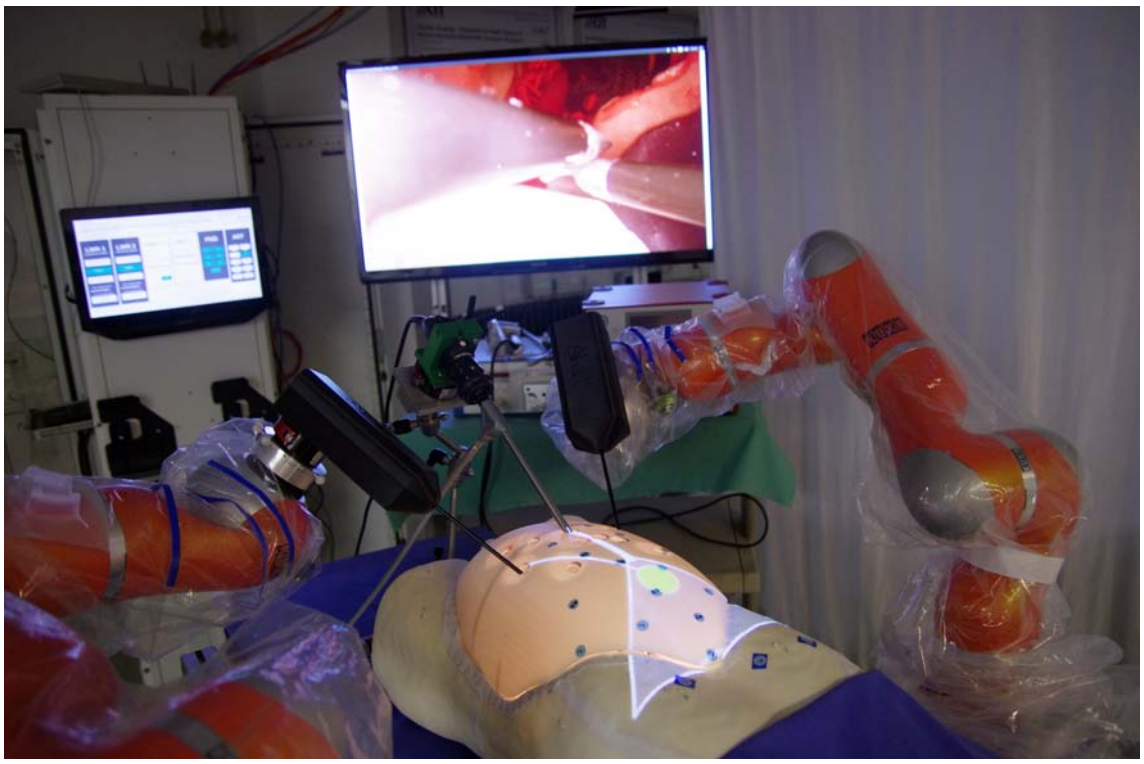


Figure 5.38.: Evaluation scenario for feedback to OR personnel. On the OpenHELP phantom, marked labels and the augmentation of the pose of an instrument inside the phantom and the vision cone of the endoscope are visible.

5.3.4.2. Projection of surgical instruments

To evaluate whether projection of the poses of surgical instruments onto the patient can assist in intuitive and effective performance of tasks related to MIRS, a trial was set up where participants had to perform instrument insertion in a laparoscopic phantom.

The pose of the endoscope as well as its frustum were projected onto the surgical phantom, as depicted in Figure 5.38. When the laparoscopic instrument was inserted into the trocar by the participant, it was also projected onto the phantom as a line between trocar point and instrument tip, using vertical surface mapping as described in section 4.6.3. Further, the depth of the instrument tip and its distance to the optical axis of the camera were visualized as described in section 4.6.5.

Trial description The task of the participants was to insert a laparoscopic instrument through a marked trocar point so that the instrument tip became visible for the endoscopic camera. The instrument was attached to an LBR which was guided by the participant using hands-on mode.

Each participant completed the trial two times under different conditions:

1. *Augmented insertion*: The poses of the instrument and of the endoscope were projected onto the laparoscopic phantom as described above.
2. *Control insertion*: No augmentation was performed. The participants had to rely on their spatial sense alone to guide the surgical instrument into a position where its tip was visible in the camera image.

To avoid learning effects, the order of insertion was randomly pre-assigned to the participants so that both augmented and control insertion were performed first by half of the participants. Directly after completing each insertion, participants filled out the *User Experience Questionnaire (UEQ)* [103], which is a subjective and multidimensional questionnaire intended to assess the quality of experience of interactive products. Participants were asked to rate their experience with the given task and not rate their overall experience with the surgical robot system.

Results The trial was conducted with eight volunteers. Their user experience was evaluated based on the UEQ obtained after the augmented insertion. Figure 5.39 shows the user experience for the augmented insertion task as absolute values on six scales. Each scale ranges from -3 to $+3$, which corresponds to a "horribly bad" to "extremely good" user experience [103]. The UEQ contains a benchmark data set with data obtained of 163 product evaluations, which allows to compare the measured user experience with a large number of products. The results of this comparison are depicted in Figure 5.39 as colored ranges on the respective scales and categories. The achieved categories are to be interpreted as follows:

- *Excellent*: In the range of the $10\,\%$ best results of all products of the benchmark data set.
- *Good*: $10\,\%$ of the results in the benchmark data set are better and $75\,\%$ of the results are worse.
- *Above average*: $25\,\%$ of the results in the benchmark data set are better than the result for the evaluated product, $50\,\%$ of the results are worse.

While all six scales show good results, the results for "novelty" and "perspicuity" stand out. The comparatively low rating of "novelty" can be expected, as the general principle of SAR is long known and participants did not have previous experience with surgical robot systems. The *excellent* result for "perspicuity" confirms that the evaluated augmentation is an effective way of visualization, contributing to an intuitive and shared understanding for potentially multiple observers.

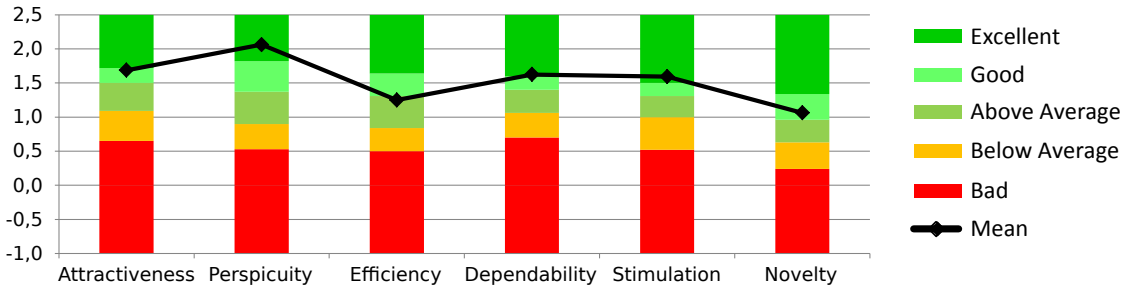


Figure 5.39.: Results of the user experience with projection of surgical instrument in laparoscopic scenario.

To verify that participants actually rated the difference between augmentation and non-augmentation, instead of rating the overall user experience of the surgical robot system for the given task, a direct comparison between the user experience in control and augmented insertion was performed. Figure 5.40 shows the results of the comparison. Results clearly show the difference between augmented and control insertions, confirming that the results discussed above directly assess the user experience of the augmentation.

Due to the low sample size, the confidence intervals for each scale are large. However, they only overlap for "stimulation" and "efficiency", meaning that for all other scales the difference is significant on the $5\,\%$ level.

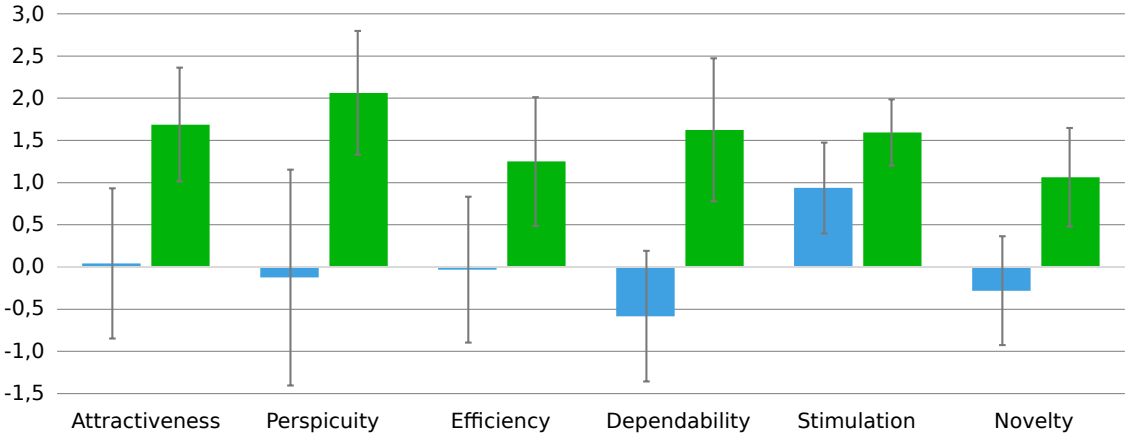


Figure 5.40.: Comparison of user experience between insertion task without *(blue)* and with *(green)* augmentation of surgical instruments.

6. Discussion, Outlook and Conclusions

This chapter provides a summary and retrospect of this thesis. The major findings and results of this work are highlighted and related to the research questions that were put forward in chapter 1.

6.1. Discussion

The evaluations and results presented in chapter 5 show that the proposed system was successfully implemented and fulfills the goals associated with monitoring the safety state of a surgical robot system. The following sections give a brief overview of the developed methods and discuss the achieved results of this thesis.

6.1.1. Supervision system

To supervise the OR, a 3D camera system has been realized that consists of two subsystems with different purposes. Both systems consist of multiple 3D cameras and can acquire a real-time representation of the OR as point clouds. The first level, a PMD camera system consists of seven industrial-grade ToF cameras that offer full external control and configuration and are therefore suitable for safety-critical applications. Their lateral resolution is low, as is the case with most ToF cameras. The second level camera system consists of four Kinect v1 cameras that offer a high lateral resolution, but do allow for external configuration or synchronization.

For comparison, all algorithms have also been evaluated using a separate camera system consisting of four recent Kinect v2 cameras that provide a higher depth resolution and a higher FoV than both other types of camera. However, the Kinect v2 cameras exhibit regular severe crosstalk between each other and offer no external control for synchronization and are therefore not applicable in the proposed supervision system.

To analyze the possibility of operating multiple 3D cameras in the same volume that are based on active measurement principles, i.e. emit IR light, extensive tests have been performed. It has been shown that the crosstalk between multiple [pmd]vision S3 cameras can be effectively eliminated using a time- and frequency-multiplexing scheme. No crosstalk has been observed between PMD and Kinect v1

cameras, confirming the feasibility of the proposed approach to combine both types of cameras into one camera system. Further, results indicate that the simultaneous use of the ARTtrack2 OTS does not influence the depth acquisition by the 3D camera systems.

In order to allow an easy and practicable registration for 3D cameras as well as other devices such as optical tracking systems, a semi-automatic projection-based method for 3D camera registration has been proposed and realized. It is specifically designed for straightforward use in an OR setting and does not require special equipment such as checkerboards or other registration targets. Instead, features are projected into the scene using visible light and detected by each camera. To assess the registration accuracy of the proposed method that can realistically be achieved in a real OR, evaluation has been performed using solely an OR table as projection surface which is in the shared working volume of all cameras of the proposed supervision system.

Registration results have been analyzed in section 5.1.2.2 in different combinations, based on ground truth obtained by a FARO platinum measurement arm. As expected, the low-resolution PMD cameras exhibit the highest registration error with a median error of $19.7\,\mathrm{mm}$, followed by the Kinect v1 cameras with a median error of $15.7\,\mathrm{mm}$. The Kinect v2 cameras show the best registration results with a median error of $7.3\,\mathrm{mm}$. This is due to both their high lateral and depth resolution, allowing for precise localization of the projected feature, as well as their high FoV that enables the detection of about $80\,\%$ of the projected features.

In a recent work by Beyl [9], who calibrated the same Kinect v1 and Kinect v2 camera system using different calibration schemes, a median registration error between $20\,\mathrm{mm}$ and $27\,\mathrm{mm}$ per camera is reported for the Kinect v1 using a pairwise registration. For the Kinect v2, Beyl performed a checkerboard-based registration w.r.t. an external OTS, resulting in a median error of between $5.6\,\mathrm{mm}$ and $10.3\,\mathrm{mm}$. Comparison of the registration methods of Beyl and the proposed projection-based registration shows that the proposed method achieves similar or better registration results while requiring significantly less involvement during the registration procedure. If only 3D cameras with a high lateral resolution need to be registered, further improvements can be expected by projecting more sophisticated features that allow for detection with sub-pixel accuracy.

A coverage analysis of the different camera systems shows that the PMD camera system only covers about $80\,\%$ of the analyzed workspace with gaps at the short end of the OR table, while the Kinect camera systems achieve near or exactly $100\,\%$ coverage of the working volume. Similarly, the Kinect v2 camera system exhibits the highest k-reliable coverage by far. These results underline the need for 3D cameras with a high FoV for applications in safe human-robot interaction where redundant coverage is required in crowded environments.

Forward propagation of semantic labelling was brought forward in this thesis to bridge the semantic gap between the first level scene model and the second level scene model. This has been achieved and evaluated with different use

cases, i.e. latency minimization and optimization of tracking robustness. Latency minimization shows a high precision and recall, even for spatially separated cameras and high delays of up $10\,\mathrm{s}$. Application to the optimization of tracking robustness shows that forward propagation allows to continue tracking even when gaps in the ground truth occur for over $20\,\mathrm{s}$.

6.1.2. Safety concept

In order to use a 3D camera-based system for monitoring safe human-robot interaction, it has to be made sure that the camera system can reliably perceive the robot and its surroundings. As surgical robots are covered in sterile draping during interventions, the effects of these drapings on image acquisition by the 3D cameras were evaluated in different experiments. The results show that despite their transparent material, the draping reflects enough IR light to be registered as valid depth measurements by all camera systems. While the ToF-based PMD and Kinect v2 cameras can accurately perceive the sterile draping, it interferes with the depth perception of the Kinect v1 cameras, which are based on structured light. This results in an increased flickering of the respective measurements.

Based on the hypothesis that distance measurements corresponding to draping result in a lower signal strength, the amplitude map obtained by the PMD cameras was analyzed for differences in signal strength related to draping. It was found that the amplitude of measurements that correspond to draping is on average $40\,\%$ lower for the examined combination of robot and draping compared to the amplitude of measurements of solid objects, even as seen through draping. While the exact results will vary for different surgical robot systems, the difference is high enough to allow for filtering depth measurements that correspond to draping based on their amplitude.

The effectiveness of amplitude-based filtering of draping was confirmed in the scope of collision avoidance tests. Analysis of different rejection factors for filtering of sterile draping shows that the PMD camera system can successfully filter the effects of sterile draping, thereby preventing "phantom collisions" with draping detected as objects close to the robot. At a rejection factor of 0.85, no phantom collisions were detected. The performance of collision avoidance slightly decreased, resulting in a smaller remaining distance to objects blocking the trajectory of the robot of $\sim 57\,\mathrm{mm}$ at an endeffector velocity of $250\,\mathrm{mm}$. The reliability of amplitude-based filtering of draping was also confirmed by analysing phantom collisions at large-scale motions of the LBR that cause more wrinkling of the sterile draping than can be expected in actual applications such as MIRS. In this experiment, occasional phantom collisions still occurred at a rejection factor of 0.85, but were completely eliminated using a rejection factor of 1.0.

Further, during collision avoidance tests it was found that at an endeffector speed between $0.35\,\mathrm{m/s}$ and $0.6\,\mathrm{m/s}$, the number of phantom collisions decreased among all outlier rejection factors. It is assumed that this is an effect of the specific

combination of cameras, their integration times and the reflectivity of the sterile draping that leads to rejection of the according distance measurements as motion blur effects for this velocity range. However, this has not been researched further because this velocity range is above the limits for safe human-robot interaction according to ISO 10218-1.

For verifying the robot positions w.r.t. accordance to preoperative planning, different localization methods based on Shape Cropping were realized and evaluated in terms of accuracy. The resulting localization error lies in the range of centimeters for all camera systems and for both localization methods, with and without sterile draping. The Kinect v2 camera system consistently achieves the best accuracy with an error of $19\,\mathrm{mm}$ and $15\,\mathrm{mm}$ for passive and active localization, being second only to the PMD camera system in quality mode for localization of draped robots. Here, the error of the Kinect v2 cameras system of $32\,\mathrm{mm}$ is larger than the error of the PMD camera system of $27\,\mathrm{mm}$. This again shows that the approach to filter sterile draping with the PMD cameras effectively preserves quality of their distance measurements, as opposed to the other camera systems where accuracy decreases in the presence sterile draping. Over all localization methods, the Kinect v1 camera system is the least accurate system for robot localization.

Comparing the accuracy of active and passive localization, it is noteworthy that active localization achieves only a lower initial accuracy, but results in better final accuracy as multiple optimization steps with different robot poses are performed. Similarly, the error range of the final active localization is significantly smaller than for passive localization, which does not move the robot during localization. Therefore, it is proposed to employ active robot localization when possible, e.g. when the robot is mounted to a fixed position in the room before the start of an intervention as proposed for the scenario of the European research project ACTIVE.

To enable a qualitative assessment of the achieved localization accuracy, comparison was performed to a recent analysis [63] of the influence of the pivot point on the reachable workspace of the LBR. According to Hutzl et al., the positions of optimal pivot points for the LBR vary by as much as $\pm 65\,\mathrm{mm}$ and $\pm 25\,\mathrm{mm}$ on the x- and y-axis of the robot. This confirms that the achieved localization accuracy of both the Kinect v2 and the PMD camera system fulfil the accuracy requirements of the specific surgical robot system, the OP:Sense platform, with which all evaluations were performed.

6.2. Future research

This thesis presents a system concept for the safe usage of surgical robot systems in the OR of the future. The current realization of the system concept with PMD and Kinect v1 cameras already shows its applicability to safe human-robot cooperation in the OR. It opens the way to future research, such as:

- *Technical enhancements*: With recent developments in the field of range imaging, the achieved results could be further improved. As an example, a higher field of view of the cameras directly contributes to the redundant coverage of the system, thereby increasing its robustness against occlusions which in turn improves its applicability for safety-related applications. Higher frame rates and/or lower latencies of the cameras would directly increase the performance, resulting in increased reaction times to adverse situations. Due to the modular nature of the system, such enhancements would not require extensive modifications, but could directly be integrated and evaluated.

- *Technical integration into the OR*: Due to the size of the cameras and the requirement of positioning them in multiple locations over and around the OR table for preventing occlusions, the current system occupies a not-too-small volume despite the spatial separation of data acquisition and processing. With current developments towards miniaturization, it might become valid to integrate miniature range imaging devices directly into the OR, for example in the ceiling and the operation lamp. Especially the integration into the OR lamp shows promise, as it is usually positioned by the surgeon above the situs with a clear field of view to patient, OR personnel and surgical robot system.

- *Semantic integration into the OR*: Integrated operating rooms of the future promise an integration of medical devices not only on a technical, but also on a semantic level. Semantic integration of the supervision system and the SAR system would offer new possibilities for both localizing medical devices in the OR, e.g. for the task of OR setup and management, and relaying information from medical devices to the OR personnel by visualization directly in the scene.

- *Case studies with upcoming surgical robot systems*: The idea of interoperability with different surgical robot systems forms the core of the proposed system concept. The actual integration with different surgical robot systems and practical assessment of the resulting advantages and/or shortcomings in a clinical environment would therefore be the logical step.

- *Integration with a knowledge-base*: The current embodiment of the system only works solely on geometric scene information, which is acquired and analyzed within the system. Integration of a knowledge base could enable higher-level reasoning about the perceived scene, allowing for scene interpretation and a better support of the OR personnel.

- *Application to non-medical scenarios*: The proposed system concept was developed and designed based on the requirements of an OR. Nevertheless, both the full system concept and the different contributions, such as forward propagation of semantic labelling, are applicable to a wide range of topics beyond the OR.

6.3. Conclusions

In this doctoral thesis, a 3D camera-based system for safe and intuitive usage of surgical robot systems has been developed, based on state-of-the-art range imaging cameras and novel algorithms. The system enables perception of the environment of a surgical robot using multiple 3D cameras and can detect potentially harmful interactions between the robot and humans, i.e. the OR personnel and the patient, as well as its environment. This allows the OR personnel to completely focus on their medical task without having to divert attention to check the correct functioning of the surgical robot system. Therefore, the proposed system has the potential to contribute to the patient outcome of robot-assisted interventions.

In summary, the main contributions of this thesis include:

Supervision system A concept for a modular, distributed 3D camera system is presented that allows to acquire a 3D scene in real-time from multiple points of view. Realization and analysis of the system show the feasibility for applications in safe human-robot interaction. Due to the spatial separation of data acquisition and processing, the system has a small footprint that renders use in a surgical environment possible. The system concept is not limited to the surgical domain, but has the potential to be applied to various scenarios.

Shape Cropping algorithm Shape Cropping allows to construct a virtual safety zone around arbitrary robot manipulators, solely based on pointclouds of the surrounding scene. In this thesis, Shape Cropping is shown to be applicable to safety-related aspects, where it can ensure that potential collisions are detected before they occur, as well as to localization of robots in unknown scenes and redundant supervision of correctness of the robot's poses.

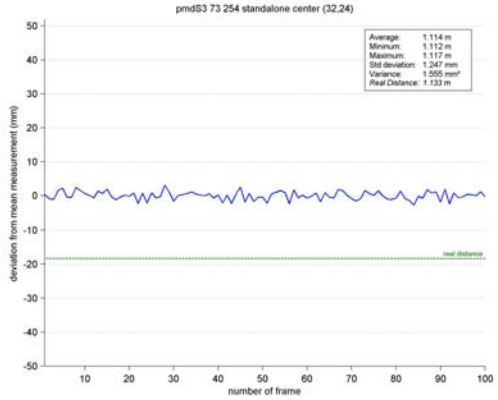
Forward propagation of semantic labelling Within this thesis, forward propagation of semantic labelling is employed to bridge the semantic gap between the two camera subsystems: The human tracking information provided by one camera system is propagated forward to the other camera system. By design, the algorithm is not limited to applications with multiple 3D camera systems or to human tracking. It allows for forward propagation of data between arbitrary data sources, requiring only a known mapping between the data sources.

Intuitive feedback to OR personnel A projector-based system for SAR is realized and fully integrated with the proposed supervision system. It allows providing information to the OR team in an intuitive and natural way by projecting information directly into the scene, where it is visible for all members of the OR team and can thereby facilitate the communication flow and a shared situation awareness.

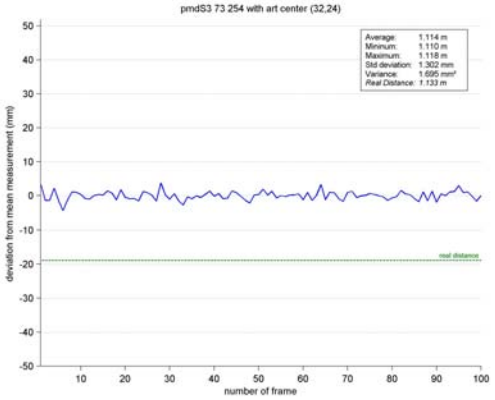
Analysis of the effects of sterile draping on perception by 3D cameras This thesis conducts the first analysis of the effects of sterile draping, in which surgical robots are covered during interventions, on their perception by 3D cameras. This is important for all future work on human-robot interaction in the OR where 3D cameras are employed for scene acquisition and analysis. A method for filtering sterile draping from scenes acquired by PMD cameras is devised and experimentally confirmed, demonstrating its applicability to safe human-robot interaction in the OR.

Appendix

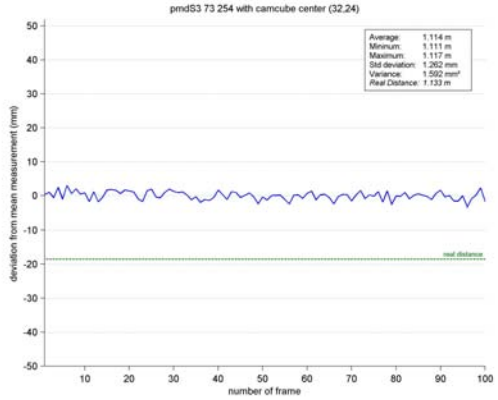
A. Interference analysis



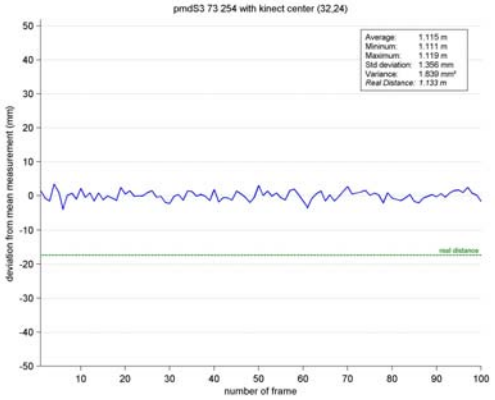
(a) standalone



(b) with ARTtrack2

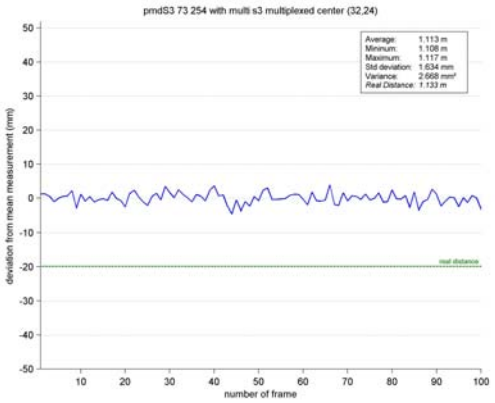


(c) with pmd[vision] CamCube 2.0



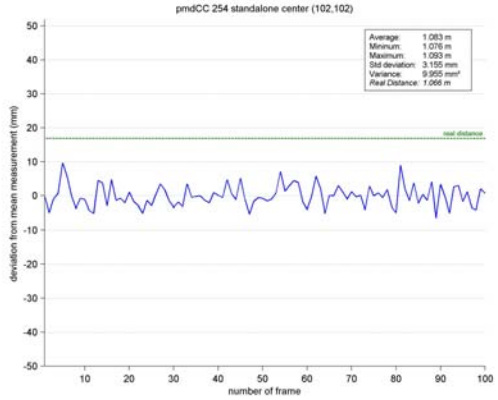
(d) with Kinect v1

(e) with unmultiplexed S3 cameras

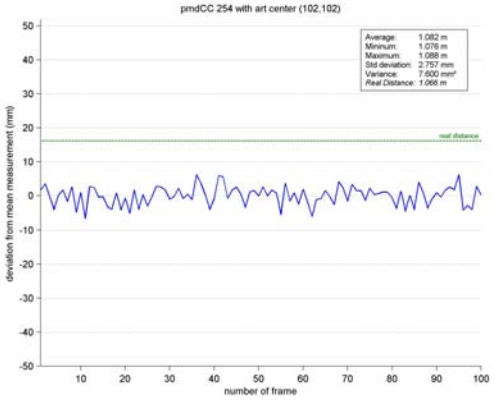


(f) with multiplexed S3 cameras

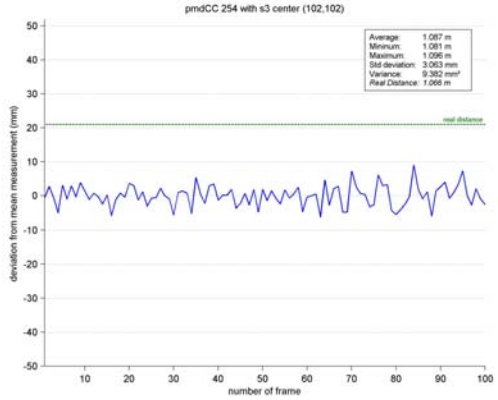
Figure A.1.: Distance observed over 100 iterations by pmd[vision] S3 for static target at a distance of $1.13\,\mathrm{m}$ in presence of different other cameras and systems.



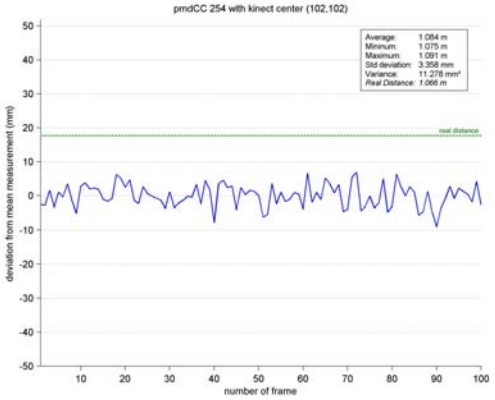
(a) standalone



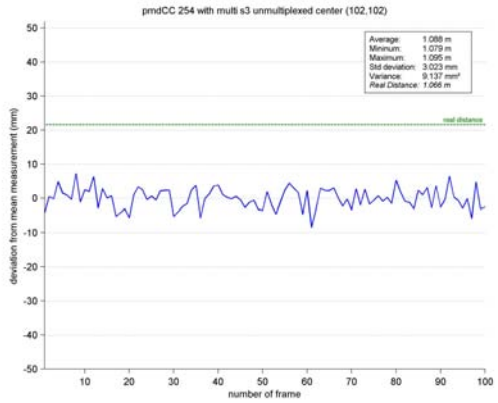
(b) with ARTtrack2



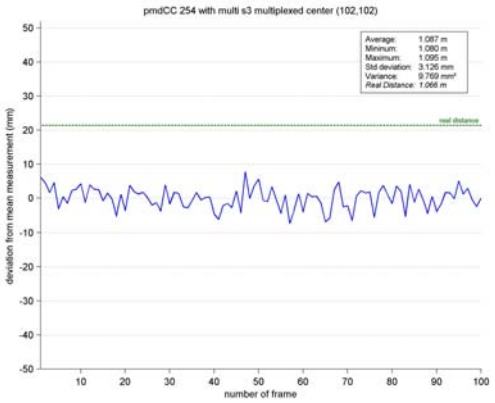
(c) with pmd[vision] S3



(d) with Kinect v1

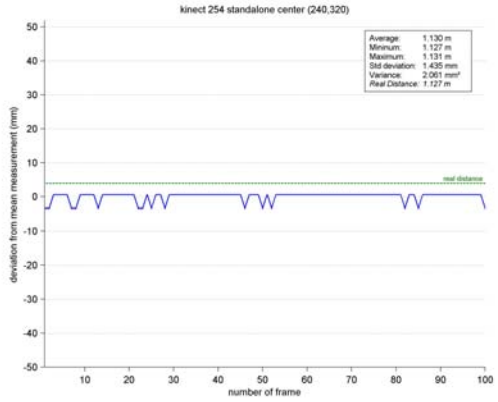


(e) with unmultiplexed S3 cameras

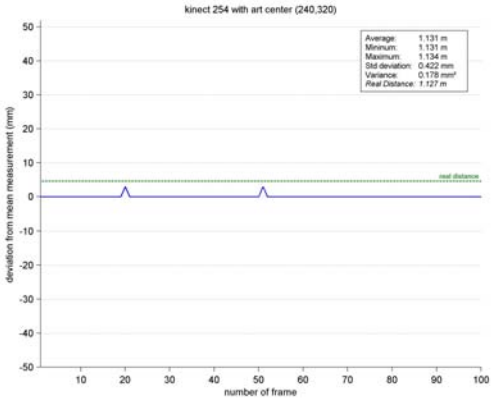


(f) with multiplexed S3 cameras

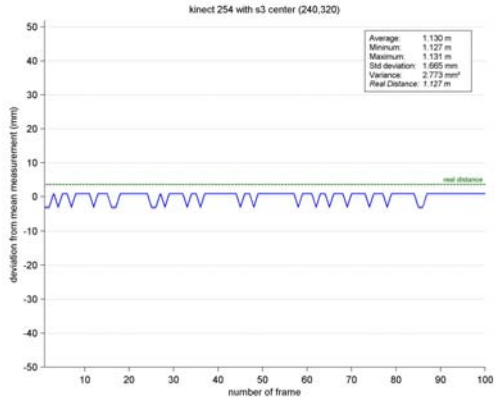
Figure A.2.: Distance observed over 100 iterations by pmd[vision] CamCube 2.0 for static target at a distance of $1.07\,\mathrm{m}$ in presence of different other cameras and systems.



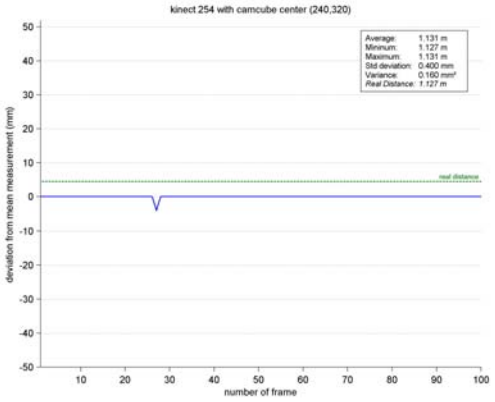
(a) standalone



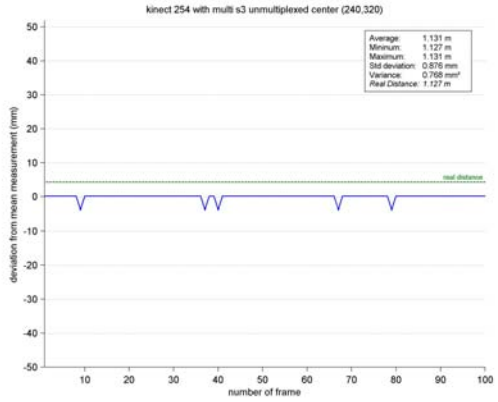
(b) with ARTtrack2



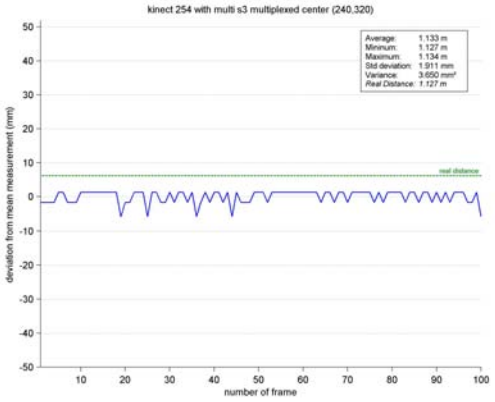
(c) with pmd[vision] S3



(d) with pmd[vision] CamCube 2.0



(e) with unmultiplexed S3 cameras



(f) with multiplexed S3 cameras

Figure A.3.: Distance observed over 100 iterations by Kinect v1 for static target at a distance of $1.13\,\mathrm{m}$ in presence of different other cameras and systems.

B. Pinhole camera model

The basic pinhole camera model calculates the 2D projection of a 3D scene point at world coordinates (X, Y, Z) based on the focal length f_x, f_y and principal point (c_x, c_y) of the camera as well as the pose of the camera in world coordinates as given by the transformation matrix R with rotational components $r_{11}, \ldots, r_{33}$ and translational vector $t = (t_1, t_2, t_3)^\top$:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{B.1}$$

The simple pinhole model does not take into account distortions of a real optical system, such as radial and tangential distortions caused by lenses. Therefore, distortion coefficients are introduced to model these properties. In this thesis, the camera model implemented for bundle adjustment is based on the OpenCV model for camera calibration [138], which in term is similar to those introduced by Claus and Fitzgibbon [21, 35].

Equation B.1 can be rewritten in the following steps by transforming world coordinates (X, Y, Z) to coordinates (x, y, z) in the camera coordinate system as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = R * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + t. \tag{B.2}$$

Provided that $z \neq 0$, the projection to pixel space is performed by

$$\begin{aligned} x' &= x/z \\ y' &= y/z \end{aligned} \tag{B.3}$$

$$\begin{aligned} u &= f_x \cdot x' + c_x \\ v &= f_y \cdot y' + c_y \end{aligned} \tag{B.4}$$

To take distortion into account, instead of directly calculating the pixel coordinate (u, v) as in Equation B.4, the radial distortion coefficients $k_1, .., k_6$ and tangential distortion coefficients p_1, p_2 are introduced as

$$\begin{aligned} x'' &= x' \frac{1+k_1 r^2+k_2 r^4+k_3 r^6}{1+k_4 r^2+k_5 r^4+k_6 r^6} + 2p_1 x' y' + p_2(r^2+2x'^2) \\ y'' &= y' \frac{1+k_1 r^2+k_2 r^4+k_3 r^6}{1+k_4 r^2+k_5 r^4+k_6 r^6} + p_1(r^2+2y'^2) + 2p_2 x' y' \\ & \text{with } r^2 = x'^2 + y'^2. \end{aligned} \tag{B.5}$$

The resulting pixel coordinates are then calculated as

$$\begin{aligned} u &= f_x * x'' + c_x \\ v &= f_y * y'' + c_y \end{aligned} . \tag{B.6}$$

Acronyms

AABB Axis-Aligned Bounding Box.

AR Augmented Reality.

CAD Computer Aided Design.

CUDA Compute Unified Device Architecture.

DLR German Aerospace Center.

DoF Degrees of Freedom.

FDA Food and Drug Administration.

FoV Field of View.

GPGPU General-Purpose Computing on Graphics Processing Unit.

GPU Graphics Processing Unit.

GUI Graphical User Interface.

HMD Head-Mounted Display.

IQR Interquartile Range.

IR Infrared.

ISO International Organization for Standardization.

LBR Light Weight Robot.

LoS Line of Sight.

MIRS Minimally Invasive Robotic Surgery.

MIS Minimally Invasive Surgery.

MR Mixed Reality.

OpenCV Open Source Computer Vision.

OR Operating Room.

OTS Optical Tracking System.

PCA Principal Component Analysis.

PCL Point Cloud Library.

PMD Photonic Mixer Device.

POV Point of View.

RANSAC Random Sample Consensus.

RASD Robot Assisted Surgical Device.

ROI Region of Interest.

ROS Robot Operating System.

RPC Remote Procedure Call.

SAFROS Patient Safety in Robotic Surgery.

SAR Spatial Augmented Reality.

SBI Suppression of Backlight Illumination.

SFF Small Form Factor.

SfP Shape from Polarization.

SNR Signal-to-Noise-Ratio.

SoC System on a Chip.

SPI Spatial Phase Imaging.

SSI Surgical Site Infection.

SVG Scalable Vector Graphics.

ToF Time-of-Flight.

VR Virtual Reality.

List of Figures

List of Tables

Bibliography

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres Solver. http://ceres-solver.org. [Online; accessed 10.12.2015].

[2] Nabeeha Ahmad, Ahmed A. Hussein, Lora Cavuoto, Mohamed Sharif, Jenna C. Allers, Nobuyuki Hinata, Basel Ahmad, Justen G. Kozlowski, Zishan Hashmi, Ann Bisantz, and Khurshid A. Guru. Ambulatory Movements, Team Dynamics and Interactions during Robot-Assisted Surgery. *BJU international*, (118):132–139, 2016.

[3] A. Alarcon and R. Berguer. A comparison of operating room crowding between open and laparoscopic operations. *Surgical Endoscopy*, 10(9):916–919, 1996.

[4] Justin Barad. Controlling Augmented Reality in the Operating Room: A Surgeon's Perspective. http://www.medgadget.com/2015/10/controlling-augmented-reality-operating-room-surgeons-perspective.html. [Online; accessed 12.01.2016].

[5] Blair A. Barbour and Himanshu Vajaria. Spatial Phase Imaging. In *Stereoscopic 3D for Media and Entertainment, SMPTE International Conference on*, pages 1–12, 2010.

[6] Basler. Engineering Sample of Time-of-Flight Cameras: Data Sheet. http://www.baslerweb.com/media/documents/BAS1508_ToF_EN__web.pdf, 12.10.2015. [Online; accessed 06.01.2016].

[7] Sebastian Bauer, Alexander Seitel, Hannes Hofmann, Tobias Blum, Jakob Wasza, Michael Balda, Hans-Peter Meinzer, Nassir Navab, Joachim Hornegger, and Lena Maier-Hein. Real-Time Range Imaging in Health Care: A Survey. In Marcin Grzegorzek, editor, *Time-of-Flight and depth imaging*, volume 8200 of *Lecture notes in computer science, state-of-the-art survey*, pages 228–254. Springer, Heidelberg, 2013.

[8] E. Bauzano, A. Fernández-Iribar, M.C. López-Casado, J. Klein, A. Rentería, and V.F. Muñoz-Martínez. Integración de Dispositivos en un Robot Quirúrgico teleoperado mediante ROS. In *Actas de las XXXVI Jornadas de Automática*, pages 815–822, 2015.

[9] Tim Beyl. *Workflow-based Context-aware Control of Surgical Robots*. PhD thesis, Karlsruher Institut für Technologie, Karlsruhe, 2015.

[10] Tim Beyl, Philip Nicolai, Jörg Raczkowsky, Heinz Wörn, Mirko D. Comparetti, and Elena de Momi. Multi kinect people detection for intuitive and safe human robot cooperation in the operating room. In *Advanced Robotics (ICAR), 2013 16th International Conference on*, pages 1–6, 2013.

[11] Tim Beyl, Luzie Schreiter, Philip Nicolai, Jörg Raczkowsky, and Heinz Wörn. 3D Perception Technologies for Surgical Operating Theatres. In J. D. Westwood, S. W. Westwood, and L. Felländer-Tsai, editors, *Medicine Meets Virtual Reality 22: NextMed / MMVR22*, pages 45–50. IOS Press, 2016.

[12] Andreas Bihlmaier, Tim Beyl, Philip Nicolai, Mirko Kunze, Julien Mintenbeck, Luzie Schreiter, Thorsten Brennecke, Jessica Hutzl, Jörg Raczkowsky, and Heinz Wörn. ROS-based Cognitive Surgical Robotics. In Anis Koubaa, editor, *Robot Operating System (ROS)*. Springer, [S.l.], 2016.

[13] Oliver Bimber and Ramesh Raskar. Modern approaches to augmented reality. In John Finnegan and Dave Shreiner, editors, *ACM SIGGRAPH 2006 Courses*, page 1.

[14] Bluetechnix GmbH. Argos3D - P100: Time-of-Flight Depth Sensor. http://datasheets.bluetechnix.at/goto/Argos/3D/P/100/Argos3D_P100_OV.pdf, 29.08.2013. [Online; accessed 15.12.2015].

[15] Felix Bork, Bernhard Fuers, Anja-Katharina Schneider, Francisco Pinto, Christoph Graumann, and Nassir Navab. Auditory and Visio-Temporal Distance Coding for 3-Dimensional Perception in Medical Augmented Reality. In *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*, pages 7–12, 2015.

[16] D. C. Bosanquet, C. N. Jones, N. Gill, P. Jarvis, and M. H. Lewis. Laminar flow reduces cases of surgical site infections in vascular patients. *Annals of the Royal College of Surgeons of England*, 95(1):15–19, 2013.

[17] G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

[18] Timo Breuer, Christoph Bodensteiner, and Michael Arens. Low-cost commodity depth sensor comparison and accuracy analysis. In Gary Kamerman, Ove Steinvall, Gary J. Bishop, Ainsley Killey, and John D. Gonglewski, editors, *SPIE Security + Defence*, SPIE Proceedings, page 92500G. SPIE, 2014.

[19] Julie Carmigniani and Borko Furht. Augmented Reality: An Overview. In Borivoje Furht, editor, *Handbook of augmented reality*, pages 3–46. Springer, New York, NY, 2011.

[20] Paul Chojecki and Ulrich Leiner. Berührungslose Gestik-Interaktion im Operationssaal: Touchless Gesture-Interaction in the Operating Room. *i-com*, 8(1), 2009.

[21] D. Claus and A. W. Fitzgibbon. A Rational Function Lens Distortion Model for General Cameras. In *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Computer Society Conference on*, pages 213–219, 2005.

[22] Colleen Culbertson (Intel). Introducing the Intel® RealSense™R200 Camera (world facing). https://software.intel.com/en-us/articles/realsense-r200-camera. [Online; accessed 06.01.2016].

[23] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M. Cortelazzo. *Time-of-flight cameras and Microsoft Kinect*. Springer briefs in electrical and computer engineering. Springer, New York, NY, 2012.

[24] A. de Luca, Alin Albu-Schaffer, Sami Haddadin, and Gerd Hirzinger. Collision Detection and Safe Reaction with the DLR-III Lightweight Manipulator Arm. In *Intelligent Robots and Systems (IROS), 2006 IEEE/RSJ International Conference on*. Institute of Electrical and Electronic Engineers, 2006.

[25] Ryan M. Dickey, Neel Srikishen, Larry I. Lipshultz, Philippe E. Spiess, Rafael E. Carrion, and Tariq S. Hakky. Augmented reality assisted surgery: a urologic training tool. *Asian Journal of Andrology*, 2015.

[26] Frank Dittrich, Stephan Puls, and Heinz Wörn. A Modular Cognitive System for Safe Human Robot Collaboration: Design, Implementation and Evaluation. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, page 10. IEEE, 2013.

[27] Benjamin J. Dixon, Michael J. Daly, Harley H. L. Chan, Allan Vescan, Ian J. Witterick, and Jonathan C. Irish. Inattentional blindness increased with augmented reality surgical navigation. *American Journal of Rhinology & Allergy*, 28(5):433–437, 2014.

[28] C. R. Dressler, T. Neumuth, M. Fischer, O. Abri, and G. Strauss. Intraoperative Bedienung einer elektronischen Patientenakte durch den Operateur. *HNO*, 59(9):900–907, 2011.

[29] Georg Eggers, Tobias Salb, Harald Hoppe, Lüder Kahrs, Sassan Ghanai, Gunther Sudra, Jörg Raczkowsky, Rüdiger Dillmann, Heinz Wörn, Stefan Hassfeld, and Rüdiger Marmulla. Intraoperative Augmented Reality: The Surgeons View. In James D. Westwood, editor, *Medicine meets virtual reality 13*, volume v. 111 of *Studies in Health Technology and Informatics*, pages 123–125. IOS, Amsterdam, Washington, DC, 2005.

[30] Technische Universiteit Eindhoven. Better surgery with new surgical robot with force feedback. https://www.tue.nl/en/university/news-and-press/news/27-09-2010-better-surgery-with-new-surgical-robot-with-force-feedback/. [Online (embedded video); accessed 19.01.2016].

[31] ESPROS photonics corporation. DATASHEET - epc660: Data Sheet. ftp://ftp.espros.com/01_Chips/Datasheet_epc660-V1.06.pdf, 17.09.2015. [Online; accessed 06.01.2016].

[32] Francesco Fanfani, Giorgia Monterossi, Anna Fagotti, Cristiano Rossitto, Salvatore Gueli Alletti, Barbara Costantini, Valerio Gallotta, Luigi Selvaggi, Stefano Restaino, and Giovanni Scambia. The new robotic TELELAP ALF-X in gynecological surgery: single-center experience. *Surgical Endoscopy*, 30(1):215–221, 2016.

[33] FDA. November 2013 Medical Product Safety Network (MedSun) Small Sample Survey - Final Report for the da Vinci Surgical System. http://www.fda.gov/downloads/MedicalDevices/ProductsandMedicalProcedures/SurgeryandLifeSupport/ComputerAssistedSurgicalSystems/UCM374095.pdf, November 2013. 22.01.2016.

[34] Marco Feuerstein. *Augmented Reality in Laparoscopic Surgery: New Concepts and Methods for Intraoperative Multimodal Imaging and Hybrid Tracking in Computer Aided Surgery*. PhD thesis, Technische Universität München, Saarbrücken, 2007.

[35] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Computer Vision and Pattern Recognition (CVPR), 2001 IEEE Computer Society Conference on*, pages I–125–I–132, 2001.

[36] Centers for Disease Control and Prevention. Growth Chart - Percentile Data Files with LMS Values. http://www.cdc.gov/growthcharts/percentile_data_files.htm. [Online; accessed 07.02.2016].

[37] Deutsches Zentrum für Luft-und Raumfahrt. Robotiksystem MiroSurge. http://www.dlr.de/Portaldata/1/Resources/standorte/oberpfaffenhofen/aktuelles/aktuelles-2/mirosurge_01.jpg. [Online; accessed 13.12.2015].

[38] Jean-Sebastien Franco, Benjamin Petit, and Edmond Boyer. 3D Shape Cropping. In Michael Bronstein, Jean Favre, and Kai Hormann, editors, *Vision, Modeling and Visualization*, 2013.

[39] Kate Alicia Gavaghan, Sylvain Anderegg, Matthias Peterhans, Thiago Oliveira-Santos, and Stefan Weber. Augmented Reality Image Overlay Projection for Image Guided Open Liver Ablation of Metastatic Liver Cancer. In Cristian A. Linte, John Moore, Elvis Chen, and David Holmes III, editors, *Augmented Enviroments for Computer-Assisted Interventions*, volume 7264 of *Lecture Notes in Computer Science / Image Processing, Computer Vision, Pattern Recognition, and Graphics Ser*, pages 36–46. Springer, New York, 2012.

[40] Kate Alicia Gavaghan, Matthias Peterhans, Thiago Oliveira-Santos, and Stefan Weber. A portable image overlay projection device for computer-aided open liver surgery. *IEEE Transactions on Bio-Medical Engineering*, 58(6):1855–1864, 2011.

[41] Todor Georgiev, Zhan Yu, Andrew Lumsdaine, and Sergio Goma. Lytro camera technology: theory, algorithms, performance analysis. In Cees G. M. Snoek, Lyndon S. Kennedy, Reiner Creutzburg, David Akopian, Dietmar

Wüller, Kevin J. Matherson, Todor G. Georgiev, and Andrew Lumsdaine, editors, *IS&T/SPIE Electronic Imaging*, SPIE Proceedings, page 86671J. SPIE, 2013.

[42] M. A. Goldstraw, K. Patil, C. Anderson, P. Dasgupta, and R. S. Kirby. A selected review and personal experience with robotic prostatectomy: implications for adoption of this new technology in the United Kingdom. *Prostate cancer and prostatic diseases*, 10(3):242–249, 2007.

[43] Jürgen Graf, Stephan Puls, and Heinz Wörn. Incorporating Novel Path Planning Method into Cognitive Vision System for Safe Human-Robot Interaction. In *2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns (COMPUTATIONWORLD)*, pages 443–447, 2009.

[44] P. Grasgruber, J. Cacek, T. Kalina, and M. Sebera. The role of nutrition and genetics as key determinants of the positive height trend. *Economics and human biology*, 15:81–100, 2014.

[45] M. Green and J. V. Odom. *Forensic Vision with Application to Highway Safety*. Lawyers & Judges Publishing Company, 2008.

[46] A. Groch, S. Haase, M. Wagner, T. Kilgus, H. Kenngott, H.-P. Schlemmer, J. Hornegger, H.-P. Meinzer, and L. Maier-Hein. Optimierte endoskopische Time-of-Flight Oberflächenrekonstruktion durch Integration eines Struktur-durch-Bewegung-Ansatzes. In Thomas Tolxdorff, editor, *Bildverarbeitung für die Medizin 2012*, Informatik aktuell, pages 39–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[47] Ayse P. Gurses, Elizabeth A. Martinez, Laura Bauer, George Kim, Lisa H. Lubomski, Jill A. Marsteller, Priyadarshini R. Pennathur, Chris Goeschel, Peter J. Pronovost, and David Thompson. Using human factors engineering to improve patient safety in the cardiovascular operating room. *Work (Reading, Mass.)*, 41 Suppl 1:1801–1804, 2012.

[48] G. S. Guthart and J. K. Salisbury. The Intuitive™ Telesurgery System: Overview and Application. In *Robotics and Automation (ICRA), 2000 IEEE International Conference on*, pages 618–621, 2000.

[49] Severine Habert, Jose Gardiazabal, Pascal Fallavollita, and Nassir Navab. RGBDX: First Design and Experimental Validation of a Mirror-Based RGBD X-ray Imaging System. In *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*, pages 13–18, 2015.

[50] Ulrich Hagn, R. Konietschke, A. Tobergte, M. Nickl, S. Jörg, B. Kübler, G. Passig, M. Gröger, F. Fröhlich, U. Seibold, L. Le-Tien, A. Albu-Schäffer, A. Nothhelfer, F. Hacker, M. Grebenstein, and G. Hirzinger. DLR MiroSurge: a versatile system for research in endoscopic telesurgery. *International Journal of Computer Assisted Radiology and Surgery*, 5(2):183–193, 2010.

[51] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with Microsoft Kinect sensor: a review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[52] Maria Hänel. *A Matter of Perspective - Three-dimensional Placement of Multiple Cameras to Maximize their Coverage*. PhD thesis, Universität Bayreuth, Bayreuth, 24.04.2015.

[53] Blake Hannaford, Jacob Rosen, Diana W. Friedman, Hawkeye King, Phillip Roan, Lei Cheng, Daniel Glozman, Ji Ma, Sina Nia Kosari, and Lee White. Raven-II: an open platform for surgical robotics research. *IEEE Transactions on Bio-Medical Engineering*, 60(4):954–959, 2013.

[54] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Horaud. *Time-of-Flight Cameras: Principles, Methods and Applications*. SpringerBriefs in Computer Science. Springer, London, 2013.

[55] Jason D. Harry, David Scott Ackerson, and Francois I. Luks. Systems and Methods for Measuring Mechanical Properties of Deformable Materials, 2011.

[56] Andrew N. Healey and Jonathan Benn. Teamwork enables remote surgical control and a new model for a surgical system emerges. *Cognition, Technology & Work*, 11(4):255–265, 2009.

[57] Heptagon. Sensing through Light Products — Heptagon. http://hptg.com/product/#imaging. [Online; accessed 06.01.2016].

[58] Heptagon. Swissranger series discontinued. http://hptg.com/industrial/. [Online; accessed 15.12.2015].

[59] D. M. Herron and M. Marohn. A consensus document on robotic surgery. *Surgical Endoscopy*, 22(2):313–25; discussion 311–2, 2008.

[60] Mathias Hoeckelmann, Imre J. Rudas, Paolo Fiorini, Frank Kirchner, and Tamas Haidegger. Current Capabilities and Development Potential in Surgical Robotics. *International Journal of Advanced Robotic Systems*, page 1, 2015.

[61] Human Science Group. Collision and injury criteria when working with collaborative robots: Health & Safety Executive (HSE) Research Report RR906. http://www.hse.gov.uk/research/rrpdf/rr906.pdf, Derbyshire, 2012.

[62] Jörn Hurtienne. *Image schemas and design for intuitive use: Exploring new guidance for user interface design: Exploring New Guidance for User Interface Design*. PhD thesis, Technische Universität Berlin, Berlin, 01.03.2011.

[63] Jessica Hutzl, Andreas Bihlmaier, Martin Wagner, Hannes Gotz Kenngott, Beat Peter Muller, and Heinz Wörn. Knowledge-based workspace optimization of a redundant robot for minimally invasive robotic surgery (MIRS).

In *Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on*, pages 1403–1408.

[64] Jessica Hutzl, Andreas Bihlmaier, Oliver Weede, and Heinz Wörn. An Automated Instrument as Component of a Cognitive Medical Robotic System for Minimal Invasive Surgery. *Conference of The International Society for Medical Innovation and Technology (iSMIT), Baden-Baden, Germany, September 5 - 7, 2013*, page 3, 2013.

[65] Jessica Hutzl and Heinz Wörn. Spatial probability distribution for port planning in minimal invasive robotic surgery (MIRS). In *Automation, Robotics and Applications (ICARA 2015), 2015 6th International Conference on*, pages 204–210, 2015.

[66] IEC. About IEC. http://www.iec.ch/about/. [Online; accessed 02.12.2015].

[67] Infineon. REAL3™ image sensor family: Data Sheet. http://www.infineon.com/dgdl/Infineon-REAL3+Image+Sensor+Family-PB-v01_00-EN.PDF?fileId=5546d462518ffd850151a0afc2302a58, 14.12.2015. [Online; accessed 06.01.2016].

[68] Swedish Standards Institute. Robots and robotic devices. http://www.sis.se/popup/iso/isotc184sc2/about_work_safety_for_medical.asp. [Online; accessed 11.12.2015].

[69] Intel. Intel® RealSense™ SDK 2015 R5 Documentation. https://software.intel.com/sites/landingpage/realsense/camera-sdk/v1.1/documentation/html/index.html?ivcamaccuracy_device_pxccapture.html. [Online; accessed 06.01.2016].

[70] Intuitive Surgical. da Vinci Xi Patient-side Cart. http://www.intuitivesurgical.com/company/media/images/systems-si/000797_si_patient_cart_arms_together.jpg. [Online; accessed 10.01.2016].

[71] Intuitive Surgical. da Vinci Xi Patient-side Cart. http://www.intuitivesurgical.com/company/media/images/xi/patient-cart-72dpi.png. [Online; accessed 10.01.2016].

[72] Intuitive Surgical. Investor Presentation Q4 2015. http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9MjcyMDc0fENoaWxkSUQ9LTF8VHlwZT0z&t=1. [Online; accessed 12.12.2015].

[73] ISO. About ISO. http://www.iso.org/iso/home/about.htm. [Online; accessed 02.12.2015].

[74] ISO. ISO 9241-11:1998: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. Published, International Organization for Standardization, Geneva, Switzerland, 19. March 1998.

[75] ISO. ISO 10218-1:2011: Robots and robotic devices – Safety requirements for industrial robots – Part 1: Robots. Published, International Organization for Standardization, Geneva, Switzerland, 01. July 2011.

[76] ISO. ISO 10218-2:2011: Robots and robotic devices – Safety requirements for industrial robots – Part 2: Robot systems and integration. Published, International Organization for Standardization, Geneva, Switzerland, 01. July 2011.

[77] ISO. ISO 8373:2012: Robots and robotic devices – Vocabulary. Published, International Organization for Standardization, Geneva, Switzerland, 01. March 2012.

[78] ISO. ISO 13482:2014: Robots and robotic devices – Safety requirements for personal care robots. Published, International Organization for Standardization, Geneva, Switzerland, 03. February 2014.

[79] ISO. IEC/NP 80601-2-77: Medical Electrical Equipment – Part 2-77: Particular requirements for the basic safety and essential performance of medical robots for surgery. Proposal, International Organization for Standardization, Geneva, Switzerland, 22. May 2015.

[80] ITU. About ITU. http://www.itu.int/en/about/Pages/default.aspx. [Online; accessed 02.12.2015].

[81] Mithun George Jacob, Yu-Ting Li, George A. Akingba, and Juan P. Wachs. Collaboration with a Robotic Scrub Nurse. *Communications of the ACM*, 56(5):68, 2013.

[82] Mithun George Jacob, Yu-Ting Li, and Juan P. Wachs. Gestonurse: A multimodal robotic scrub nurse. In Holly Yanco, Aaron Steinfeld, Vanessa Evers, and Odest Chadwicke Jenkins, editors, *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, page 153, 2012.

[83] Joshua Nipper. Robotically-Assisted Surgical Devices (RASD): An FDA Perspective. In U.S. Food and Drug Administration, editor, *Public Workshop - Robotically-Assisted Surgical Devices: Challenges and Opportunities*, 2015.

[84] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3D: High-Quality Depth Sensing with Polarization Cues. In *Computer Vision (ICCV), 2015 International Conference on*. IEEE – Institute of Electrical and Electronics Engineers, 2015.

[85] Lüder Kahrs, Harald Hoppe, Georg Eggers, Jörg Raczkowsky, Rüdiger Marmulla, and Heinz Wörn. Visualization of Surgical 3D Information with Projector–based Augmented Reality. In James D. Westwood, editor, *Medicine meets virtual reality 13*, volume v. 111 of *Studies in Health Technology and Informatics*, pages 243–246. IOS, Amsterdam, Washington, DC, 2005.

[86] H. G. Kenngott, J. J. Wünscher, M. Wagner, A. Preukschas, A. L. Wekerle, P. Neher, S. Suwelack, S. Speidel, F. Nickel, D. Oladokun, L. Maier-Hein, R. Dillmann, H. P. Meinzer, and B. P. Müller-Stich. OpenHELP (Heidelberg laparoscopy phantom): Development of an open-source surgical evaluation and training tool. *Surgical Endoscopy*, 29(11):3338–3347, 2015.

[87] Marta Kersten-Oertel, Pierre Jannin, and D. Louis Collins. The state of the art of visualization in mixed reality image guided surgery. *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society*, 37(2):98–112, 2013.

[88] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, 12(2):1437–1454, 2012.

[89] Sami G. Kilic, M. Faruk Kose, and Kubilay Ertan. *Robotic surgery: Practical examples in gynecology*. De Gruyter, Berlin, 2013.

[90] Young Min Kim, Derek Chan, Christian Theobalt, and Sebastian Thrun. Design and calibration of a multi-view TOF sensor fusion system. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 1–7, 2008.

[91] Bojan Kocev, Felix Ritter, and Lars Linsen. Projector-based surgeon-computer interaction on deformable surfaces. *International journal of computer assisted radiology and surgery*, 9(2):301–312, 2014.

[92] Thomas Köhler, Sven Haase, Sebastian Bauer, Jakob Wasza, Thomas Kilgus, Lena Maier-Hein, Christian Stock, Joachim Hornegger, and Hubertus Feußner. Multi-sensor super-resolution for hybrid range imaging with application to 3-D endoscopy and open surgery. *Medical Image Analysis*, 24(1):220–234, 2015.

[93] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time–of–Flight Cameras in Computer Graphics. *Computer Graphics Forum*, 29, 2010.

[94] Rainer Konietschke, Tim Bodenmüller, Christian Rink, Andrea Schwier, Berthold Bäuml, and Gerd Hirzinger. Optimal setup of the DLR MiroSurge telerobotic system for minimally invasive surgery. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 3435–3436, 2011.

[95] Anis Koubaa, editor. *Robot Operating System (ROS): The Complete Reference*. Springer, [S.l.], 1st ed. 2016 edition, 2016.

[96] Robert Krempien, Harald Hoppe, Lüder Kahrs, Sascha Daeuber, Oliver Schorr, Georg Eggers, Marc Bischof, Marc W. Munter, Juergen Debus, and Wolfgang Harms. Projector-based augmented reality for intuitive intraoperative guidance in image-guided 3D interstitial brachytherapy. *International journal of radiation oncology, biology, physics*, 70(3):944–952, 2008.

[97] Franziska Kühn and Martin Leucker. OR.NET: Safe Interconnection of Medical Devices. In Jeremy Gibbons and Wendy MacCaull, editors, *Foundations of health information engineering and systems*, volume 8315 of *LNCS sublibrary. SL 2, Programming and software engineering*, pages 188–198. Springer, 2014.

[98] A. Kurmann, M. Peter, F. Tschan, K. Mühlemann, D. Candinas, and G. Beldi. Adverse effect of noise in the operating theatre on surgical-site infection. *The British journal of surgery*, 98(7):1021–1025, 2011.

[99] Stawros Ladikos. *Real-Time Multi-View 3D Reconstruction for Interventional Environments*. PhD thesis, Technische Universität München, 2011.

[100] F. Lai and E. Entin. Robotic Surgery and the Operating Room Team. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(11):1070–1073, 2005.

[101] Pablo Lamata, Wajid Ali, Alicia Cano, Jordi Cornella, Jerome Declerck, Ole J. Elle, Adinda Freudenthal, Hugo Furtado, Denis Kalkofen, Edvard Naerum, Eigil Samset, Patricia Sánchez-Gonzalez, Francisco M. Sánchez-Margallo, Dieter Schmalstieg, Mauro Sette, Thomas Stüdeli, Jos Vander Sloten, and Enrique J. Gómez. Augmented Reality for Minimally Invasive Surgery: Overview and Some Recent Advances. In Soha Maad, editor, *Augmented reality*. InTech, Rijek, Crotia, 2010.

[102] Benjamin Langmann, Klaus Hartmann, and Otmar Loffeld. Depth Camera Technology Comparison and Performance Evaluation. In *Pattern Recognition Applications and Methods, 2012 International Conference on*, pages 438–444, 2012.

[103] Bettina Laugwitz, Theo Held, and Martin Schrepp. *Construction and evaluation of a user experience questionnaire*. Springer, 2008.

[104] Damien Lefloch, Rahul Nair, Frank Lenzen, Henrik Schäfer, Lee Streeter, Michael J. Cree, Reinhard Koch, and Andreas Kolb. Technical Foundation and Calibration Methods for Time-of-Flight Cameras. In Marcin Grzegorzek, editor, *Time-of-Flight and depth imaging*, volume 8200 of *Lecture notes in computer science, state-of-the-art survey*, pages 3–24. Springer, Heidelberg, 2013.

[105] Wim Lemkens, Prabhdeep Kaur, Koen Buys, Peter Slaets, Tinne Tuytelaars, and Joris de Schutter. Multi RGB-D camera setup for generating large 3D point clouds. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1092–1099. IEEE, 2013.

[106] Cristian A. Linte, Katherine P. Davenport, Kevin Cleary, Craig Peters, Kirby G. Vosburgh, Nassir Navab, Philip Eddie Edwards, Pierre Jannin, Terry M. Peters, David R. Holmes, and Richard A. Robb. On mixed reality environments for minimally invasive therapy guidance: systems architecture, successes and challenges in their implementation from laboratory to

clinic. *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society*, 37(2):83–97, 2013.

[107] Nicolas Loy Rodas and Nicolas Padoy. Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, Monte Carlo simulations and wireless dosimeters. *International journal of computer assisted radiology and surgery*, 10(8):1181–1191, 2015.

[108] Lytro. Lytro Illum. https://store.lytro.com/products/lytro-illum. [Online; accessed 11.01.2016].

[109] Kamal Maheshwari. Operating Room Design Manual: Chapter 9: Room Ventilation Systems. https://www.asahq.org/resources/resources-from-asa-committees/operating-room-design-manual, 2012. [Online; accessed 05.02.2016].

[110] Aruna D. Mane, Sirkazi Mohd Arif, and Waleed Abdu Rahiman. An Advanced Robot -Robin Heart (A Surgeon without Hand Tremor). *International Journal of Engineering and Advanced Technology (IJEAT)*, 2013(5):242–251, 2013.

[111] R. Marmulla, Harald Hoppe, J. Mühling, and G. Eggers. An augmented reality system for image-guided surgery. *International Journal of Oral and Maxillofacial Surgery*, 34(6):594–596, 2005.

[112] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge Mass. u.a., 2010.

[113] Manuel Martinez and Rainer Stiefelhagen. Kinect Unleashed: Getting Control over High Resolution Depth Maps Vision Applications. In *Machine Vision Applications (MVA 2013), Proceedings of the 13. IAPR International Conference on*, pages 247–250, 2013.

[114] Manuel Martinez and Rainer Stiefelhagen. Kinect Unbiased. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 5791–5795, 2014.

[115] Sergio E. Martinez Herrera, Abed Malti, Olivier Morel, and Adrien Bartoli. Shape-from-Polarization in laparoscopy. In *Biomedical Imaging (ISBI 2013), 2013 IEEE 10th International Symposium on*, pages 1412–1415, 2013.

[116] M. J. Matarić. *The Robotics Primer*. MIT Press, 2007.

[117] Bjoern Matthias. The Role of Collision Experiments in Safety Standardization and in the Characterization of Collaborative Robots, Systems and Applications. http://www.researchgate.net/profile/Bjoern_Matthias/publication/282778869_The_Role_of_Collision_Experiments_in_Safety_Standardization_and_in_the_Characterization_of_Collaborative_Robots_Systems_and_Applications/links/561c345d08aea80367243a4d.pdf.

[118] John J. Meehan and Anthony Sandler. Pediatric robotic surgery: A single-institutional review of the first 100 consecutive cases. *Surgical Endoscopy*, 22(1):177–182, 2008.

[119] Mesa Imaging AG. SR4000 Data Sheet. http://downloads.mesa-imaging.ch/dlm.php?fname=pdf/SR4000_Data_Sheet_rev1.5.pdf, 15.05.2009. [Online; accessed 15.12.2015].

[120] Microsoft. Introduccing the Microsoft HoloLens Development Edition. https://www.microsoft.com/microsoft-hololens/en-us/development-edition#de-accordion-tech-spec-panel. [Online; accessed 08.04.2016].

[121] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, 77(12):1321–1329, 1994.

[122] Stephen Miller, Alex Teichman, and Sebastian Thrun. Unsupervised extrinsic calibration of depth sensors in dynamic scenes. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2695–2702. IEEE, 2013.

[123] Stefan Mitrasinovic, Elvis Camacho, Nirali Trivedi, Julia Logan, Colson Campbell, Robert Zilinyi, Bryan Lieber, Eliza Bruce, Blake Taylor, David Martineau, Emmanuel L. P. Dumont, Geoff Appelboom, and E. Sander Connolly. Clinical and surgical applications of smart glasses. *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine*, 23(4):381–401, 2015.

[124] Holger Mönnich, Philip Nicolai, Tim Beyl, Jörg Raczkowsky, and Heinz Wörn. A supervision system for the intuitive usage of a telemanipulated surgical robotic setup. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 449–454, 2011.

[125] Holger Mönnich, Philip Nicolai, Jörg Raczkowsky, and Heinz Wörn. A semi-autonomous robotic teleoperation surgery setup. *International Journal of Computer Assisted Radiology and Surgery*, 6(S1):132–133, 2011.

[126] Carlos Morato, Krishnanand N. Kaipa, Boxuan Zhao, and Satyandra K. Gupta. Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-time Human Tracking. *Journal of Computing and Information Science in Engineering*, 14(1):11006, 2014.

[127] Daniel Moreno and Gabriel Taubin. Simple, Accurate, and Robust Projector-Camera Calibration. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 464–471, 2012.

[128] Michael Müller, Marie-Claire Rassweiler, Jan Klein, Alexander Seitel, Matthias Gondan, Matthias Baumhauer, Dogu Teber, Jens J. Rassweiler,

Hans-Peter Meinzer, and Lena Maier-Hein. Mobile augmented reality for computer-assisted percutaneous nephrolithotomy. *International Journal of Computer Assisted Radiology and Surgery*, 8(4):663–675.

[129] Matteo Munaro, Alex Horn, Randy Illum, Jeff Burke, and Radu Bogdan Rusu. OpenPTrack: People Tracking for Heterogeneous Networks of Color-Depth Cameras. In *3D Robot Perception with Point Cloud Library at Intelligent Autonomous Systems (IAS Workshop), 2013 First International Workshop on*, pages 235–247, 2013.

[130] Anja Naumann, Jörn Hurtienne, Johann Habakuk Israel, Carsten Mohs, Martin Christof Kindsmüller, Herbert A. Meyer, and Steffi Hußlein. Intuitive Use of User Interfaces: Defining a Vague Concept. In Don Harris, editor, *Engineering Psychology and Cognitive Ergonomics*, volume 4562 of *Lecture Notes in Computer Science*, pages 128–136. Springer Berlin Heidelberg, 2007.

[131] Nassir Navab, Tobias Blum, Lejing Wang, Aslı Okur, and Thomas Wendler. First Deployments of Augmented Reality in Operating Rooms. *Computer*, 45(7):48–55, 2012.

[132] Stefan Escaida Navarro, Maximiliano Marufo, Yitao Ding, Stephan Puls, Dirk Göger, Björn Hein, and Heinz Wörn. Methods for safe human-robot-interaction using capacitive tactile proximity sensors. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1149–1154. IEEE, 2013.

[133] Philip Nicolai and Jörg Raczkowsky. Operation Room Supervision for Safe Robotic Surgery with a multi 3D-Camera Setup, 29.09.2011.

[134] Philip Nicolai, Jörg Raczkowsky, and Heinz Wörn. Continuous Pre-Calculation of Human Tracking with Time-delayed Ground-truth - A Hybrid Approach to Minimizing Tracking Latency by Combination of Different 3D Cameras. In *Informatics in Control, Automation and Robotics (ICINCO), 12th International Conference on*, pages 121–130, 2015.

[135] Philip Nicolai, Jörg Raczkowsky, and Heinz Wörn. Model-Free (Human) Tracking Based on Ground Truth with Time Delay: A 3D Camera Based Approach for Minimizing Tracking Latency and Increasing Tracking Quality. In Joaquim Filipe, Oleg Gusikhin, Kurosh Madani, and Jurek Sasiadek, editors, *Informatics in control, automation and robotics*, volume 383 of *Lecture Notes in Electrical Engineering*, pages 247–266. Springer, Cham, 2016.

[136] odos imaging. OI-VS-1000 - time-of-flight 3D vision system: Data Sheet. http://www.odos-imaging.com/?ddownload=4124, 17.03.2015. [Online; accessed 06.01.2016].

[137] Kenton O'Hara, Neville Dastur, Tom Carrell, Gerardo Gonzalez, Abigail Sellen, Graeme Penney, Andreas Varnavas, Helena Mentis, Antonio Criminisi, Robert Corish, and Mark Rouncefield. Touchless interaction in surgery. *Communications of the ACM*, 57(1):70–77, 2014.

[138] OpenCV. Camera calibration and 3d reconstruction. http://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html. [Online; accessed 18.02.2016].

[139] Orbbec. Persee — Orbbec. https://orbbec3d.com/product-persee/. [Online; accessed 11.01.2016].

[140] Björn Ostermann. *Entwicklung eines Konzepts zur sicheren Personenerfassung als Schutzeinrichtung an kollaborierenden Robotern*. PhD thesis, Wuppertal, Univ., Diss., 2014, 2014.

[141] Hanhoon Park, Moon-Hyun Lee, Sang-Jun Kim, and Jong-Il Park. Surface-Independent Direct-Projected Augmented Reality. In P. J. Narayanan, Shree K. Nayar, and Heung-Yeung Shum, editors, *Computer vision - ACCV 2006*, volume 3852 of *Lecture Notes in Computer Science*, pages 892–901. Springer, Berlin, 2006.

[142] Andrew Payne, Andy Daniel, Anik Mehta, Barry Thompson, Cyrus S. Bamji, Dane Snow, Hideaki Oshima, Larry Prather, Mike Fenton, Lou Kordus, Pat O'Connor, Rich McCauley, Sheethal Nayak, Sunil Acharya, Swati Mehta, Tamer Elkhatib, Thomas Meyer, Tod O'Dwyer, Travis Perry, Vei-Han Chan, Vincent Wong, Vishali Mogallapu, William Qian, and Zhanping Xu. A 512×424 CMOS 3D Time-of-Flight image sensor with multi-frequency photo-demodulation up to 130MHz and 2GS/s ADC. In *Solid-State Circuits (ISSCC), 2014 IEEE International Conference on*, pages 134–135, 2014.

[143] Priyadarshini R. Pennathur, David Thompson, James H. Abernathy, Elizabeth A. Martinez, Peter J. Pronovost, George R. Kim, Laura C. Bauer, Lisa H. Lubomski, Jill A. Marsteller, and Ayse P. Gurses. Technologies in the wild (TiW): human factors implications for patient safety in the cardiovascular operating room. *Ergonomics*, 56(2):205–219, 2013.

[144] Jochen Penne, Kurt Höller, Michael Stürmer, Thomas Schrauder, Armin Schneider, Rainer Engelbrecht, Hubertus Feußner, Bernhard Schmauss, and Joachim Hornegger. Time-of-Flight 3-D Endoscopy. In Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, volume 5761 of *Lecture Notes in Computer Science*, pages 467–474. Springer Berlin Heidelberg, 2009.

[145] Jochen Penne, Christian Schaller, Rainer Engelbrecht, Lena Maier-Hein, Bernhard Schmauss, Hans-Peter Meinzer, and Joachim Hornegger. Laparoscopic Quantitative 3D Endoscopy for Image Guided Surgery - Anwendungen, Aachen, Germany, March 14-16, 2010. In Thomas Martin Deserno, Heinz Handels, Hans-Peter Meinzer, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2010 - Algorithmen - Systeme - Anwendungen, Aachen, Germany, March 14-16, 2010*, volume 574 of *CEUR Workshop Proceedings*, pages 16–20. CEUR-WS.org, 2010.

[146] Christian Perwass and Lennart Wietzke. Single lens 3D-camera with extended depth-of-field. In *IS&T/SPIE Electronic Imaging*, SPIE Proceedings, page 829108. SPIE, 2012.

[147] Photon-X. All-in-One 3D Biometrics: Recognize multiple biometrics simultaneously. http://ideasorlando.com/wp-content/uploads/2013/04/Photon-X-Party.pdf, 04.04.2013. [Online; accessed 08.01.2016].

[148] Dario Piatti and Fulvio Rinaudo. SR-4000 and CamCube3.0 Time of Flight (ToF) Cameras: Tests and Comparison. *Remote Sensing*, 4(12):1069–1089, 2012.

[149] Pilz GmbH & Co. KG. Safe Camera System SafetyEye. https://www.pilz.com/download/open/Leaflet_SafetyEYE_EN_2014_05_low.pdf. 11.11.2015.

[150] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, volume 3, page 5, 2009.

[151] Rebecca Randell. Are robot surgeons in the operating theatre as safe as they could be? http://theconversation.com/are-robot-surgeons-in-the-operating-theatre-as-safe-as-they-could-be-45113, 28.07.2015. 22.01.2016.

[152] Rebecca Randell, Joanne Greenhalgh, Jon Hindmarsh, Dawn Dowding, David Jayne, Alan Pearman, Peter Gardner, Julie Croft, and Alwyn Kotze. Integration of robotic surgery into routine practice and impacts on communication, collaboration, and decision making: a realist process evaluation protocol. *Implementation science : IS*, 9:52, 2014.

[153] Shreedhar Rangappa, Mitul Tailor, Jon Petzing, Peter Kinnell, and Michael Jackson. The suitability of lightfield camera depth maps for coordinate measurement applications. In Antanas Verikas, Petia Radeva, and Dmitry Nikolaev, editors, *Eighth International Conference on Machine Vision*, SPIE Proceedings, page 987523. SPIE, 2015.

[154] Michael F. Rayo and Susan D. Moffatt-Bruce. Alarm system management: evidence-based guidance encouraging direct measurement of informativeness to improve alarm response. *BMJ quality & safety*, 24(4):282–286, 2015.

[155] Christopher Reardon, Huan Tan, Balajee Kannan, and Lynn DeRose. Towards safe robot-human collaboration systems using human pose detection. In *Technologies for Practical robot Applications (TePRA), 2015 IEEE International Conference on*, pages 1–6, 2015.

[156] Thorsten Ringbeck and Bianca Hagebeuker. A 3D Time of Flight Camera for Object Detection. In Armin Grün, editor, *Optical 3-D measurement techniques*, 2007.

[157] Nicolas Loy Rodas, Fernando Barrera, and Nicolas Padoy. Marker-Less AR in the Hybrid Room Using Equipment Detection for Camera Relocalization. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical image computing and computer-assisted intervention – MICCAI 2015*, volume 9349 of *LNCS sublibrary. SL 6, Image processing, computer vision, pattern recognition, and graphics*, pages 463–470. Springer, Cham, 2015.

[158] J. M. Rodriguez-Ramos, J. G. Marichal-Hernandez, J. P. Luke, J. Trujillo-Sevilla, M. Puga, M. Lopez, J. J. Fernandez-Valdivia, C. Dominguez-Conde, J. C. Sanluis, F. Rosa, V. Guadalupe, H. Quintero, C. Militello, L. F. Rodriguez-Ramos, R. Lopez, I. Montilla, and B. Femenia. New developments at CAFADIS plenoptic camera. In *Information Optics (WIO), 2011 10th Euro-American Workshop on*, pages 1–3, 2010.

[159] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4, 2011.

[160] Paul Rybski, Peter Anderson-Sprecher, Daniel Huber, Chris Niessl, and Reid Simmons. Sensor fusion for human safety in industrial workcells. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3612–3619, 2012.

[161] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3:137–150, 2013.

[162] Bernard Schmidt and Lihui Wang. Depth camera based collision avoidance via active robot control. *Journal of Manufacturing Systems*, 33(4):711–718, 2014.

[163] Sue Sendelbach and Marjorie Funk. Alarm fatigue: a patient safety concern. *AACN advanced critical care*, 24(4):378–386, 2013.

[164] Byung-Kuk Seo, Moon-Hyun Lee, Hanhoon Park, Jong-Il Park, and Young Soo Kim. Direct-Projected AR Based Interactive User Interface for Medical Surgery. In *17th International Conference on Artificial Reality and Telexistence (ICAT 2007)*, pages 105–112, 2007.

[165] SOFAR S.p.A. TELELAP ALF-X surgery robot presentation ENG. http://www.alf-x.com/en/. [Online; accessed 19.01.2016].

[166] Dominik Stengel, Thomas Wiedemann, and Birgit Vogel-Heuser. Efficient 3D voxel reconstruction of human shape within robotic work cells. In *Mechatronics and Automation (ICMA), 2012 IEEE International Conference on*, pages 1386–1392, 2012.

[167] STMicroelectronics. Micro-mirrors from STMicroelectronics Provide Precision in Perceptual Computing. http://www.st.com/web/en/press/p3657. [Online; accessed 06.01.2016].

[168] Jan Stühmer, Sebastian Nowozin, Andrew Fitzgibbon, Richard Szeliski, Travis Perry, Sunil Acharya, Daniel Cremers, and Jamie Shotton. Model-Based Tracking at 300Hz using Raw Time-of-Flight Observations. In *Computer Vision (ICCV), 2015 International Conference on*. IEEE – Institute of Electrical and Electronics Engineers, 2015.

[169] Dan Sunday. Intersection of rays and triangles. http://geomalgorithms.com/a06-_intersect-2.html. [Online; accessed 11.02.2016].

[170] Jefferey Too Chuan Tan, Feng Duan, Ryu Kato, and Tamio Arai. Safety Strategy for Human–Robot Collaboration: Design and Development in Cellular Manufacturing. *Advanced Robotics*, 24(5-6):839–860, 2010.

[171] Jeffrey Too Chuan Tan and Tamio Arai. Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing. In *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*, pages 1–6, 2011.

[172] J.-P. Tardif, S. Roy, and J. Meunier. Projector-based augmented reality in surgery without calibration. In *25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 548–551, 2003.

[173] Russell H. Taylor. Robotically Assisted Surgical Devices (RASD): Key Non-Clinical Performance Characteristics. In U.S. Food and Drug Administration, editor, *Public Workshop - Robotically-Assisted Surgical Devices: Challenges and Opportunities*, 2015.

[174] TedCas. Natural User Interfaces for healthcare: Product Brochure. http://www.tedcas.com/sites/default/files/TedCas-TedCube_Brochure.pdf, 27.02.2015. [Online; accessed 07.01.2016].

[175] Alex Teichman, Jake T. Lussier, and Sebastian Thrun. Learning to Segment and Track in RGBD. *IEEE Transactions on Automation Science and Engineering*, 10(4):841–852, 2013.

[176] Texas Instruments. OPT8320 3D Time-of-Flight Sensor: Data Sheet. http://www.ti.com/lit/ds/symlink/opt8320.pdf, 17.12.2015. [Online; accessed 05.01.2016].

[177] Tim Duncan (Intel). Can Your Webcam Do This? - Exploring the Intel® RealSense™3D Camera (F200). https://software.intel.com/en-us/blogs/2015/01/26/can-your-webcam-do-this. [Online; accessed 06.01.2016].

[178] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision algorithms*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer, Berlin and New York, 2000.

[179] Jocelyne Troccaz, U. Hagn, M. Nickl, S. Jörg, G. Passig, T. Bahls, A. Nothhelfer, F. Hacker, L. Le-Tien, A. Albu-Schäffer, R. Konietschke, M. Grebenstein, R. Warpup, R. Haslinger, M. Frommberger, and G. Hirzinger. The DLR MIRO: A versatile lightweight robot for surgical applications. *Industrial Robot: An International Journal*, 35(4):324–336, 2008.

[180] U.S. Food and Drug Administration. MAUDE Adverse Event Report: INTUITIVE SURGICAL DAVINCI SURGICAL ROBOT. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=1660340. [Online; accessed 11.03.2016].

[181] U.S. Food and Drug Administration. MAUDE Adverse Event Report: INTUITIVE SURGICAL, INC. DAVINCI SYSTEM, SURGICAL, COMPUTER CONTROLLED INSTRUMENT. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=2973586. [Online; accessed 11.03.2016].

[182] U.S. Food and Drug Administration. MAUDE Adverse Event Report: INTUITIVE SURGICAL,INC. DA VINCI SI SURGICAL SYSTEM ENDOSCOPIC INSTRUMENT CONTROL SYSTEM. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=2552349. [Online; accessed 11.03.2016].

[183] U.S. Food and Drug Administration. MAUDE Adverse Event Report: INTUITIVE SURGICAL,INC. DA VINCI SI SURGICAL SYSTEM ENDOSCOPIC INSTRUMENT CONTROL SYSTEM. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=3215235. [Online; accessed 11.03.2016].

[184] U.S. Food and Drug Administration. MAUDE Adverse Event Report: INTUITIVE SURGICAL,INC. DA VINCI SURGICAL SYSTEM ENDOSCOPIC INSTRUMENT CONTROL SYSTEM. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=2580527. [Online; accessed 11.03.2016].

[185] U.S. Food and Drug Administration, editor. *Public Workshop - Robotically-Assisted Surgical Devices: Challenges and Opportunities*, 2015.

[186] Linda van den Bedem. *Realization of a demonstrator slave for robotic minimally invasive surgery*. PhD thesis, Eindhoven University of Technology, Eindhoven, 2010.

[187] Kartik Venkataraman, Dan Lelescu, Jacques Duparré, Andrew McMahon, Gabriel Molina, Priyam Chatterjee, Robert Mullis, and Shree Nayar. PiCam: An Ultra-Thin High Performance Monolithic Camera Array. *ACM Transactions on Graphics*, 32(6):1–13, 2013.

[188] C. Vogel, M. Poggendorf, C. Walter, and N. Elkmann. Towards safe physical human-robot collaboration: A projection-based safety system. In *Intelligent*

Robots and Systems (IROS 2011), 2011 IEEE/RSJ International Conference on, pages 3355–3360, 2011.

[189] Christian Vogel, Christoph Walter, and Norbert Elkmann. A projection-based sensor system for safe physical human-robot collaboration. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 5359–5364. IEEE, 2013.

[190] Rong Wen. *Projection-based Spatial Augmented Reality for Interactive Visual Guidance in Surgery*. PhD thesis, National University of Singapore, Singapore, 2013.

[191] Rong Wen, Chee-Kong Chui, Sim-Heng Ong, Kah-Bin Lim, and Stephen Kin-Yong Chang. Projection-based visual guidance for robot-aided RF needle insertion. *International Journal of Computer Assisted Radiology and Surgery*, 8(6):1015–1025, 2013.

[192] Rong Wen, Binh P. Nguyen, Chin-Boon Chng, and Chee-Kong Chui. In Situ Spatial AR Surgical Planning using Projector-Kinect System. In Thang Huynh Quyet, Binh Nguyen Thanh, Tien Do Van, Marc Bui, and Son Ngo Hong, editors, *Proceedings of the Fourth Symposium on Information and Communication Technology*, pages 164–171, 2013.

[193] Rong Wen, Wei-Liang Tay, Binh P. Nguyen, Chin-Boon Chng, and Chee-Kong Chui. Hand gesture guided robot-assisted surgery based on a direct augmented reality interface. *Computer Methods and Programs in Biomedicine*, 116(2):68–80, 2014.

[194] Heinz Wörn and Harald Hoppe. Augmented Reality in the Operating Theatre of the Future. In Wiro J. Niessen and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 1195–1196. Springer, Berlin, Heidelberg, 2001.

[195] ximea. Real-time 3D camera: Data Sheet. http://www.ximea.com/files/brochures/PhoXI-Realtime-3D-camera-2015-brochure.pdf, 28.10.2014. [Online; accessed 06.01.2016].

[196] Theo Watson Zachary Lieberman, Arturo Castro, and Others. openFrameworks. http://openframeworks.cc. [Online; accessed 13.12.2015].

[197] N. Zeller, F. Quint, and U. Stilla. Calibration and accuracy analysis of a focused plenoptic camera. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:205–212, 2014.

[198] Xinran Zhang, Guowen Chen, and Hongen Liao. A high-accuracy surgical augmented reality system using enhanced integral videography image overlay. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 4210–4213, 2015.

[199] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.