Augmented Light Field Visualization and Real-Time Image Enhancement for Computer Assisted Endoscopic Surgery

Der Technischen Fakultät der
Universität Erlangen–Nürnberg
zur Erlangung des Grades

DOKTOR–INGENIEUR

vorgelegt von

Florian Vogt

Erlangen — 2006

Als Dissertation genehmigt von
der Technischen Fakultät der
Universität Erlangen-Nürnberg

Tag der Einreichung:	31.10.2005
Tag der Promotion:	12.7.2006
Dekan:	Prof. Dr. A. Leipertz
Berichterstatter:	Prof. Dr. H. Niemann
	Prof. Dr. N. Navab

To my family

Acknowledgements

I am deeply grateful that I had the opportunity to write this thesis while working at the Chair for Pattern Recognition within the project B6 of the Sonderforschungsbereich 603 (funded by Deutsche Forschungsgemeinschaft). Many people contributed to this work and I want to express my gratitude to all of them.

In particular, I thank my advisors Prof. Dr. Niemann, Prof. Dr. Paulus, Prof. Dr. Hohenberger, PD Dr. Schick, and Prof. Dr. Greiner for the idea for this project and the willingness to support me during my work. I additionally appreciate the good cooperation of the three involved institutions of the University Erlangen-Nuremberg: the Chair for Pattern Recognition, the Chair for Computer Graphics, and the Department of Surgery.

Besides my supervisors, my colleagues were ready to discuss any question anytime. I appreciate sharing my office with Jochen Schmidt and Ingo Scholz for several years. Their knowledge and advice helped me solve several problems and the three of us had a lot of fruitful discussions. Concerning medical questions I worked closely together with Dr. Sophie Krüger and I want to thank her for supporting me. Concerning questions on computer graphics, my colleagues Christian Vogelgsang and Marco Winter were a great help. I also want to acknowledge the technical support of Fritz Popp and Walter Fentze. At the beginning of my work there were days when I walked more than once into Christopher Drexler's office because of some C++ or Linux problem I could not solve myself. Thank you Chris for helping me. I supervised several students who contributed to this work. Especially, I want to mention Marco Weber who worked four years for me as student assistant and whose diploma thesis I finally supervised.

A large number of people read the manuscript, most of them read it completely although I did not force them to do so. Thank you Jochen, Ingo, Sophie, Christoph, Achim, Elmar, Dietrich, Beni, and Thomas. I am especially thankful that Sandra and my best friend Markus agreed to read the manuscript although they presumably understood only a tiny part of the theoretical considerations. I additionally want to thank Prof. Dr. Niemann and Prof. Dr. Navab for reviewing my thesis.

Some people supported me within the scope of a research cooperation: thank you Andrea Schenk (MeVis), Dr. Marc Schneberger (A.R.T.), PD Dr. Peter Hastreiter (Neurocenter), Dr. Wittenberg (Fraunhofer IIS), and Dr. Cavallaro (Institute of Radiology).

At last I want to thank my family, especially my beloved wife Uta, for supporting me during my years at university.

Florian Vogt

Contents

1 Introduction 1

1.1 Problem Statement and Medical Importance . . . 2
1.2 Contribution of this Work . . . 4
1.3 Image Modalities . . . 6
1.4 Data Fusion . . . 8
1.5 Outline . . . 10

2 State of the Art in Computer Assisted Endoscopic Surgery 11

2.1 Conventional Surgery . . . 12
2.2 Minimally Invasive Surgery . . . 14
2.3 Robot Assisted Interventions . . . 17
2.4 Image Enhancement . . . 20
2.5 Augmented Reality (AR) . . . 24
  2.5.1 3-D Visualization . . . 25
  2.5.2 Pose Determination . . . 27
  2.5.3 Virtual Data . . . 29
  2.5.4 Registration and Fusion . . . 29
  2.5.5 Comparison of Medical AR and 3-D Visualization Approaches . . . 30

3 Light Field Theory 33

3.1 Definition and Concept . . . 33
  3.1.1 Samples of the Plenoptic Function . . . 36
  3.1.2 Depth Maps and Confidence Maps . . . 38
3.2 Light Field Visualization . . . 39
  3.2.1 Two-Plane Light Fields . . . 39
  3.2.2 Free Form Light Fields . . . 43

3.2.3 Inclusion of Confidence Maps 48
3.3 Light Field Reconstruction 48
3.3.1 Mechanical Calibration 48
3.3.2 Structure-From-Motion 50
3.4 Dynamic Light Fields 53

4 Image Enhancement 55
4.1 Distortion Correction 57
4.1.1 Distortion Model 59
4.1.2 Camera Calibration 61
4.2 Color Normalization 68
4.3 Temporal Filtering 71
4.4 Filtering the Region of Interest 75
4.5 Image Geometry Transformations 75
4.6 Image Enhancement by Light Fields 77
4.6.1 Highlight Detection 79
4.6.2 Highlight Substitution 80
4.7 Summary 81

5 Light Fields of the Operation Area 83
5.1 Preprocessing 84
5.1.1 De-Interlacing 84
5.1.2 Distortion Correction 86
5.1.3 Cropping 87
5.2 Intrinsic Parameters 87
5.3 Extrinsic Parameters from Point Correspondences 90
5.4 Extrinsic Parameters from a Robot Arm 94
5.4.1 AESOP's Kinematics 95
5.4.2 Hand-Eye Calibration of AESOP 97
5.5 Extrinsic Parameters from an Optical Tracking System 101
5.5.1 Target Design 103
5.5.2 Hand-Eye Calibration of smARTtrack1 106
5.6 Computation of Depth and Confidence Maps 111
5.7 Light Field Visualization 117
5.8 Comparison 118

5.9 Summary 119

6 Augmented Reality: Registration and Fusion of 3-D Data with Light Fields 121

6.1 Important Anatomical Structures 122

6.2 Anatomical Database 125

6.3 Estimation of Registration Transformation Parameters 126

6.3.1 Selection of 3-D Points in the Endoscope Coordinate System 126

6.3.2 Selection of 3-D Points in the CT Data Coordinate System 127

6.3.3 Computation of Registration Transformation Parameters 129

6.4 Fusion by Multi-Modality Visualization 133

6.5 Summary 135

7 Experiments and Evaluation 137

7.1 Methodology for Evaluation 140

7.1.1 Real-Time Image Enhancement 140

7.1.2 Pose Error 142

7.1.3 Quality of Light Fields and Image Quality 145

7.2 Endoscopic Image Enhancement 147

7.2.1 Camera Calibration and Distortion Correction 149

7.2.2 Color Normalization 151

7.2.3 Temporal Filtering 152

7.2.4 Image Geometry Transformations 156

7.2.5 Evaluation 156

7.3 Light Field Reconstruction and Visualization 161

7.3.1 Static Light Fields Using Structure-From-Motion 161

7.3.2 Static Light Fields Using AESOP 163

7.3.3 Static Light Fields Using smARTtrack1 166

7.3.4 Dynamic Light Fields Using smARTtrack1 188

7.4 Image Enhancement by Light Fields 189

7.5 Augmented Reality: Registration and Fusion with 3-D Data 194

7.6 Discussion 198

8 Summary and Outlook 201

8.1 Summary 201

8.2 Outlook 204

A Homogeneous coordinates 207

B Singular Value Decomposition 209

C Mathematical Symbols 213

D German Title, Contents, Introduction, and Summary 223

D.1 Inhaltsverzeichnis 223
D.2 Einleitung 227
  D.2.1 Problembeschreibung und medizinische Relevanz 228
  D.2.2 Beitrag dieser Arbeit 230
  D.2.3 Bildmodalitäten 232
  D.2.4 Datenfusion 235
  D.2.5 Übersicht 237
D.3 Zusammenfassung 239

List of Figures 243

List of Tables 247

Bibliography 249

Index 273

Curriculum Vitae 277

Chapter 1

Introduction

The contentment of people depends mainly on their health. Therefore, societies spend a lot of money on research aimed at developing optimal techniques for disease treatment. The focus of this thesis lies on diseases where surgery is currently seen as the optimal treatment. Common examples for such diseases are inflammations of the gall bladder or the appendix, which result in strong pain. The removal of the affected anatomical structure is most often the only way to treat these patients. The goal of research in this area is to treat the patient as well as possible while decreasing the patient's trauma. Trauma can be defined as the amount of injury caused by the treatment, the intra- and post-operative pain, the cosmetic impairment, and the time of convalescence. The tendency in the field of surgery is to move towards so-called *minimally invasive operations* which traumatize the patient considerably less than conventional surgery.

The idea of minimally invasive surgery is to access the operation site through small "keyholes" which require only small incisions with a diameter of about 1 to 2 cm. Manipulation is performed by special surgical instruments, whereas the image of the operation site is obtained by an endoscope onto which a camera is mounted. Light is introduced through the endoscope, which provides, together with the camera, the image of the operation site displayed on a video monitor. Three terms are employed synonymously for this kind of operation: *minimally invasive surgery*, *keyhole surgery*, and *endoscopic surgery*.

Compared to conventional surgery, performing a minimally invasive operation involves training and dealing with a lot of drawbacks: unconventional instruments, no direct sense of touch or only through the surgical instruments, restricted freedom of movement, limited vision, image degradations caused by highlights, smoke, or small flying particles, and loss of stereoscopic depth perception due to displaying the endoscope's image on a video monitor. However, the reduced patient's trauma justifies such a human and technical effort. More and more minimally invasive operations replace the conventional operation as *gold standard*, i. e., the treatment that is currently seen as optimal. Examples are cholecystectomy (removal of the gall bladder), appendectomy (removal of the appendix), inguinal and diaphragmatic hernia, gastro-esophageal reflux disease (GERD), and bowel surgery (resection of the bowel in inflammatory or malignant diseases).

In this thesis techniques for supporting the surgeon during endoscopic surgery by computer vision methods are investigated. The following sections give a detailed description of the problems that arise during minimally invasive operations (Section 1.1) and the contribution of this work to reduce these problems (Section 1.2). The contributions are also related to conventional imaging technologies (Section 1.3) and to data fusion (Section 1.4). Finally, Section 1.5 outlines the structure of this thesis.

1.1 Problem Statement and Medical Importance

Compared to conventional surgery, many challenging problems arise during endoscopic surgery:

Image degradations: The endoscope's light fiber bundles emerge directly next to its distal lense. Hence, tissue surfaces perpendicular to the viewing direction show highlights, especially if the tissue is wet. The amount of light that can be introduced through the endoscope into visceral cavities is restricted. Too much light would cause too much heat and thus burn tissue which is close to the endoscope's tip. Under these conditions, visceral cavities, especially large ones like the abdomen, may only be illuminated inhomogeneously and with low contrast in some areas. Furthermore, close tissue surfaces may be over-exposed by the amount of light necessary to illuminate the rear part of the visceral cavity. Smoke, small flying particles, and a reddish coloring due to bleeding are the result of cutting tissue with high frequency diathermy or ultrasound dissectors. Finally, endoscope lenses have a very small focal length, e. g., 7 mm for a $1/2\,''$ CCD chip with PAL resolution (768×576 pixels), and it is particularly lenses with a small focal length that give rise to image distortions, especially at the border of the image, due to the manufacturing.

Limited vision: The problem of limited vision can be understood if one imagines the task of obtaining a clear impression of a room that can only be viewed through a camera. One would probably choose the smallest available focal length, walk into the middle of the room and try to look around. Imagine now the same task but with the camera mounted onto a large rod which has to be moved from the outside through the keyhole of the door.

If the focal length of the camera is fixed, which is common in the case of endoscopic surgery, a close examination of objects requires moving the camera close towards them. However, only a very small part of the room becomes then visible, and to navigate around the room becomes difficult.

Loss of stereoscopic depth perception: Human depth perception is mainly based on having and using two eyes. The depth information is then extracted from the eye's images according to the position of an object in the left and right image. The larger the difference of the two positions the nearer the object. Regarding projected images, i. e., photos, television, or computer and video monitors, this kind of depth perception is not possible as the image is displayed flatly. Other clues correlated to the depth of objects are then used: occlusion, illumination, and size information for still images, speed of object movement in relation to its size for moving images like in television. However, the impression is not the same as with *real* stereoscopic vision. During minimally invasive operations this difference becomes relevant. The simple task of grasping an object with an endo-grasper illustrates the difference: while this task is simple with normal vision, it becomes extremely difficult when the projected image on a video monitor has to be used.

Unconventional surgical instruments: All surgical instruments that are used differ from conventional ones, e. g. they are longer and smaller. Therefore, their use must be practiced.

Restricted freedom of movement: The main reduction of the patient's trauma is achieved by accessing the operation site through small "keyholes". This leads to a restriction of possible movements: each instrument and the endoscope has to be inserted and moved through such a "keyhole".

Limited sense of touch: During conventional surgery the surgeon is additionally able to examine the operation site by his sense of touch. The already described minimally invasive techniques do not allow this palpation. Only a very limited sense of touch is possible by using a surgical instrument: the elasticity of tissue can be examined by poking it with an instrument.

Difficult hand-eye coordination: The paradigm of hand-eye coordination describes the act of moving a hand (holding a surgical instrument) to a certain location. In the case of minimally invasive operations, this task becomes very difficult due to the following reasons: apart from the loss of stereoscopic depth perception, the viewing direction of the endoscope may not correspond to the surgeon's view, and the movement has to be performed with a long instrument through a “keyhole”. For instance, the tip of the instrument moves left when the surgeon moves the instrument to the right due to the manipulation through a keyhole.

With regard to the stated objective to develop optimal techniques for disease treatment, it is important to reduce the mentioned problems as far as possible. Thus, the conditions for the surgeon will improve, e. g., by improving the image quality or the site overview, which will lead to reduced stress. As a result, the performance of the operation will improve and the operation time can be reduced. Altogether this leads to a reduced patient’s trauma and recovery time.

In this thesis methods to remedy or reduce the problems of *image degradations, limited vision*, and *loss of stereoscopic depth perception* will be addressed. The following section elaborates the contributions.

1.2 Contribution of this Work

In order to reduce the problems during endoscopic surgery a novel system has been developed, which provides real-time image enhancement, 3-D visualization of the operation site, and augmented reality, i. e., registration and fusion with CT/MRI data. It allows removing or reducing several image degradations, reconstructing a 3-D model of the operation site (a so-called *light field*) which can be regarded in 3-D from arbitrary positions, and augmenting either the 2-D live image or the light field with CT/MRI data after performing a registration based on the reconstructed 3-D information. In the following, the contributions are described in more detail.

Image degradations Except for the work presented recently in [Fis04], only solutions for single image degradations have been published, e. g., see [Grö01, Hel01, Mün04], and most approaches were not developed for usage in the operating room. Therefore, first of all a system that allows processing and displaying endoscopic images is composed. The main component is a typical video-endoscopic system. For real-time image enhancement, the system is extended by a PC with a S-VHS frame grabber card and a second monitor. This setup allows grabbing the image from the endoscopic camera as well as processing and displaying it on the second monitor. Optimized algorithms enable real-time processing. Image distortions are corrected by calibrating the intrinsic camera parameters of the endoscope and feeding these parameters into a distortion correction algorithm. A color normalization method that was originally applied to improve object localization and classification, the *color cluster rotation* algorithm [Pau98], is employed to display illumination independent images in which different tissue types can also be separated in difficult situations. Small flying particles and smoke disturb the surgeon while cutting tissue. These degradations are reduced by temporal color median filtering. A method that allows using fast *spatial* median filters for *temporal* filtering is developed. When the pose, i. e., the position and orientation, of the endoscope is known, rotating the image according to a predefined horizon allows the horizon to be kept steady for almost arbitrary movements of the endoscope. Up to now, no evaluation of endoscopic image enhancement methods has been published. All methods developed here are therefore evaluated by surgeons.

Loss of stereoscopic depth perception and limited vision The proposed solution for both problems is the reconstruction of a light field [Lev96, Gor96] of the operation site. Light fields are a relatively new image-based method for modeling and visualizing 3-D scenes. Although other techniques have been used for 3-D reconstruction of the operation site [Tho02, Küb02, Dey02, Dev01], light fields have not yet been examined. The main problems during the reconstruction of light fields from endoscopic images are the determination of the endoscope's pose and the computation of 3-D scene geometry. Light field visualization would also be possible without knowledge of scene geometry, but with poor quality. Thus the light fields reconstructed in this thesis always contain 3-D scene geometry. Three different solutions for light field reconstruction are presented: based only on the input images, by using a robot arm that moves the endoscope and provides pose information, and by using an optical tracking system for endoscope pose determination. Each method has advantages and disadvantages. The visualization of the operation site using light fields permits the operation site to be viewed in 3-D, e. g., on a 3-D monitor or a head-mounted display (HMD). The viewing position is thereby not restricted to the original endoscope poses, e. g., if a part of a scene was captured by moving the endoscope very close to it, the overview can be increased by virtually decreasing the focal length and moving the endoscope backwards. This is especially helpful for coping with the problem of limited vision. Since the 3-D scene is represented in the computer, all movements are virtual and do not require moving the real endoscope. Finally, a method for substituting arbitrary image degradations with the help of light fields is developed. Three prerequisites are necessary to apply the technique: a light field of a scene has to exist, the degradation does not remain at the same position with respect to the scene while the endoscope moves, and the degradation can be detected in the image.

Augmented reality The information available in a light field, namely the 3-D scene geometry, allows providing 3-D augmented reality during an endoscopic surgery. For this purpose the light field is registered with other 3-D data like CT and MRI using anatomical landmarks which are identified in the scene. Then, CT/MRI data can be overlayed onto the light field visualization, allowing to "see" beyond the surface, through organs and tissue. Due to the scene geometry information, anatomical landmarks can be employed and markers are not necessary. Here the focus is on the computation of the necessary information to reconstruct a light field together with 3-D scene geometry and to register it with other 3-D data. Rendering techniques for the light field as well as for augmented reality, i. e., the light field overlayed with CT/MRI data, are not examined. Up to now, only augmented reality systems using markers for registration are known, e. g., [Sch03a] where the registered CT image is overlayed onto the 2-D monitor image. Naturally this kind of live 2-D AR is also possible after the light field was registered with 3-D data.

Now that the contributions of this work have been described, the next two sections relate them to conventional imaging technologies and to data fusion.

1.3 Image Modalities

Imaging technologies providing information about the inside of the patient are widely used in modern medicine. For instance, modern surgery is almost always preceded by some kind of image acquisition to obtain detailed information about the anatomy and the disease of the patient. Especially when an exact diagnosis of a disease is not possible, some kind of imaging technology might help to clarify the diagnosis. Apart from images that are obtained by *looking directly* into the patient using an endoscope, the discovery of X-rays has provided the possibility of generating images of human anatomy without injuring the patient or only with the potentially harmful radiation. Wilhelm Röntgen discovered X-rays in 1895. The first X-ray machines have been available shortly afterwards and were used all over the world for medical purposes. Nowadays, a lot of well established imaging technologies exist, e. g., computer tomography (CT), magnetic resonance imaging (MRI), functional MRI (fMRI), angiography, digital subtraction angiography (DSA), positron emission tomography (PET), single photon emission computer tomography (SPECT), or (3-D) ultrasound. The development of all these imaging technologies still continues. In the following paragraphs CT, MRI, and PET will be explained in more detail.

Computer tomography is based on X-rays. A conventional X-ray machine consists of an X-ray tube that emits a bundle of X-rays, and an X-ray film. The X-rays are propagated through the patient. Different types of tissue absorb different amounts of X-rays. The X-ray film is located opposite to the X-ray tube, the patient being in between. The "shadow" of the emitted X-rays is recorded by the film, i. e., the denser an anatomical structure, the more X-rays are attenuated and the brighter is the appearance on the film. Since X-rays are usually propagated through several tissue types, the resulting image becomes the "sum" of all those tissues. The spatial resolution of X-ray images is very high, details with a diameter of 0.1 mm can be distinguished. The idea of CT is to collect a large amount of data from each side of the patient by using X-rays. During the examination an X-ray tube rotates around the patient and emits a 1-D bundle of X-rays. Opposite the tube, a large number of electronic X-ray detectors detect the amount of radiation that was propagated through the patient. The tube and the detector perform a full circle around the patient and the detectors capture several thousand X-rays. The knowledge of X-ray physics and acquisition geometry allows reconstructing a 2-D image (slice) from the 1-D X-ray projections. A 3-D volume is obtained by acquiring several 2-D slices at different positions. It is now also possible to generate 2-D slices from any angle at any location (multi-planar reconstruction). Nowadays, the resolution of modern CTs is $0.2 \times 0.2 \times 0.4\,\mathrm{mm}^3$ $(x \times y \times z)$. CT and conventional X-ray images show density differences, i. e., morphological differences. These techniques are therefore particularly well suited for examining fractures, tumors, respiratory diseases such as tuberculosis, and other abnormalities that are accompanied by deviations in tissue density.

Magnetic resonance imaging employs a powerful magnetic field to obtain 2-D slice images of the patient. The spins of all atomic nuclei in the slice are aligned by the magnetic field. Radio frequency pulses perpendicular to the slice then cause some of the hydrogen nuclei to change their alignment. When the radio frequency is turned off, the hydrogen nuclei release radio frequency energy as they return to their original configuration. Detectors, in this case coils, wrapped around the patient, record these radio frequency signals. Again several 2-D slices provide 3-D anatomical information. The resolution of modern MRI scans is $0.8 \times 0.8 \times 0.8\,\mathrm{mm}^3$ $(x \times y \times z)$. In general, hydrogen nuclei are used in medicine but other nuclei could also be employed as it is done for instance in chemical research. The benefit of MRI scans is the possibility of discriminating anatomical structures and fluids with similar density but different amounts of hydrogen nuclei, e. g., fatty tissues with little water can be separated from blood vessels and other fluid-filled areas.

Positron emission tomography is a medical imaging technology where radioactive tracers are injected into the patient. A tracer consists of positron emitting radionucleids that are incorporated into normal body compounds such as glucose or water. The tracer is usually injected into the patient's blood circuit. The emitted positrons meet an electron after traveling one millimeter at the most. The reaction produces a pair of gamma ray photons in opposite directions. These gamma rays are recorded by a ring of detectors, and only simultaneous signals in opposite directions are further processed, the others being treated as noise. The resolution of PET scans is

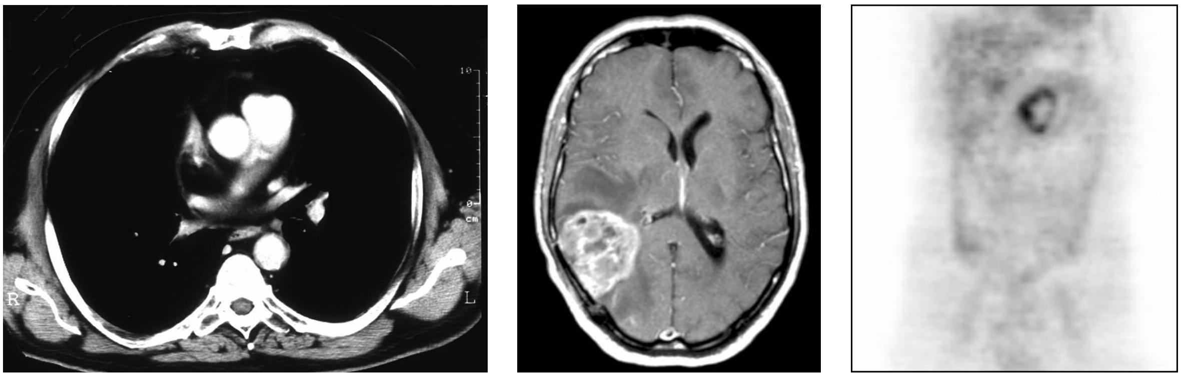


Figure 1.1: Examples of CT (left), MRI (middle), and PET image (right). The CT image shows a section through the thorax (lung), the MRI image shows a section through the head, and the PET image shows a "vertical" (transversal) section through the abdomen (Images by courtesy of the Department of Nuclear Medicine, University of Erlangen-Nuremberg).

very low: $4 \times 4 \times 6\,\mathrm{mm}$ ($x \times y \times z$). The benefit of PET scans is the possibility of studying biochemical processes, e. g., the activity of the brain or the absorption of glucose by tissue which can indicate a tumor. PET is therefore a functional imaging technology.

Figure 1.1 shows example images of the described image modalities. All three modalities depend on computers to reconstruct the 3-D data from lower-dimensional data that are collected electronically by some kind of detector.

The reconstruction of light fields can be regarded as a new 3-D imaging technology, where, similar to CT, 3-D information is reconstructed from lower-dimensional data. The common principle is the reconstruction of n-dimensional data from $n-1$ dimensional projections, assuming the projection parameters are known: on one hand the pose of X-ray tube and detectors as well as the equations describing the projection of X-rays through an object (X-ray attenuation law) allow the reconstruction of 2-D slice images from 1-D projections. The 3-D volume consists of consecutive 2-D slices. On the other hand the pose of the endoscope, the intrinsic camera parameters, and the equations describing the optical projection (pinhole camera model, perspective projection) allow the reconstruction of the 3-D scene from 2-D camera images.

1.4 Data Fusion

When regarding the benefits of CT, MRI, and PET, it becomes clear that no perfect imaging technology exists. Each type has advantages and disadvantages. CT scans allow distinguishing structures of the body with different density. Especially the anatomy of bony structures (high density) can be judged. MRI scans are well suited for soft tissues. A high contrast allows to

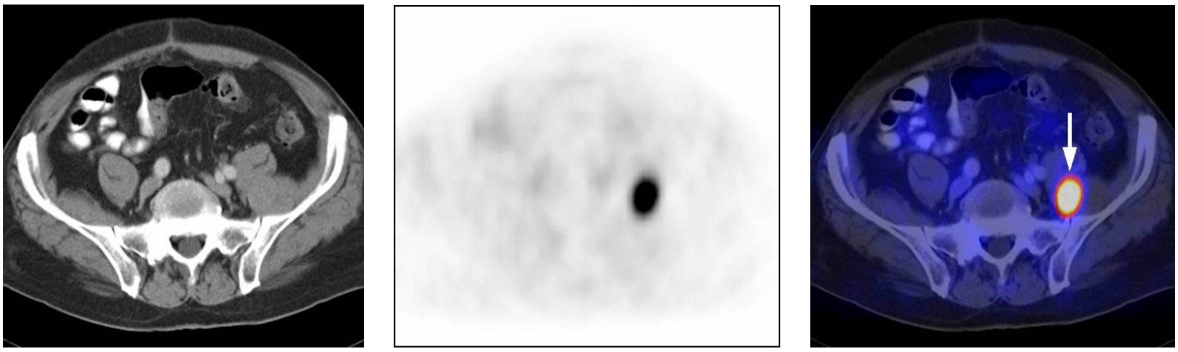


Figure 1.2: The advantages of fusing different image modalities can be seen: the tumor (a metastasis in the abdomen) which is not — or only for experts — visible in the CT image (left) is clearly visible in the PET scan as a black dot (middle). The fusion of both modalities (right) allows the exact localization of the tumor in the CT image (Images by courtesy of the Department of Nuclear Medicine, University of Erlangen-Nuremberg).

detect pathological abnormalities in blood vessels and organs, e. g., heart and prostate gland. Additionally, according to current medical knowledge, an MRI examination is harmless since non-ionizing radiation in the radio frequency range is employed. In contrast to CT and MRI, which both visualize anatomical structures, PET scans visualize biochemical processes which allow for the detection of abnormalities before changes are apparent in CT or MRI. PET is particularly suited to detect several types of tumors and metastasis, e. g., tumors in the liver, the lung, the breast, and the pancreas. The disadvantages of PET are the introduction of radioactive material into the patient and the low resolution of the retrieved images.

Data fusion is motivated by the wish to combine the advantages of different image modalities. Given two image modalities and a number of 2-D slices for each modality, the question arises which voxel of the first modality corresponds to which voxel of the second. More formally: a transformation which maps a coordinate system in the first modality into the second has to be determined. This process is called registration. A major problem for registration algorithms is the movement of the patient and his vital functions that deform tissue, such as breathing or heart beat. Especially soft tissue and deformable organs will in general not be located at exactly the same position after the patient has moved. However, if a transformation can be found, the two data sets can be fused, e. g., a transformed 2-D slice of a PET scan can be overlayed onto a corresponding slice of a CT scan, and it is possible to localize a tumor in the CT scan that is only, or more clearly, visible in the PET scan (see Figure 1.2). Of course, complete 3-D data sets can also be fused and displayed together. Then, one of the data sets is usually displayed semi-transparently. Volume rendering algorithms which exploit graphics hardware allow a fast visualization.

Similarly to the fusion of CT and PET, a light field can be fused with other available 3-D data like CT and MRI. The fused dataset can then be displayed to support the surgeon during the operation by augmented reality.

Summarizing the last two sections, light fields can be seen as a new kind of 3-D imaging technology which is generated directly in the operating room using 2-D endoscopic images. Furthermore, light fields can be fused with 3-D data available from conventional imaging technologies. Finally, the reconstruction of light fields may not only be performed during surgery but also for diagnostic purposes.

1.5 Outline

This thesis is structured as follows: Chapter 2 describes the state of the art in computer assisted endoscopic surgery; especially the development from conventional to minimally invasive surgery is pointed out. This chapter also summarizes the most recent developments in this area: robot assisted interventions, endoscopic image enhancement methods, and medical augmented reality systems.

The theory of light fields is introduced in Chapter 3. After defining light fields, known reconstruction and visualization techniques are summarized.

Chapter 4 develops solutions to reduce disturbing image degradations that occur during endoscopic surgery. A complete system for real-time endoscopic image enhancement is described. Additionally, a method for image enhancement based on light fields is presented.

Chapter 5 discusses three different ways of light field reconstruction. Pose determination systems are employed to reduce computation time and to increase robustness. Two pose determination systems are examined: a robot arm and an optical tracking system. Additionally, the reconstruction of light fields based only on the input video stream is described and compared to the other two methods.

In order to provide 3-D (and 2-D) augmented reality in the operating room, the light field is registered and fused with CT data. The developed methods are described in Chapter 6. First, important anatomical structures are identified, segmented, and examples are stored in a database. Then the registration parameters are estimated and the fused datasets are visualized.

Experiments and evaluations for the methods developed in Chapters 4 to 6 are shown in Chapter 7. The work is summarized and concluded with an outlook in Chapter 8.

Chapter 2

State of the Art in Computer Assisted Endoscopic Surgery

Surgical interventions were refined during the past centuries. After the “father of medicine”, Hippocrates (460-377 BC), the development of medicine went slowly until the High Middle Ages (1200-1400). From then, different factors led to a great growth and development, e. g., university degrees were required to practice medicine. Schools of surgery were founded and this was also the time of the first research about disinfection of wounds. Advances were made beginning with the Late Middle Ages. More serious injuries, e. g., due to the use of guns in battles and wars, lead to further advancement of the art of surgery. Furthermore, printing became available which resulted in a great increase in medical literature. In this period Leonardo da Vinci (1442-1519) and Michelangelo (1475-1564) published their well known anatomical descriptions of the human body. The art of surgery profited from the great progress in the anatomical knowledge. The introduction of narcosis in 1846 and local anesthesia introduced a new epoch for surgery. The discovery of X-rays in 1895 was a huge step: now it was possible to obtain information from inside the body without injuring the patient and to medicate according to this information.

The wish of physicians to obtain information from inside the body without injuring the patient is very old. Endoscopy was already described by Hippocrates, who referenced a rectal speculum [Reu98]. The Greek word “endo” means within, inside and internal; “scope” is derived from the Greek word “skopein”, (to) look at. Simple specula for gynecological endoscopy were found in the ruins of Pompeii, e. g., a three-bladed vaginal speculum [Reu98]. However, with the lack of a suitable light source, endoscopy was impractical. Sunlight, candlelight or an oil lamp were the only available light sources. In the 19-th century, kerosene, carbide, and gas lamps became available. Modern endoscopy had to solve three problems: the access to the inside of the body, the introduction of light into cavernous organs, and the transfer of the image to the eye. The main difficulty was the introduction of light while simultaneously transferring the image to the eye. The development of a light conductor called *Lichtleiter* by Philipp Bozzini (1773-1809) in 1806 and the invention of electric light by Edison in 1879 enabled further progress. Bozzini was the first who constructed an endoscope with an optical part that included a light conductor. The construction of endoscopes was refined during the 20-th century leading to the endoscope as it is known today.

Endoscopic surgery was already performed during the classical antiquity, e. g., inside the urethra. However, the surgeons operated blindly. The experiments of Bozzini in 1806 were the first "real" endoscopic interventions. The urethra was the operation site. Bladder tumors were firstly treated endoscopically in 1885, bladder stones in 1891. Since the experiments by Nitze in 1891, surgical instruments were introduced through separate ports. Commercial video-endoscopic systems have been available since 1961 (Siemens). In 1969 the first minimally invasive video-endoscopic surgery was performed: a thoracoscopic sympathectomy (severing of the sympathicus nerve). Since 1971 the camera could be mounted directly onto the optics of the endoscope. In 1985 Erich Mühe performed the first minimally invasive cholecystectomy (removal of the gall bladder) on a human being in Erlangen, Germany. Two years later, the first minimally invasive cholecystectomy *using a video-endoscopic system* was performed by Phillipe Mouret in France. The benefit of a minimally invasive operation becomes clear when it is compared to conventional surgery.

This chapter describes the state of the art in computer assisted endoscopic surgery. Sections 2.1 and 2.2 exemplarily depict the state of the art in conventional and minimally invasive surgery for a cholecystectomy. Sections 2.3 to 2.5 summarize the state of the art in *computer assisted* endoscopy: Section 2.3 refers to the use of robots, Section 2.4 describes computer vision algorithms used to enhance endoscopic images, and Section 2.5 shows recent developments in the area of augmented reality.

2.1 Conventional Surgery

This section describes a typical surgery by the example of an open cholecystectomy. Strong pain in the abdomen, either in the center beneath the breastbone or below the right ribs is the most frequent symptom indicating a problem with the gall bladder. The pain usually occurs following meals, especially fatty ones, and can last from minutes to a few hours. If the pain is caused by a gall bladder that is infected, inflamed, blocked, or filled with gallstones, the removal of the gall

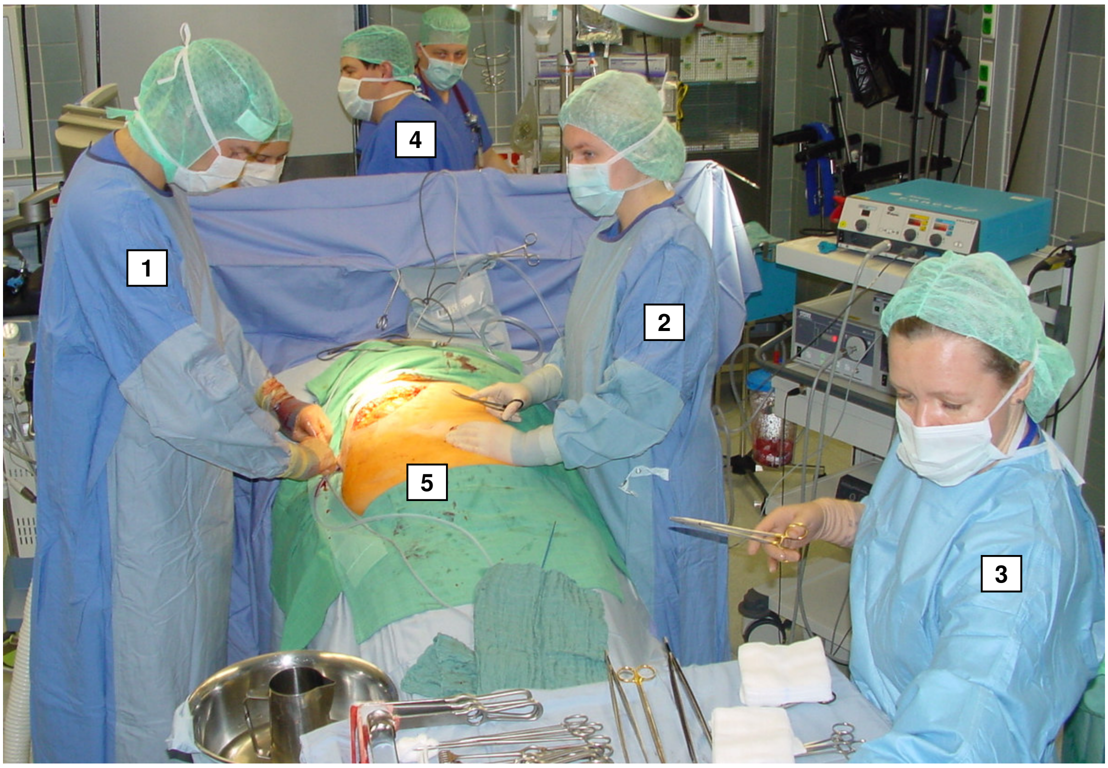


Figure 2.1: Setup for a conventional open cholecystectomy: the surgeon (1) performs the operation together with the assisting surgeon (2), the theater nurse (3) assists the surgeons, and the anesthesiologist (4) is responsible for the patient's (5) narcosis.

bladder is often the choice of treatment.

The standard open operation is carried out under general anesthesia. It starts with an incision of several centimeters just below the rib on the right side of the abdomen. The gall bladder is located below the liver which has to be moved to expose the gall bladder. Now the dissection begins. The vessels and tubes, namely cystic artery and cystic duct, to and from the gall bladder are identified, ligated, and cut. The gall bladder is removed. The incision is closed.

Figure 2.1 shows the setup in the operating room for a conventional open cholecystectomy. Four people are required to carry out the operation: the surgeon and the assisting surgeon, who perform the surgery, the theater nurse, and finally the anesthesiologist who is responsible for the patient's narcosis.

The next section describes the same operation using minimally invasive techniques. However, it should be noted that minimally invasive surgery is not always possible. In rare cases, for instance, when the gall bladder is extremely inflamed, infected or has very large gallstones, conventional surgery is recommended.

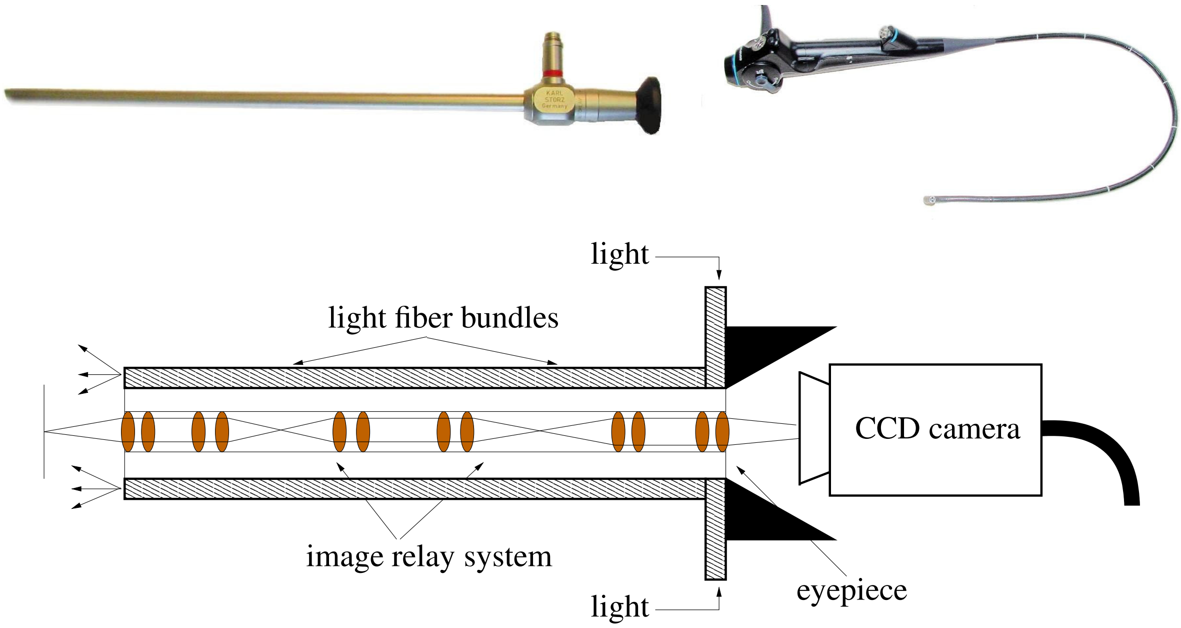


Figure 2.2: Endoscopes are either rigid (top left) or flexible (top right) tubular instruments, equipped with an illumination system and an optical image relay system. The image relay system of rigid endoscopes consists of several lenses and an eyepiece (bottom). The image of the operation site is either relayed to the eye or a Charge-Coupled Device (CCD) camera [Bop99].

2.2 Minimally Invasive Surgery

In a medical context the term *endoscopy* denotes the illumination and inspection of visceral cavities and cavernous organs with the help of an endoscope [Psc98]. It is either motivated by a diagnostic purpose with the possibility of obtaining a tissue probe for further histological examination, or by a surgical purpose to provide the image of the operation site during minimally invasive operations. An endoscope is a tubular instrument, equipped with an illumination system and an optical image relay system [Bop99]. Rigid and flexible fiber-optic devices are available. The rigid design permits the use of glass lenses and rods which results in better light transmission and image quality compared to fiber-optic-based designs. Here, rigid endoscopes are employed. Figure 2.2 displays rigid and flexible endoscopes and shows a detailed sketch of a rigid endoscope.

If endoscopy is used to provide the image of an operation site, the endoscope itself as well as the required surgical instruments can be introduced into the body through small ports, so-called *trocars*, that require merely small incisions: usually about 1 to 2 cm, depending on the diameter of endoscope and instruments. Due to the reduced trauma for the patient this kind of surgery is called *minimally invasive*. Depending on the intended purpose, endoscopes of different length and diameter are employed with specific names referring to the location where they are used. Some examples are laparoscope (abdomen, rigid endoscope), thoracoscope (thoracic cavity, rigid endoscope), arthroscope (cavities of joints, e. g., knee, rigid endoscope), gastroscope (stomach, flexible endoscope), colonoscope (large intestine, flexible endoscope) and bronchoscope (respiratory organs, flexible endoscope).

Approximately 30% of all interventions at the Department of Surgery, University of Erlangen-Nuremberg, are performed minimally invasive. Almost 100% of all cholecystectomies are performed minimally invasive, which is currently the standard for this operation [Sop92]. In the following paragraphs the workflow during a minimally invasive operation for laparoscopic cholecystectomy is exemplarily described.

Figure 2.3 shows the typical situation in the operating room. At least four persons are generally required: the surgeon who manipulates with surgical instruments, the assisting surgeon who moves the endoscope according to the needs of the surgeon, the theater nurse, and finally the anesthesiologist who is responsible for the patient's narcosis. The following steps are performed sequentially (see also [Coo92]):

1. A so-called mini-laparotomy is performed through an incision below the belly button (subumbilical) to introduce the first optic trocar (10 mm in diameter) into the abdomen. Through this port CO_2 gas is insufflated up to 15 mm mercury pressure to create a pneumoperitoneum. The gas pressure leads to the required space in the abdominal cavity. After warming up to body temperature, the laparoscope is introduced through the trocar for a first round view to inspect the abdomen. The endoscope's images of the abdomen are displayed on a video monitor.

2. Under direct vision from the abdominal cavity, three other trocars are inserted: in the upper abdomen (mid abdominal line, trocar diameter of 10 mm), the right upper abdomen (mid clavicular line, trocar diameter of 5 mm), and the right side (axillary line, trocar diameter of 5 mm). Through these trocars laparoscopic instruments, e. g., scissors, coagulators, graspers, etc. are introduced.

3. After insertion of the instruments the cholecystectomy begins with the dissection of the gall bladder from the liver bed. An endothermic coagulator is used for dissection and stopping of bleedings. The dissection procedure is the same as in open cholecystectomy: the vessels and tubes, namely cystic artery and cystic duct, to and from the gall bladder are

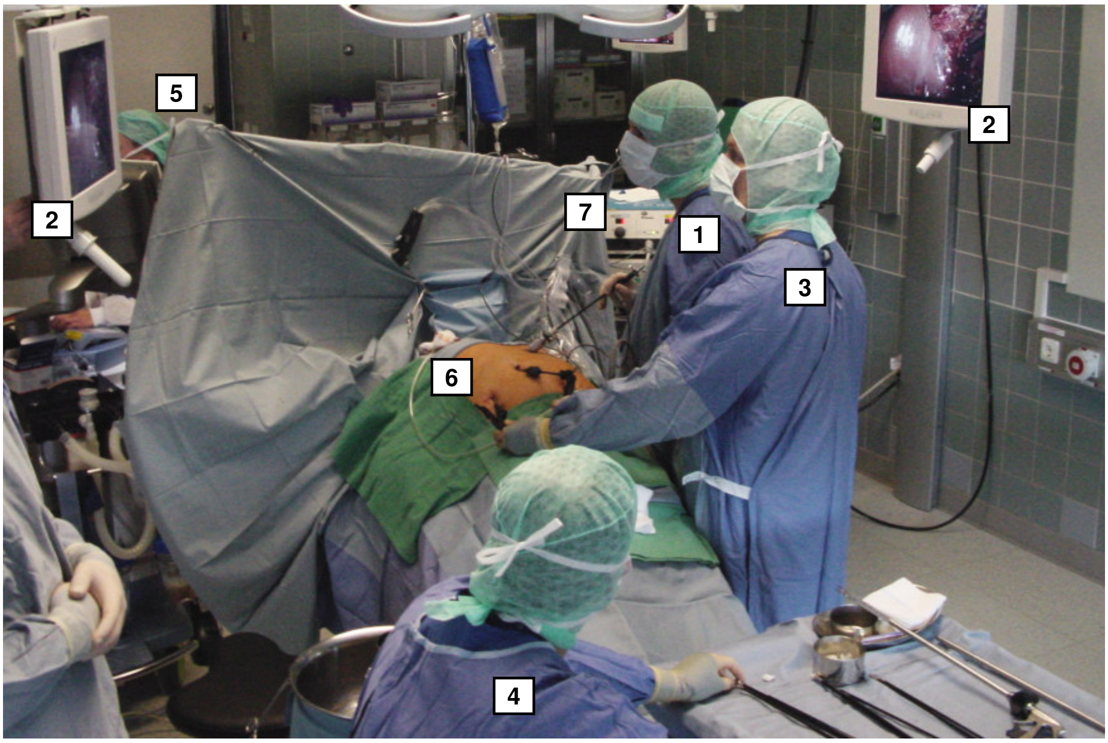


Figure 2.3: Setup for a laparoscopic cholecystectomy in a modern operating room: the surgeon (1) manipulates with surgical instruments and looks at the image displayed on a video-monitor (2), the assisting surgeon (3) moves the endoscope according to the needs of the surgeon, the theater nurse (4) assists the surgeon, and the anesthesiologist (5, scarcely visible) is responsible for the patient's (6) narcosis. The video-endoscopic system (7, scarcely visible) includes a rack, an endoscopic camera, a light source, a carbon dioxide insufflator and one or more video monitors (2) displaying the image of the endoscope.

identified, ligated by titanium clips, and cut with endoscissors. In case of bleeding, blood is removed by a suction-irrigation device to avoid an imbibition of all tissues with blood which would have a permanent reddish coloring effect. The use of a thermic coagulator burns the tissue which leads to smoke and small flying particles in the abdominal cavity. In order to remove these degradations an exchange of the inflated gas is necessary.

4. The gall bladder is removed through one of the larger incisions, usually the mini-laparotomy port, after removing the optic trocar. A loss of gall stones in case of an organ rupture can be avoided by using a retrieval bag.

5. Finally, the CO_2 gas is sucked off and the incisions are closed with sutures and dermal glue.

Robots	**Humans**
$\oplus$ Good spatial accuracy	$\ominus$ Limited spatial accuracy
$\oplus$ High speed of action*	$\ominus$ Limited speed, especially for high accuracy tasks*
$\oplus$ Untiring, stable	$\ominus$ Prone to tremor and fatigue
$\oplus$ Results are reproducable*	$\ominus$ Results vary from human to human*
$\oplus$ Can be designed for a wide range of scales	$\ominus$ Limited dexterity outside human scale
$\oplus$ May be sterilized	$\ominus$ Limited sterility
$\oplus$ Resistant to radiation and infection	$\ominus$ Susceptible to radiation and infection
$\ominus$ Limited hand-eye coordination	$\oplus$ Strong hand-eye coordination
$\ominus$ Limited dexterity	$\oplus$ Dexterous (at human scale)
$\ominus$ Have to be programmed*	$\oplus$ Flexible and adaptable
$\ominus$ Limited to relatively simple procedures	$\oplus$ Can integrate extensive and diverse information
$\ominus$ Use only quantitative information	$\oplus$ Able to use qualitative information
Are expensive but may reduce personnel costs*	-

Table 2.1: Comparison of robot and human characteristics which are important for surgery (adapted from [How99], own extensions labeled with *).

2.3 Robot Assisted Interventions

During the last two decades more and more robot systems have been employed in operative medicine. Master-slave manipulators are distinguished from robots. Robots are defined as "*automatically controlled multitask manipulators, which are freely programmable in three or more axes*" [Fed01]. In contrast to this, master-slave manipulators are in general robot arms which perform simple actions according to human input. The terms *manipulator* and *robot arm* are used synonymously.

The reasons for using robots in medicine are the same as for using them in industrial applications: their spatial precision, their lack of fatigue, and their speed of action are advantageous compared to human beings. Additionally they may reduce personnel costs. Table 2.1 compares human and robot characteristics showing that robotic systems extend the human capabilities but are limited to specific tasks.

Orthopedic operations where bones have to be shaped for prothesis to fit in are perfectly suited for robots: they can mill, drill, and saw with a precision less than $0.1\,\mathrm{mm}$ [Fed01]. Therefore, hip replacement surgery was one of the ground-breaking applications for robotic systems in medicine [How99]. The surgeon plans the task for the robot based on pre-operative 3-D CT data. The CT data are registered to the patient by placement of markers (so-called *fiducials*) in suitable bony anatomical structures. The markers are visible in the CT scan of the patient and thus allow registration between patient and CT data. The robot is placed in relation to the patient at a fixed known location. The markers are used to register the robot with the patient. Finally, the robot shapes the hip bone with a high-speed milling device based on CT data. Two representatives of orthopedic robotic systems are ROBODOC [Bar98, Hon03] and CASPAR [Pet00]. In [Bar98] an improved prosthetic fit, i. e., the fit between the bone and the replacement joint, and a reduced overall complication rate was shown for operations with CASPAR. However, results of more recent clinical trials challenge the benefit of orthopedic robotic systems [Maz04].

Neurosurgery is another area where high precision robots support the surgeons. In the beginning, stereotactic frames were attached to the patient's skull for registration. The relationship between the frame and pre-operative CT data was used to guide instruments within the brain. Nowadays less invasive markers or video images can be used for registration, and optical tracking systems enable high accuracy navigation of hand-held instruments [Kon98, Hol01, Lié01]. Furthermore, several robotic systems have been developed to enhance stability, accuracy, and ease of use [How99], e. g., Minerva [Gla95] and Neuromate [Fed01]. Robots similar to CASPAR also have been used in neurosurgery [Fed01]: the hexapod robot "Evolution 1" is employed to move the endoscope and instruments through the nose into the head cavity [Zim02, Nim04].

The following paragraphs describe master-slave manipulators for endoscopic surgery in greater detail. Programmable robots that perform tasks automatically are not yet used. Commercially available systems differ in the input device and the number of manipulators, i. e., the number of robot arms, that are employed. Simple systems consist of only one robot arm which generally moves the endoscope. More complex systems include several robot arms allowing surgical instruments to be guided according to "joystick" movements of the surgeon, where quivering of the human hand can be removed electronically [Fed01]. These systems also allow telesurgery, i. e., operations where the surgeon is far away from the patient. Two simple and two complex systems are described.

The Automated Endoscope System for Optimal Positioning (AESOP) [Met98, Jac97, Bac97] is the most frequently sold manipulator [Fed01]. AESOP is a robot arm that provides voice-controlled movement of the endoscope. In comparison to foot control, which is commonly used in operating rooms, e. g., to enable cutting tissue with the high frequency diathermy, voice control of AESOP was found to be more accurate [All98] with the advantage of not requiring the surgeon to look away from the video monitor. AESOP's robot arm has seven degrees of freedom. The arm and the endoscope plug are magnetically connected. This provides mechanical disconnection in case of mechanical obstruction. Before each operation the lower safety limit for robot arm movements has to be set manually, i. e., below this limit the robot would touch the patient. The controller of the robot arm then prevents such movements. The surgeon has to wear a head-set with a microphone to provide a high quality voice signal (cf. Figure 2.4). Initially, the classifier for the voice commands is trained for each surgeon and the results are then stored on a voice card that has to be inserted into the controller. The movements are restricted according to the available commands [Bal02]. AESOP provides commands for incremental movements, e. g., "up" or "left", as well as commands for continuous movements, e. g., "move down" or "move right". Commands for continuous movements have to be terminated with "stop". Three endoscope positions can be stored, e. g., with the command "save one" the current position is stored, and the robot returns to this position when the corresponding command "return one" is given. This feature simplifies and accelerates moving the endoscope. AESOP positions the endoscope without quivering and more accurately than a surgical assistant. The Department of Surgery of the University of Erlangen-Nuremberg uses the AESOP model 3000 for minimally invasive thoracoscopic and laparoscopic surgery. Figure 2.4 shows the setup in one of Erlangen's operating rooms for a minimally invasive operation using AESOP 3000. Compared to the conventional procedure where four people are required to carry out the operation (cf. Figure 2.3), three are now sufficient for minimally invasive procedures: AESOP substitutes the surgical assistant who normally moves the endoscope. Setting up AESOP in the operating room requires only about three minutes.

Another manipulator system for the controlled movement of an endoscope is EndoAssist [Aio02]. The movement is not directed by voice but is derived from the movement of the surgeon's head by a special sensor. The sensor has to be activated by a foot switch. This increases safety since the surgeon himself decides which movements should be interpreted as commands and which should be ignored. Consequently, misinterpretations by unwanted head movements are avoided which allows surgical tasks to be performed faster compared to AESOP [Neb03].

Two complex master-slave systems are currently used in minimally invasive surgery [Nio01, Fed01]: ZEUS and DaVinci [Int05b]. With these systems, the surgeon operates not directly at the operating table but several meters away, sitting at a control console. From there, the physician controls the robot arms with devices similar to joysticks. The robot arms allow the use of different types of surgical instruments. The new environment requires an enormous training effort for the physician. ZEUS consists of three interactive robotic arms based on AESOP technology. One arm positions the endoscope to provide the view of the operation site, while the other two arms are reserved for manipulation of surgical instruments under the surgeon's control. The en-

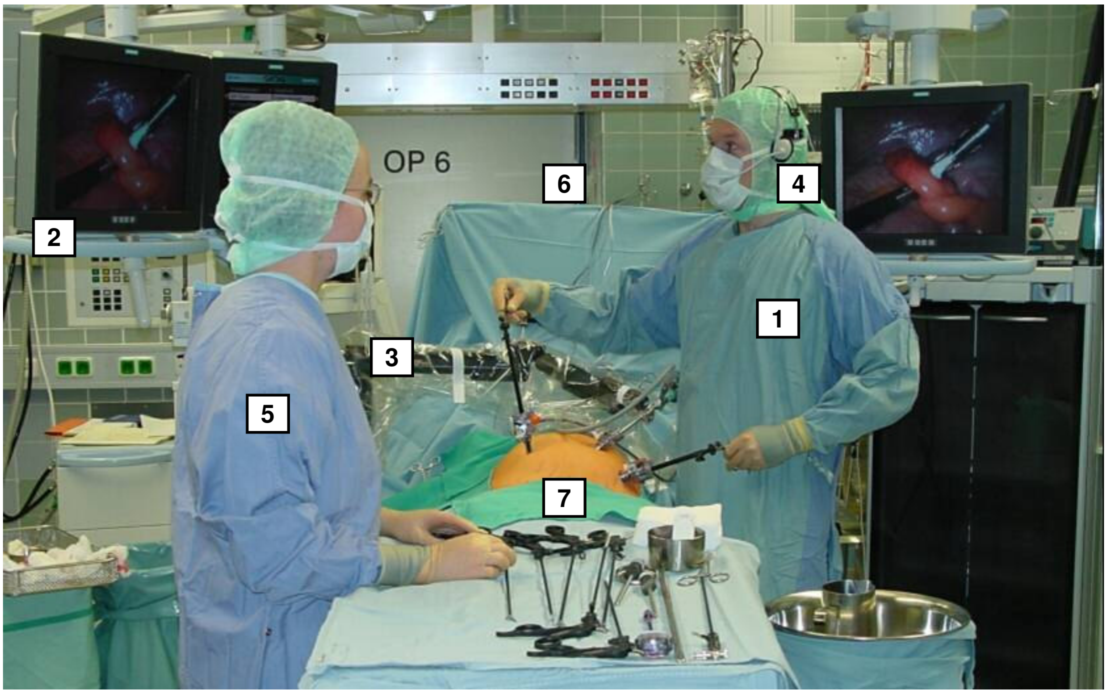


Figure 2.4: Setup for a minimally invasive operation at the Department of Surgery of the University of Erlangen-Nuremberg using the robot arm AESOP 3000: the surgeon (1) manipulates with two surgical instruments and looks at the image displayed on a video-monitor (2), the robot arm AESOP 3000 (3) moves the endoscope according to the voice commands of the surgeon, who wears a headset (4), the theater nurse (5) assists the surgeon, and the anesthesiologist (located behind the blanket at number 6, not visible here) is responsible for the patient's (7) narcosis.

doscope is moved by spoken commands. The surgeon seated at the console can choose between 2-D and 3-D view of the operation site [Pra02]. Quivering of the hand movements of the operating surgeon is removed electronically and large movements of the hand are translated into small movements of the instruments. The control devices resemble conventional surgical instruments. For some years the surgeons have been able to operate minimally invasive on the beating heart, e. g., perform bypass operations with the DaVinci system [Sel00]. The movements of the surgeon's hand can be scaled arbitrarily, so that the surgical instruments can also be moved with high accuracy at micro-surgical operations [Get02]. DaVinci also provides quivering removal. The operation site is viewed ten times magnified and in 3-D.

2.4 Image Enhancement

Minimally invasive surgery is carried out by the surgeon viewing the image of the operation site displayed on a video monitor (cf. Figures 2.3 and 2.4). The advantages for the patient due to the

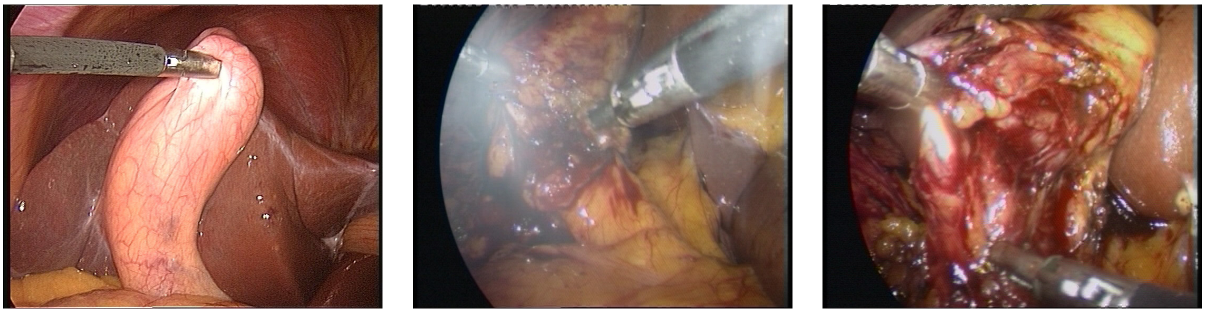


Figure 2.5: Examples of degradations in endoscopic images. The left image shows the result of image distortion: the surgical instrument is bent instead of straight. In the middle image smoke hampers vision. Highlights are visible in all three images, but particularly in the right one. Bleeding leads to an imbibition of the tissues with blood leading to a reddish coloring, which can be seen in the image on the right. The three images were captured during laparoscopic cholecystectomies.

minimally invasive technique are made possible using special equipment: endoscopes, cameras, video monitors, and several surgical instruments such as endoscissors, endograspers, high frequency diathermy for cutting tissue, etc. However, this equipment leads to disadvantages for the surgeon. The surgeon's already difficult task due to the lack of sense of touch (only through the surgical instruments), restricted freedom of movement, limited vision, and loss of stereoscopic depth perception as a result of the displaying of the endoscope's images on a video monitor, is complicated by degradations in the displayed image. Figure 2.5 depicts three images of laparoscopic cholecystectomies that illustrate some of the occurring image degradations. Table 2.2 summarizes all common degradations, their cause, and already published approaches to remove or reduce these degradations.

Two types of endoscopic image enhancement can be distinguished: real-time image enhancement (pre-processing) and offline image enhancement (post-processing). The objective of the first type is to help facilitate the difficult conditions during minimally invasive operations. The goal of the latter one is to improve the quality of captured and stored images, e. g., to improve the automatic classification of tissue into benign or malign, respectively [Mün03, Mün04]. Algorithms for real-time image enhancement may only require 40 msec of computation time since 25 images per second are displayed. The available time for offline image processing depends on the task but usually several seconds or even minutes are acceptable.

Although the number of minimally invasive operations is growing, the companies providing the endoscopy equipment do not yet offer systems for real-time image enhancement. White balancing, which is done once at the beginning of an operation, and the common methods to adapt the image of a video monitor (contrast, brightness, color, aperture) are the only possibil-

Degradation	**Cause**	**Solutions**
Highlights	The light fiber bundles are located directly beneath the distal lens of the endoscope (cf. Figure 2.2). Tissue surfaces perpendicular to the viewing direction show highlights, especially when the tissue is wet.	Highlight removal [Bor02, Grö01, Sch00, Pal99] or highlight detection with subsequent coloring of highlight regions [Fis04, Vog02b, Vog01a, Stö00, Gev00].
Over-exposure	The amount of light required to illuminate the rear parts of visceral cavities can lead to an over-exposure of near tissue surfaces.	Highlight removal methods which are not based on separating diffuse from specular reflection can be used, e. g., [Bor02, Grö01, Vog02b, Vog01a]
Distortions	Optical lenses with small focal lengths are used to enlarge the visible area and gain clarity; inaccuracies during the manufacturing process of the optics occur. The resulting image distortion, e. g., straight lines get bent, increases towards the borders of the image.	Distortion correction by modeling lens distortion and determining the distortion parameters using a calibration pattern [Vog03a, Zha02, Sal02, Vog01a, Hel01, Zha00, Tsa87].
Color errors, reddish coloring	A bad white balancing leads to unnatural image colors. Bleeding leads to an imbibition of all tissue with blood leading to a reddish coloring.	Color correction for static causes like bad white balancing [Fis04, Mün03, Mün04] or color normalization [Vog03a, Vog01a, Pau98].
Inhomogeneous illumination, low contrast	The available light often is not sufficient for illuminating the operation site optimally	Highpass filtering in the frequency domain, histogram equalization [Fis04].
Smoke and small flying particles	Tissue is cut with high frequency diathermy.	Temporal filtering [Vog03a, Vog01a, Vog01b].

Table 2.2: Common degradations found in endoscopic images, causes, and proposed methods to reduce or remove the degradations. Apart from the own publications on real-time endoscopic image enhancement [Vog03a, Vog02b, Vog01a, Vog01b], the only other real-time image enhancement method is described in [Hel01]. The objective of real-time processing was not yet reached in [Fis04]. The special area of endoscopic image processing is addressed in [Fis04, Mün04, Mün03, Vog03a, Vog02b, Zha02, Vog01a, Vog01b, Hel01, Grö01, Pal99]. All other methods were developed for general image processing.

ities to change the appearance of the image. Most industrial research seems to be done in the field of camera and monitor development, but new techniques like progressive scan cameras, which capture the whole image at once and not in interlaced mode, are not (yet) offered. Simple real-time processing of endoscopic images was already performed in 1996 when a real-time system for moving the endoscope automatically was presented [Arb96]: The endoscope positioning robot AESOP moves the endoscope according to a tracked color marker at the tip of a surgical instrument. A comparison between robotic and human camera control was presented in [Omo99]. The frequency of camera correction and lens cleaning is reduced significantly by using robotic camera control. Additionally, the subjective impression of the surgeons was that the robot performs better than a human assistant. It is understandable that not many publications about real-time image enhancement exist. Only since the beginning of the new millennium have affordable computers become fast enough for more complex real-time image processing. RAVE, a system for real-time autonomous video enhancement, was presented in 2002 [Abl02]. The system was designed to enhance low-quality or corrupted streaming video data. The focus was on detection and removal of artefacts like blurring, snow noise, brightness flicker, and ghosting. The goal of real-time processing was not yet reached. An approach for real-time distortion correction in endoscopic images was published in 2001 [Hel01]. In the same year the methods for real-time endoscopic image enhancement developed in this thesis were published for the first time [Vog01a, Vog01b], including distortion correction, color normalization, and temporal filtering. The complete system was then presented in 2003 together with a subjective evaluation of the image enhancement methods by physicians [Vog03a]. One year later, a system for real-time endoscopic image enhancement was described in [Fis04], but according to the authors the goal of real-time processing was not yet reached. In a pre-operative calibration phase information about already present degradations is collected and computations for later corrections are performed. Look-up tables are used to store the results. During the runtime phase these tables are used to perform fast transformations of image colors. The proposed system detects highlights, compensates inhomogeneous illumination and low contrast, and removes color errors. Unfortunately no details about the applied algorithms and computation times are provided.

In the area of offline endoscopic image processing/enhancement methods a larger number of publications exists. Usually no computation times are provided. For diagnostic purposes the goal is the classification of endoscopic images, e. g., to detect tumors in colonoscopic images [Kar03, Mar03a, Wan01] or to detect cancer in images of lung tissue [Gal99]. Classification results of esophagus images can be enhanced significantly by applying color shading correction and color calibration [Mün03, Mün04]. An algorithm that automatically rotates the endoscopic image and keeps the horizon steady is described in [Kop01, Kop04]. It is based on the computation of camera motion by the 8-point algorithm [LH81, Har97]. The required 2-D point correspondences are established by tracking points from image to image. Highlight detection and removal was addressed in [Pal99]. Based on the di-chromatic reflectance model for di-electric inhomogeneous material [Sha85], the specular and diffuse part of image regions is computed. Suppressing the specular part removes highlights. Two alternative approaches are presented in [Grö01]. Highlights in endoscopic images of a beating heart are eliminated by linear interpolation with the help of gradient information or with an iterative filling-in approach employing anisotropic diffusion.

Conventional image processing methods may be applied for endoscopic image enhancement. *Distortion correction* is achieved by computing distortion parameters with camera calibration algorithms. An overview over such algorithms together with an accuracy evaluation is given in [Sal02]. Tsai's algorithm [Tsa87] is widely used. Zhang's algorithm [Zha00] models more parameters (two for radial and two for tangential distortion) and requires at least two images of a calibration pattern in contrast to [Tsa87] where one image is sufficient. All camera calibration algorithms are based on the pinhole camera model which will be described in Section 3.1.1.

If *highlight regions* are defined as missing data, these regions can be filled by the algorithm presented in [Bor02]. The approach is inspired by texture synthesis techniques. The missing data are iteratively filled from the borders to the middle. For each pixel a neighborhood is defined and the color value of other pixels with similar neighborhood are used to fill the missing data. Several publications address the topic of *highlight detection* in color images [Gev00, Sch00, Stö00]. However, these algorithms have to be combined with a substitution method, e. g., the method mentioned above [Bor02]. A method for highlight detection and substitution based on a *light field* was presented in [Vog02b].

Illumination correction is most commonly applied to improve the results of feature tracking [Grä03, Fus99, Jin01], object localization [Pau98], and object recognition [Fin98].

Apart from the system developed here, no convenient system for real-time endoscopic image enhancement is currently available, especially none that can be used during minimally invasive operations, neither commercially nor as a research project. In fact only one out of all publications describes an algorithm running in real-time: Helferty's distortion correction approach [Hel01]. The methods proposed here are the only ones that were evaluated by physicians [Krü04, Krü03a, Vog03a]. For all other methods an evaluation of the benefit by physicians is lacking completely.

2.5 Augmented Reality

Virtual reality (VR) denotes completely computer generated environments and is the computer vision/graphics counterpart to our physical environment. It is also possible to mix both kinds of reality which is then denoted as *mixed reality* [Mil99]. Probably the most important part of mixed reality is *augmented reality (AR)*: real scenes are augmented by computer generated, i. e., virtual, objects. As an example imagine an architect who wishes to view his virtual design in the real environment. Another example is AR in medicine: computer generated medical information like MRI and CT data would be helpful for surgeons if it were available in the sense of augmented reality, e. g., if the vision of the surgeon would be augmented by the location of a tumor that is clearly visible in the CT data but not in the real scene.

In the following sections the components of typical medical AR systems like the ones described in [Sch01a, Vog04c, Vog04d] are summarized. A head-mounted display (HMD) is generally used for *3-D visualization* of the real scene (Section 2.5.1). A *pose determination system* (Section 2.5.2) provides the viewer's pose and allows displaying the corresponding view of the virtual scene or object. In advance, virtual reality (Section 2.5.3) and reality are *registered* to each other so that the correct views of the virtual scene can be fused and displayed together with the real scene (Section 2.5.4).

2.5.1 3-D Visualization

Humans observe their 3-D environment with two eyes that allow obtaining stereoscopic depth information. In contrast to this, computer images are usually displayed on a flat monitor. It is impossible to obtain stereoscopic depth information from such images. If views of AR/VR scenes are projected onto a flat monitor, other depth cues like occlusion and speed of movement in relation to the observer's viewpoint provide information about depth. However, the 3-D impression is not realistic. Two solutions are currently available: 3-D monitors and head-mounted displays (HMDs). The idea of both is to provide separate images for the left and the right eye and thus simulating "normal" stereoscopic depth perception. HMDs display the corresponding left and right images by two small displays directly in front of the eyes. These displays are mounted onto the head-device used by the observer, e. g., a special helmet. Using 3-D monitors, both images are projected onto the monitor and separated afterwards: either the monitor itself separates the images (autostereoscopic display), e. g., by using prisms that provide two images according to the viewing direction of the observer's eyes [Dod95, Dod00, See05], or the viewer wears special glasses which provide the correct image for each eye. State of the art in this area is the simultaneous projection of the left and right image with polarized light onto a screen and using glasses with corresponding filters [Ind05]. I-Max 3-D cinemas are also based on this technique.

Both, AR and VR applications use HMDs or 3-D monitors for realistic stereoscopic 3-D visualization. HMDs as well as 3-D monitors require an already available 3-D scene which allows obtaining stereo images. VR applications simply render two images of the virtual scene with known stereo parameters: distance of viewpoints (baseline) and convergence. The visualization of augmented reality depends on the employed technique. On the one hand optical see-through HMDs allow normal perception of a scene through a semi-transparent display on which virtual images can be displayed, e. g., see [Aue99, Sal01]. This type of HMD has the disadvantage that a displacement of the helmet leads to a displacement of the virtual scene. On the other hand video-see-through HMDs acquire the required images of the scene by two digital cameras, fuse these images with the VR, and display them afterwards. Usually a stereo camera system mounted on the helmet is used, e. g., as in [Vog04c, Vog04d], but other designs are also possible [Fuc98]. Stereo camera systems are also suitable for visualization on a 3-D monitor.

In the case of endoscopic surgery it is very difficult to provide stereo images of the operation site. A possibility is to use stereo endoscopes which are also called 3-D endoscopes. Two lens systems integrated into the optical cylinder of the 3-D endoscope provide the stereo image. The surgeon wears an HMD. Although HMDs became lighter compared to the first prototypes, wearing an HMD is not convenient for many surgeons. An HMD supports the surgeon with a realistic 3-D impression but complicates other tasks like changing instruments or communication with colleagues, and wearing an HMD for several hours is not very pleasant. Probably 3-D monitors will be used in future but the quality of the 3-D impression is currently not comparable to HMDs. A system design with a multi camera endoscope and a 3-D monitor was described in [Dod95]: The manufacturing of an endoscope with six lenses was identified as main challenge and still no such endoscope has been manufactured. Moreover, the smallest diameter of available stereo endoscopes is 10 mm. Another disadvantage of 3-D endoscopes is the reduced image quality due to the smaller lens systems compared to monocular endoscopes with the same diameter. Companies like Karl Storz [Sto05] offer 3-D endoscopes for normal endoscopic surgery, but the demand is not very high. 3-D endoscopes are mainly used in minimally invasive robotic master-slave systems like DaVinci and ZEUS.

The disadvantages of stereo endoscopes motivate research with monocular endoscopes. Realistic 3-D visualization with stereo images in real-time by using monocular endoscopes is not possible. After acquiring a sequence of monocular images, a 3-D model of the scene has to be generated which then allows the rendering of stereo images by texture mapping or image based rendering algorithms. The simplest 3-D model consists of the scene geometry (depth) obtained from two images together with texture information (acquired images). A more sophisticated 3-D model integrates the information from several images. Special endoscopes were developed to simplify the computation of scene geometry. In [Hay01] a laser-pointing endoscope together with a conventional endoscope allows the triangulation of the projected laser spot. Another laser device was presented in [Mül02]: a ring of laser light is used to measure the 3-D geometry during tracheoscopies. A structured light approach was presented in [Fuc98]. A method which exploits the known shape of surgical instruments was published in [Cab04]. The instruments function as calibration patterns. Depth information is thereby obtainable for the area in the image where the instruments are located. There are also solutions for geometry reconstruction based solely on the captured images. The basic idea is the application of *structure-from-motion* algorithms to endoscopic images (cf. Section 3.3.2), in [Tho02] to a sequence acquired during a coloscopy and in [Kop04] to a sequence acquired during a laparoscopy. Kübler et al. [Küb02] applied their structure-from-motion approach to a simulated sequence of a coloscopy. The idea of improving these approaches by computing the pose of the endoscope with the help of a pose determination system (cf. Section 2.5.2) is obvious but was not yet investigated. The computation of depth information may be simplified by using stereo endoscopes despite their drawbacks [Mou01, Cor00, Hof02]: after calibrating the endoscope, stereo matching algorithms can be applied leading to a dense depth map.

Due to the problematic nature of generating a 3-D model from monocular endoscopic images, many minimally invasive AR approaches are not providing a realistic stereoscopic 3-D visualization. Their objective is to augment the 2-D image displayed on the monitor [Olb05, Feu05, Tra04, Sch01a, Hel01, Wes99]. Therefore, no 3-D model of the scene has to be generated. A pose determination system and the registration and fusion of the tracked endoscope's image with the virtual data is sufficient. The images of the VR are projected according to the known viewing position and overlayed onto the real image that is displayed on the monitor.

2.5.2 Pose Determination

Three technical possibilities exist for determining the pose of an endoscope in the operating room: electro-magnetic tracking systems, optical tracking systems, and electro-mechanic positioning systems. An electro-magnetic tracking system is based on a sender and a receiver component. The sender generates pulsed electro-magnetic fields, e. g., with 100 Hz, and the pose of the coil of the receiver can be determined by the use of electro-magnetic induction physics [Sch03a]. Electro-magnetic tracking systems have been used for pose determination of rigid and flexible endoscopes [Sch01c, Sch01a, Sch03a, Wes99, Ell03]. Two representative systems are MINIBIRD by Ascension Technology [Asc05] and AURORA by Northern Digital [Nor05].

A typical optical tracking system consists of two or more cameras and a so-called *target* that is tracked. The target is built out of markers that can easily be identified in the images captured by the cameras. For instance, spheres with a retro-reflective surface (passive tracking) or light-emitting diodes (LEDs) are employed (active tracking). Infrared light may be used to simplify marker identification. The 3-D position of each visible marker is calculated by the tracking system. The knowledge of the geometry of the target then allows calculating its pose. Optical tracking systems are utilized for non-endoscopic medical AR approaches [Vog04c, Vog04d, Aue99, Fuc98, Hol01, Lié01, Kor04] as well as for endoscope tracking [Vog05b, Olb05, Feu05, Tra04, Dey00, Dey02, Fuc98, Kon98, Sch98, DB01]. Two representative systems are POLARIS by Northern Digital [Nor05] and smARTtrack1 by Advanced Realtime Tracking [Adv05]. Both use infrared light.

Electro-mechanic tracking systems usually consist of a mechanical arm which is built out of several joints. The orientation of each joint is measured by potentiometers. Passive systems exist [Mar03b], but active systems, i. e., robot arms like AESOP, are more widespread. The state of the art of robot arms was described in Section 2.3.

The advantage of optical tracking systems lies in their high accuracy. Northern Digital specifies the root mean square (RMS) error of the position determined by their POLARIS system as $\leq 0.35\,\mathrm{mm}$ which was also shown in [Sal01], together with an RMS orientation error of $\leq 1\,^\circ$. Advanced Realtime Tracking specifies the RMS error of their smARTtrack1 system as $\leq 0.2\,\mathrm{mm}$ and $\leq 0.12\,^\circ$. Optical tracking systems provide a higher accuracy in comparison to magnetic tracking systems [Aue99, Sal01, Sch03a]. The disadvantage of optical tracking systems is the required visibility of the target. For applications where this cannot be guaranteed, electro-magnetic tracking systems are better suited. However, in that case several drawbacks have to be accepted: The accuracy is lower with an RMS error of $\leq 1.8\,\mathrm{mm}$ and $\leq 1.7\,^\circ$ [Nor05, Sch03a], the operation range is smaller and metallic objects influence the measurement and further reduce the accuracy [Hum02].

High accuracy electro-mechanic tracking systems such as CASPAR exist; however, systems like AESOP were not designed to provide high pose accuracy, but to allow voice-controlled endoscope positioning (cf. Section 2.3). The accuracy of AESOP is specified by Computer Motion Inc. with an RMS error of $1.5\,\mathrm{mm}$, no orientation error is provided. In general the accuracy of modern electro-mechanic and optical tracking systems is comparable [Mar03b]. The disadvantage of electro-mechanic tracking systems is their limited freedom of motion and the reduced accuracy for specially designed endoscope positioning systems like AESOP.

A fourth possibility of pose determination would have been acoustical tracking systems but this technique is not common, especially not for tracking endoscopes, and is therefore not elaborated on in more detail.

2.5.3 Virtual Data

In general, virtual data for medical AR systems are obtained from CT or MRI scans of the patient. The quality of 3-D ultrasound is not yet good enough in comparison with CT/MRI, but it has already been used for the intra-operative registration with CT/MRI data [Wu03]. Since CT/MRI data contain a lot of information that is not relevant for the specific task, either a precedent segmentation of the relevant data is performed or rendering using transfer functions that show only the interesting parts is employed. The advantage of a precedent segmentation is that geometric models can be built out of the segmented data which can then be used for fast rendering. The simplest geometric model consists of a set of 3-D points. If the interior of the object is irrelevant, the surface can be extracted from the segmented data and for instance be represented as a triangular mesh (see Chapter 4.2.2-7 in [Gir00]). The disadvantage of both models is the large number of parameters that have to be stored. A reduction of the parameters may be possible, e. g., reducing the number of vertices of the triangular mesh [Cam99], or the object could be approximated by simple geometric forms like ellipses or cylinders.

A very exact anatomical model of the human body is the “VOXEL-MAN” [Höh00]. It is based on images of 770 cryotom slices and corresponding CT images. Large anatomical structures are modeled as a set of colored 3-D points, very small structures like small vessels are modeled as polygons that are fitted to the anatomy.

The modeling of anatomical structures like stomach, intestines, and vessels for endoscopic training systems is described in [Küh00]. The proposed method also allows for the modeling of deformations. An object is represented by a set of control points. Each control point has a weight and is connected to other control points. The computation of a deformation requires the solving of a differential equation of second order. A similar approach for modeling vessels was presented in [Abd98]. Curves and surface patches that connect 3-D points model the surface of the object.

2.5.4 Registration and Fusion

The simultaneous visualization of reality and VR, e. g., the display of a *fused* image in an HMD, requires a *registration* between the two domains. In general a coordinate system is assigned to each domain. The task of registration is to determine a transformation between the two coordinate systems. For AR applications the viewing position in the real world has to be mapped to the viewing position in the virtual world. Based on the registration transformation, the virtual and real data can be fused and visualized.

An overview over registration methods is given in [Mai98]. Registration techniques are usually classified according to a number of attributes, e. g., see [Mai98, Haj01]. A main discrimination criterion is the type of transformation: namely rigid or non-rigid. Other interesting criteria are: intrinsic (intensity based) vs. extrinsic (marker based), intra-modal vs. inter-modal, and the dimension of the domains (single images or image sequences, 2-D/2-D, 2-D/3-D, 3-D/3-D). Only rigid registration techniques will be regarded in the following.

Extrinsic 3-D/3-D registration was used in [Feu05, Tra04, Sch01a] to register the endoscope with CT data by placing markers onto the skin of the patient. In [Vog04c, Vog04d] a calibration pattern provides the necessary information to register an HMD with the pose of an instrument. Intrinsic registration techniques based on mutual information are summarized in [Plu03].

The registration of two datasets is mostly performed in two steps: a coarse registration followed by a fine registration. The coarse registration is thereby usually performed manually or manual-interactively [Hub03]. For instance, three corresponding points, which can be selected by hand, are sufficient for a rigid 3-D/3-D registration. Based on the transformation estimated by the coarse registration, an iterative-closest-point (ICP) algorithm [Bes92, Che92, Rus01] is normally applied for fine registration. Very robust variations of the ICP algorithm lead to good results even with a bad coarse registration [Sha99]. A fully automatic registration approach is presented in [Hub03]: *spin-images* [Joh97] provide the coarse registration which is refined by an ICP algorithm. The spin images are employed to compute the necessary point correspondences automatically, where local statistical features or geometric features like curvature [Yam02] are used.

Once the registration transformation is known, the two domains have to be visualized together. In the case of AR systems the virtual information is overlayed onto the reality visualization, either in 3-D with an HMD [Vog04c, Vog04d, Wen03], or in 2-D projected onto a monitor [Feu05, Tra04, DB01, Sch01a]. For a realistic visualization correct occlusions of the virtual data have to be computed. Note that this is not possible without having a 3-D model of the real scene. For pure VR systems which register two 3-D datasets like MRI and CT, more complex solutions have to be found for a suitable visualization, e. g., see [Has99].

2.5.5 Comparison of Medical AR and 3-D Visualization Approaches

Table 2.3 summarizes the properties of some medical AR and 3-D visualization approaches. In addition the properties of the approach presented in this thesis are listed.

In [Cor00] only a concept of a system is presented, the system was never implemented. This publication is therefore ignored in the following. All minimally invasive AR approaches utilize

Approach	**Goal**	**Appl.**	**Pose**	**Reality vis.**	**Reg.**	**End.**	**Model**
S. Vogt [Vog04d]	AR	ISV	o	3-D	e	-	-
T. Thormählen [Tho02]	3-D rv	MIS	-	3-D m	-	m	3-D
C. Kübler [Küb02]	3-D rv	MIS	-	3-D m	-	m	3-D
D. Dey [Dey02]	3-D rv	MIS	o	3-D m	e	m	3-D
W. Konen [Kon98]	3-D r	MIS	o	2-D	i	m	3-D
J. Cortadellas [Cor00]	AR	MIS	em	3-D	i	s	3-D
F. Devernay [Dev01]	AR	MIS	r	3-D m	e	s	3-D
M. Scheuering [Sch03a]	AR	MIS	em	2-D	e	m	-
B. Olbrich [Olb05] and S. De Buck [DB01]	AR	MIS	o	2-D	e	m	-
M. Feuerstein [Feu05] and J. Traub [Tra04]	AR/P	MIS	o	2-D	e	m	-
F. Vogt, this thesis and [Vog04a, Vog05b]	AR/3-D rv	MIS	o/r/-	2-D/3-D m	i	m	3-D

Table 2.3: Comparison of medical AR and 3-D visualization approaches: **Goal:** augmented reality (AR), surgery planning (P), 3-D reconstruction and visualization (3-D rv), 3-D reconstruction (3-D r); **Application (Appl.):** minimally invasive surgery (MIS) or in-situ visualization (ISV); **Pose determination system (Pose):** optical (o), electro-magnetic (em), robot arm (r), none (-); **Reality visualization (Reality vis.):** 3-D live (3-D), 3-D scene model (3-D m) or 2-D live (2-D); **Registration (Reg.):** extrinsic, with some kind of markers (e), intrinsic, without markers (i), none (-); **Endoscope type (End.):** monocular (m), stereo (s), or none (-); **Model of the operation site (Model):** 3-D or no model (-).

some kind of pose determination system and markers for (extrinsic) registration [Olb05, Feu05, Tra04, Dev01, Sch03a, DB01]. The 2-D live image is augmented when monocular endoscopes are employed [Olb05, Feu05, Tra04, DB01, Sch03a]. In [Dev01] a 3-D model is computed by using a stereo endoscope, which also allows providing stereoscopic 3-D perception of the augmented reality. Optical tracking systems and extrinsic registration are also employed in other medical AR systems, e. g., for in-situ visualization [Vog04d]. Several techniques for the reconstruction of a 3-D model from monocular endoscopic images exist: by utilizing an optical tracking system [Dey02, Kon98] or only based on the image sequence [Tho02, Küb02].

Apart from the methods developed here [Vog03b, Vog04b, Vog04a, Vog05b], up to now *light fields* have not been employed for image based 3-D modeling and augmented reality in endoscopic surgery, and no minimally invasive AR approach employs *intrinsic registration*.

Chapter 3

Light Field Theory

This chapter introduces the concept of light fields (Section 3.1), light field visualization techniques (Section 3.2), and light field reconstruction (Section 3.3). It further describes the relatively new concept of dynamic light fields (Section 3.4).

A detailed description of light field reconstruction and visualization can be found in [Hei04].

3.1 Definition and Concept

As mentioned in the introduction (Section 1.1) one of the main goals in computer assisted endoscopic surgery is to support the surgeon with a 3-D visualization of the operation site. In this thesis light fields are used for modeling and visualizing 3-D scenes. An alternative way would be geometry-based modeling and visualization: the scene is modeled by geometric primitives composed of different materials and a set of lights [Lev96]. Based on the model an image of the scene is generated. In contrast to this, modeling with light fields is an image-based method: even with unknown scene geometry, new photo-realistic views of the scene can be generated based solely on pre-acquired images.

In 1996 light fields have been introduced into computer vision and graphics [Gor96, Lev96]. In general they describe a set of samples of the plenoptic function [Ade91], which identifies everything that can potentially be seen within a scene:

$$\psi(\theta, \phi, \lambda, \tau, \boldsymbol{p}) = I \, . \tag{3.1}$$

The function value I measures the intensity for wavelength λ at point $\boldsymbol{p} = (x, y, z)^{\mathrm{T}}$ in direction $\boldsymbol{n}$ specified by the two angles θ and ϕ at any point in time τ (cf. Figure 3.1). In accordance with

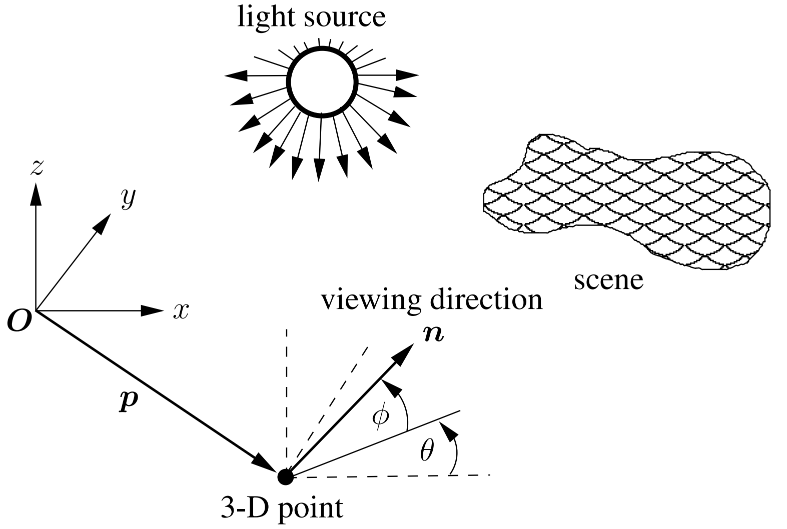


Figure 3.1: The plenoptic function $\psi(\theta, \phi, \lambda, \tau, \boldsymbol{p})$ measures the intensity for wavelength λ at point $\boldsymbol{p}$, in direction $\boldsymbol{n}$ specified by the two angles θ and ϕ, at any point in time τ.

[Hei04] $\boldsymbol{n}$ is defined as follows:

$$\boldsymbol{n} = \begin{pmatrix} \cos(\theta)\cos(\phi) \\ \sin(\theta)\cos(\phi) \\ \sin(\phi) \end{pmatrix} . \tag{3.2}$$

For the acquisition of light fields with digital cameras the complexity of the 7-dimensional function ψ has to be reduced [Hei04]. One simplification and two assumptions are made:

1. **Simplification of ψ:** The plenoptic function ψ can be regarded as a function that measures the spectral energy distribution, which is a function over λ, at each 6-tuple $(\theta, \phi, \tau, \boldsymbol{p})$. This real-valued function is represented by a discrete 3-tuple $(I_\mathrm{r}, I_\mathrm{g}, I_\mathrm{b})$, according to the three color channels of digital cameras, *red*, *green*, and *blue*, which result from the usual sampling of the energy distribution by the spectral sensitivity curves of the camera's CCD chip. Thus, the simplified plenoptic function is defined as:

$$\boldsymbol{\psi_6}(\theta, \phi, \tau, \boldsymbol{p}) = \begin{pmatrix} I_\mathrm{r} \\ I_\mathrm{g} \\ I_\mathrm{b} \end{pmatrix} . \tag{3.3}$$

In the following the dimension of plenoptic functions is written as an index, e. g., $\boldsymbol{\psi_6}$ denotes the 6-dimensional plenoptic function.

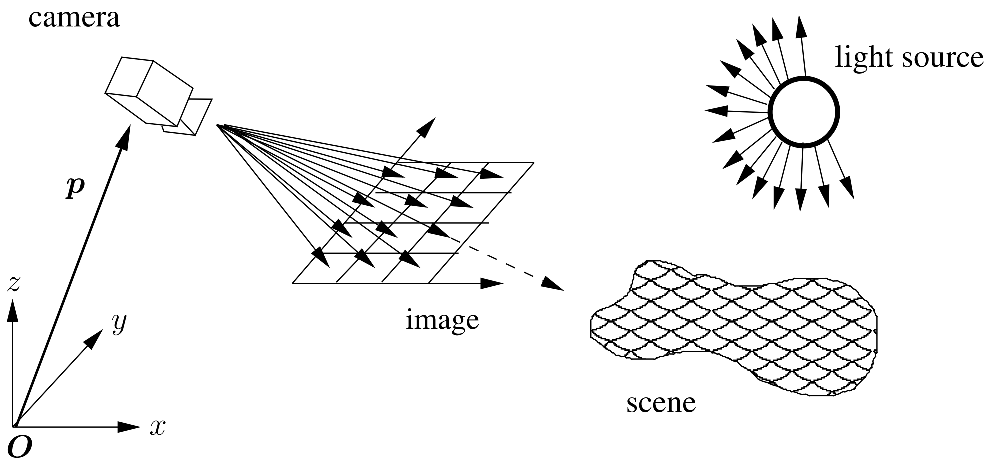


Figure 3.2: A sample of the plenoptic function captured by a camera. A whole bundle of directions (light rays) is recorded simultaneously. Each pixel corresponds to one light ray.

2. **Static scene assumption:** The scene is assumed to be static, i. e., the only object that moves is the camera and all observed objects are static. Given this assumption the sampling of the plenoptic function is independent of the point in time τ, i. e.,

$$\boldsymbol{\psi_5}(\theta, \phi, \boldsymbol{p}) = \boldsymbol{\psi_6}(\theta, \phi, \tau, \boldsymbol{p}), \forall\, \tau\,. \tag{3.4}$$

3. **Transparent medium assumption**: a transparent medium (air) is assumed to fill the space between the camera and the scene. Furthermore, degradations like fog or smoke may not occur. Formally, the transparent medium assumption can be expressed as

$$\boldsymbol{\psi_5}(\theta, \phi, \boldsymbol{p}) = \boldsymbol{\psi_5}(\theta, \phi, \boldsymbol{p} + s \cdot \boldsymbol{n}), \forall s \in \mathbb{R}\,. \tag{3.5}$$

   This means the value of $\boldsymbol{\psi_5}(\theta, \phi, \boldsymbol{p})$ along the light ray $\boldsymbol{p} + s \cdot \boldsymbol{n}$ does not change. The 5-D parameter vector $(\theta, \phi, \boldsymbol{p})$ then has only 4 degrees of freedom. A 4-D parameterization of $\boldsymbol{\psi_5}(\theta, \phi, \boldsymbol{p})$ is presented in Section 3.2.1

If images are captured by a digital camera, a whole bundle of directions (light rays) is recorded simultaneously at a 3-D point for each image (see Figure 3.2). Each pixel corresponds to one light ray. By taking several images from different points in space in different directions, a more or less dense sampling of the plenoptic function is obtained.

A *light field* consists of all captured images of a scene together with the projection parameters. The information contained in a light field allows the computation of the plenoptic function

$\psi_5(\theta, \phi, \boldsymbol{p})$ for the bundle of light rays corresponding to the pixels of each captured image.

3.1.1 Samples of the Plenoptic Function

This section explains the computation of samples of the plenoptic function given a light field. A pinhole camera model is assumed as physical model of the perspective projection of a world point to pixel coordinates. This is a very common assumption made in computer vision, e. g., see [Tsa87, Tru98, Zha00]. The extrinsic camera parameters describe the pose of the camera in world coordinates by a rotation matrix $\boldsymbol{R} = [\boldsymbol{r}_{\mathrm{x}}, \boldsymbol{r}_{\mathrm{y}}, \boldsymbol{r}_{\mathrm{z}}] \in \mathbb{R}^{3\times 3}$ with $\boldsymbol{r}_{\mathrm{x}}, \boldsymbol{r}_{\mathrm{y}}, \boldsymbol{r}_{\mathrm{z}} \in \mathbb{R}^3$ and a translation vector $\boldsymbol{t} \in \mathbb{R}^3$. Without loss of generality the coordinate system of the camera is defined as follows:

- The origin $\boldsymbol{t}$ coincides with the camera's projection center.
- The x-axis $\boldsymbol{r}_{\mathrm{x}}$ is parallel to the horizontal axis of the image plane and points to its right side.
- The y-axis $\boldsymbol{r}_{\mathrm{y}}$ is parallel to the vertical axis of the image plane and points to its bottom.
- The z-axis $\boldsymbol{r}_{\mathrm{z}}$ is chosen as the cross product of $\boldsymbol{r}_{\mathrm{x}}$ and $\boldsymbol{r}_{\mathrm{y}}$ to obtain a right-handed coordinate system. The vector $\boldsymbol{r}_{\mathrm{z}}$ points to the viewing direction of the camera.
- The image plane is parallel to the plane spanned by $\boldsymbol{r}_{\mathrm{x}}$ and $\boldsymbol{r}_{\mathrm{y}}$.

The intrinsic camera parameters $(F_{\mathrm{x}}, F_{\mathrm{y}}, C_{\mathrm{x}}, C_{\mathrm{y}})$ describe the intrinsic projection properties of the camera. F_{x} and F_{y} are the *effective focal lengths* in x- and y-direction, given in pixels. The focal length F in mm is obtained by multiplying F_{x} by the pixel size dx [mm/pixel] on the sensor chip in x-direction ($F = F_{\mathrm{x}} \cdot dx$). $(C_{\mathrm{x}}, C_{\mathrm{y}})^{\mathrm{T}}$ is the intersection of the optical axis (viewing direction $\boldsymbol{r}_{\mathrm{z}}$) with the sensor chip of the camera, called *principal point* (in pixels). As in most real acquisition systems the *image skew* is assumed to be zero, i. e., the angle between $\boldsymbol{r}_{\mathrm{x}}$ and $\boldsymbol{r}_{\mathrm{y}}$ is 90°.

If all camera parameters are known, they generally describe the transformation from a 3-D world point $\boldsymbol{w}$ to camera coordinates ${}^{\mathrm{c}}\boldsymbol{w}$ and the projection from camera coordinates to pixel coordinates $\boldsymbol{q}$:

$$ {}^{\mathrm{c}}\boldsymbol{w} = \begin{pmatrix} {}^{\mathrm{c}}w_{\mathrm{x}} \\ {}^{\mathrm{c}}w_{\mathrm{y}} \\ {}^{\mathrm{c}}w_{\mathrm{z}} \end{pmatrix} = \boldsymbol{R}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}) = \begin{pmatrix} \boldsymbol{r}_{\mathrm{x}}{}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}) \\ \boldsymbol{r}_{\mathrm{y}}{}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}) \\ \boldsymbol{r}_{\mathrm{z}}{}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}) \end{pmatrix} , \tag{3.6} $$

$$\boldsymbol{q} = \begin{pmatrix} q_{\mathrm{x}} \\ q_{\mathrm{y}} \end{pmatrix} = \frac{1}{{}^{\mathrm{c}}w_{\mathrm{z}}} \begin{pmatrix} F_{\mathrm{x}}\, {}^{\mathrm{c}}w_{\mathrm{x}} \\ F_{\mathrm{y}}\, {}^{\mathrm{c}}w_{\mathrm{y}} \end{pmatrix} + \begin{pmatrix} C_{\mathrm{x}} \\ C_{\mathrm{y}} \end{pmatrix} . \tag{3.7}$$

Introducing the so-called *calibration matrix*

$$\boldsymbol{K} = \begin{pmatrix} F_{\mathrm{x}} & 0 & C_{\mathrm{x}} \\ 0 & F_{\mathrm{y}} & C_{\mathrm{y}} \\ 0 & 0 & 1 \end{pmatrix} , \tag{3.8}$$

and using homogeneous coordinates, equations (3.6) and (3.7) can be written as:

$$\underline{\boldsymbol{q}} \sim \boldsymbol{K}\boldsymbol{R}^{\mathrm{T}} \left(\boldsymbol{w} - \boldsymbol{t}\right) \tag{3.9}$$

or expressed as a single matrix multiplication:

$$\underline{\boldsymbol{q}} \sim \underbrace{\boldsymbol{K}\left[\boldsymbol{R}^{\mathrm{T}}, -\boldsymbol{R}^{\mathrm{T}}\boldsymbol{t}\right]}_{=:\boldsymbol{P}} \underline{\boldsymbol{w}} = \boldsymbol{P}\underline{\boldsymbol{w}} . \tag{3.10}$$

The homogeneous vector $\underline{\boldsymbol{q}}$ corresponds to the Euclidean vector $\boldsymbol{q}$ (see Appendix A). The sign "$\sim$" means the equality of homogeneous vectors up to an unknown scalar. The *projection matrix* $\boldsymbol{P}$ contains the extrinsic $(\boldsymbol{R}, \boldsymbol{t})$ and intrinsic $(\boldsymbol{K})$ camera parameters. The knowledge of $\boldsymbol{P}$, which can be decomposed into $\boldsymbol{K}$, $\boldsymbol{R}$, and $\boldsymbol{t}$, allows computing the parameters of the plenoptic function $\boldsymbol{\psi}_{\mathbf{5}}(\theta, \phi, \boldsymbol{p})$ [Hei04]:

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{t} && (3.11) \\ \boldsymbol{n} &= \begin{pmatrix} n_{\mathrm{x}} \\ n_{\mathrm{y}} \\ n_{\mathrm{z}} \end{pmatrix} = \boldsymbol{R}\boldsymbol{K}^{-1}\underline{\boldsymbol{q}} && (3.12) \\ \theta &= \mathrm{sarctan}(n_{\mathrm{y}}, n_{\mathrm{x}}) && (3.13) \\ \phi &= \mathrm{sarctan}\left(n_{\mathrm{z}}, \sqrt{n_{\mathrm{x}}^2 + n_{\mathrm{y}}^2}\right) , && (3.14) \end{aligned}$$

with sarctan(x, y) defined as

$$\mathrm{sarctan}(x, y) = \begin{cases} \arctan\left(\frac{x}{y}\right) & , \quad y > 0 \\ \pi + \arctan\left(\frac{x}{y}\right) & , \quad y < 0 \\ \frac{\pi}{2} \cdot \mathrm{sign}(x) & , \quad y = 0 \end{cases} . \tag{3.15}$$

Equation (3.12) can be derived from equation (3.9) by first solving for $\boldsymbol{w}$, leading to the back-projection of an image point $\boldsymbol{q}$ to all possible locations of the corresponding 3-D point $\boldsymbol{w}$:

$$\boldsymbol{w} \sim \boldsymbol{R}\boldsymbol{K}^{-1}\underline{\boldsymbol{q}} + \boldsymbol{t} \tag{3.16}$$

For $\boldsymbol{t} = (0,0,0)^{\mathrm{T}}$, $\boldsymbol{w}$ points to the direction of the light ray passing through $\boldsymbol{q}$. Since the only interest is the direction, the unknown scalar can be fixed to 1, which finally leads to equation (3.12). Equations (3.13) and (3.14) can be verified by using equation (3.2).

If the measured color value of $\boldsymbol{q}$ is given as $\boldsymbol{f}(\boldsymbol{q}) = (I_{\mathrm{r}}(\boldsymbol{q}), I_{\mathrm{g}}(\boldsymbol{q}), I_{\mathrm{b}}(\boldsymbol{q}))^{\mathrm{T}}$, the equations (3.11) to (3.14) provide the parameters to store a sample of the plenoptic function:

$$\boldsymbol{\psi}_5(\theta, \phi, \boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{q}) \,. \tag{3.17}$$

3.1.2 Depth Maps and Confidence Maps

A light field can be extended by additional information. A *depth map* and a *confidence map* may be available for each captured image. The depth map $d(\boldsymbol{q})$ stores the distance to the surface of the scene/object for each pixel $\boldsymbol{q}$. Either the *range*

$$d_t(\boldsymbol{q}) = \|\boldsymbol{w} - \boldsymbol{t}\| \,, \tag{3.18}$$

i. e., the Euclidean distance of the 3-D point $\boldsymbol{w}$ to the camera center, or the depth

$$d_z(\boldsymbol{q}) = \boldsymbol{r}_{\mathrm{z}}{}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}) \,, \tag{3.19}$$

i. e., the length of the projection of the 3-D point $\boldsymbol{w}$ onto the viewing direction $\boldsymbol{r}_{\mathrm{z}}$ of the camera, is stored. For a known pixel $\boldsymbol{q}$, $d_t(\boldsymbol{q})$ can be converted into $d_z(\boldsymbol{q})$ and vice versa. If the representation is unimportant, the index is omitted, i. e., $d(\boldsymbol{q})$ is used. Depth maps are a per image description of the scene's surface geometry.

The values of the confidence map $c(\boldsymbol{q})$ allow storing a confidence value for each pixel that represents the reliability of the color and depth information for this pixel. The range of the confidence value is $[0, 1]$, where a larger value means that the information is more reliable. For instance, if some region in the captured image is known to be corrupted, the confidence value for all pixels of this region can be set to zero.

A light field that additionally contains a depth map and a confidence map for each captured

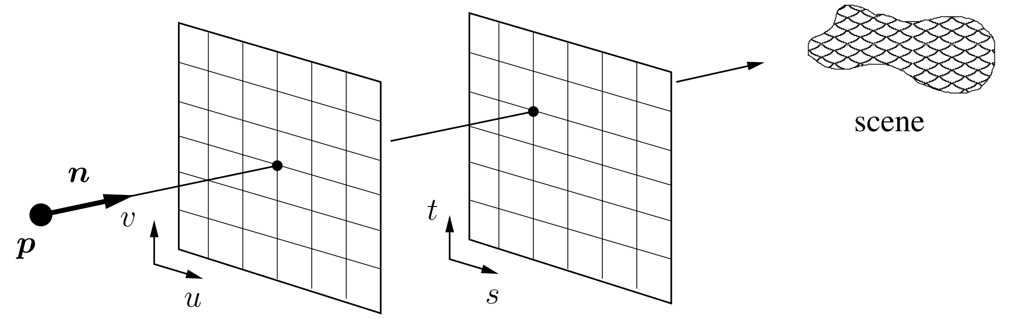


Figure 3.3: Two-plane parameterization of light rays. A light ray through $\boldsymbol{p}$ with direction $\boldsymbol{n}$ is defined by connecting a point on the uv-plane to the st-plane.

image will be denoted as *DC light field*. For storing samples of a DC light field, the *DC plenoptic function* $\boldsymbol{\psi}_{\boldsymbol{5,dc}}$ is defined as:

$$\boldsymbol{\psi}_{\boldsymbol{5,dc}}(\theta, \phi, \boldsymbol{p}) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{q}) \\ d(\boldsymbol{q}) \\ c(\boldsymbol{q}) \end{pmatrix}, \tag{3.20}$$

with $d(\boldsymbol{q}) \in \mathrm{I\!R}$ and $c(\boldsymbol{q}) \in [0, 1]$. Each confidence map is initialized with $c(\boldsymbol{q}) = 1 \ \forall \ \boldsymbol{q}$, i. e., all captured pixels are assumed to be correct. The additional information contained in a DC light field can be used to improve the quality of its visualization (see Section 3.2).

3.2 Light Field Visualization

This section summarizes some selected visualization methods for light fields. The last approach presented in this section (*unstructured lumigraph rendering*, page 46) is the most sophisticated among these. It combines the advantages of light field rendering with those of view-dependent texture mapping (an alternative image-based rendering approach). A comparison of image-based rendering approaches can be found in [Büh01].

3.2.1 Two-Plane Light Fields

As shown in Section 3.1, the parameter vector $(\theta, \phi, \boldsymbol{p})$ of the plenoptic function $\boldsymbol{\psi}_{\boldsymbol{5}}$ has only 4 degrees of freedom. In order to overcome the redundancy of the 5-D representation, Levoy [Lev96] proposes to parameterize light rays (plenoptic function samples) by their intersections with two planes (see Figure 3.3). The planes can be fixed arbitrarily in space, but usually parallel planes are used. A local coordinate system is defined for each plane: (u, v) for the first and (s, t) for the second plane. A light ray is then defined by connecting a point on the uv-plane to the st-plane. This leads to the 4-D plenoptic function:

$$\boldsymbol{\psi}_{\mathbf{4}}(u, v, s, t) = \boldsymbol{f}(\boldsymbol{q})\,. \tag{3.21}$$

Each pair of planes is called *light slab*. A light field that uses this kind of parameterization will be denoted as *two-plane light field* or *PP light field*. Intuitively, six light slabs, i. e., one light slab for each side of a cube, are sufficient for representing all possible samples of the plenoptic function for a scene. Only one decision has to be made: is the scene/object *inside* the cube or *outside*? For flyarounds of a (small) object, the light slab cube has to be defined in such a way that the object lies inside. The light slab cube would lie in the middle of a scene for a panoramic view. A flyaround for a small object was realized by [Lev96] with 4 light slabs (the top and bottom light slab was not recorded). This is easier than generating a light field for a panoramic view.

So far, the parameter space of the 4-D plenoptic function is continuous. However, the representation of this function in a computational framework requires a discretization [Gor96]. Therefore, a discrete subdivision of each plane has to be chosen. One method is to move the camera on an arbitrary regular grid on the uv-plane and define the st-plane to be parallel to the uv-plane with a distance of the focal length F. This means the image plane of the camera is identical to the st-plane and the resolution of the chip defines the grid. Then each captured image defines a bunch of light rays through one point $(i, j)^{\mathrm{T}}$ on the uv-plane. Another technique was used in [Gor96]: the discretization is defined by choosing N_{uv} subdivisions in the u and v dimensions and N_{st} subdivisions in s and t. A quadrilinear basis function $B_{i,j,k,l}$ is defined which has a value of 1 at grid point $(i, j, k, l)^{\mathrm{T}}$ and drops off to zero at all neighboring grid points. If $\boldsymbol{\psi}_{4,d}(i, j, k, l)$ denotes the discrete plenoptic function value at the 4-D grid point $(i, j, k, l)^{\mathrm{T}}$, the computation of the discrete function values $\boldsymbol{\psi}_{4,d}(i, j, k, l)$ is done by integrating $\boldsymbol{\psi}_{\mathbf{4}}$ against the *duals* of the basis functions, where in [Gor96] the original basis functions $B_{i,j,k,l}$ are used as approximation of their own duals. This step can be interpreted as point sampling $\boldsymbol{\psi}_{\mathbf{4}}$ after it has been low pass filtered with the dual basis function, in this case with $B_{i,j,k,l}$. The continuous plenoptic function $\boldsymbol{\psi}_{4,r}$ is then reconstructed as the linear sum

$$\boldsymbol{\psi}_{4,r}(u, v, s, t) := \sum_{i=0}^{N_{uv}} \sum_{j=0}^{N_{uv}} \sum_{k=0}^{N_{st}} \sum_{l=0}^{N_{st}} \boldsymbol{\psi}_{4,d}(i, j, k, l) B_{i,j,k,l}(u, v, s, t)\,. \tag{3.22}$$

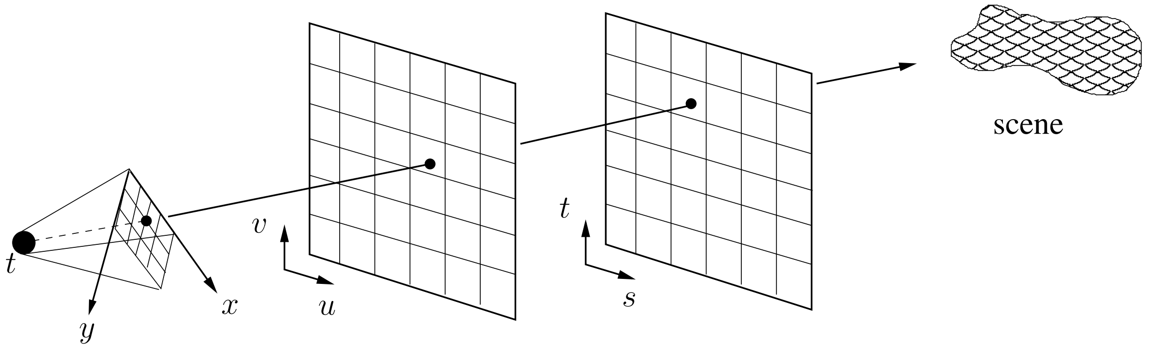


Figure 3.4: Rendering of a two-plane light field. For each pixel the corresponding light ray is projected into the uv- and st-plane. The 16 nearest samples of $\psi_{4,d}(i,j,k,l)$ are used for quadrilinear interpolation.

The additional index r labels the plenoptic function $\psi_{4,r}$ as *reconstructed* from discrete values of $\psi_{4,d}$.

The main advantage of PP light fields is the possibility of generating new views of the scene in real-time (≥ 25 frames per second). Given a PP light field, a 2-D slice of light rays must be resampled from the 4-D light slabs. The process can be divided into two steps: first, the continuous parameters (u, v, s, t) for the pixel $\boldsymbol{q}$ are computed; then, the color value $\boldsymbol{f}(\boldsymbol{q})$ is resampled for those parameters. The idea for implementing the first step is to use a projective mapping which can be implemented on graphics hardware (in real-time). The light ray corresponding to $\boldsymbol{q}$ is projected into the uv- and st-plane, respectively. The second step is achieved by interpolating ψ_4 from the nearest samples of $\psi_{4,d}$: the 16 nearest samples are used for a quadrilinear interpolation (see Figure 3.4). For images of a collection of light slabs each light slab is drawn sequentially. If the light slabs do not overlap, each pixel is only drawn once.

Gortler [Gor96] extended the light field by approximative 3-D shape information to correct errors during the rendering process. The techniques used in [Gor96] for reconstruction of the 3-D shape are described in Section 3.3, page 48. Approximate 3-D shape information is used in such a way that for a given light ray, defined by the continuous vector $(u, v, s, t)^{\mathrm{T}}$, and a given discrete neighbor point $(i, j)^{\mathrm{T}}$ of $(u, v)^{\mathrm{T}}$, a new light ray $(i, j, s', t')^{\mathrm{T}}$ can be calculated that intersects the same geometric location on the object as the original ray $(u, v, s, t)^{\mathrm{T}}$. This new light ray is calculated for each discrete neighbor of $(u, v)^{\mathrm{T}}$. These four light rays are then used for interpolation. If $z = d(u, v, s, t)$ denotes the depth value at which the light ray $(u, v, s, t)^{\mathrm{T}}$ first intersects a surface of an object (see Figure 3.5), then [Gor96]:

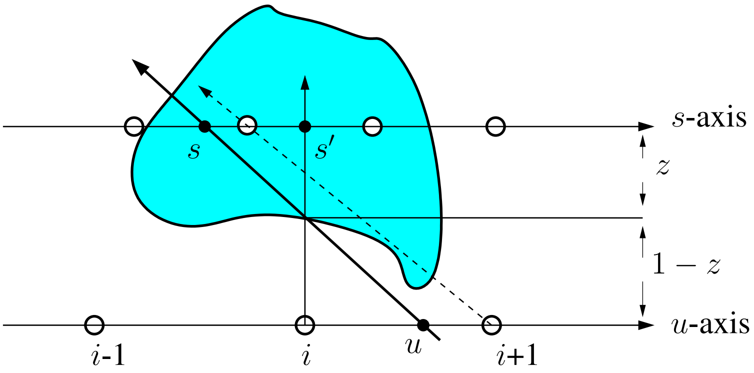


Figure 3.5: Depth correction visualized for two dimensions of a light field. The light ray $(u,s)^{\mathrm{T}}$ intersects the object at depth z, with $z=0$ for points on the s-axis. The distance between s- and u-axis is 1. A new light ray $(i,s')^{\mathrm{T}}$ through a discrete neighbor i of u, intersecting the object at the same surface point, is obtained by examining similar triangles: $\frac{s'-s}{z}=\frac{u-i}{1-z}$. This equation can be solved for s'. The new light ray $(i,s')^{\mathrm{T}}$ is more accurate than the one of the nearest discrete neighbors (dashed line).

$$s' = s + (u-i)\frac{z}{1-z}\,, \tag{3.23}$$

$$t' = t + (v-j)\frac{z}{1-z}\,. \tag{3.24}$$

The 3-D shape information is accounted for by defining the (new) basis functions

$$B'_{i,j,k,l}(u,v,s,t) = B_{i,j,k,l}(u,v,s',t') \tag{3.25}$$

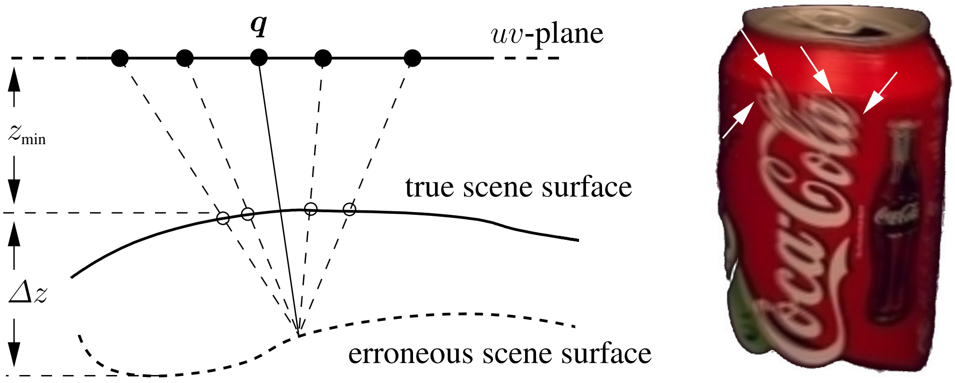
for the reconstruction of the plenoptic function $\psi_{4,r}$, i. e., first s' and t' are computed, then the "old" basis function is evaluated. In the system proposed by Gortler [Gor96], depth corrected quadrilinear basis functions are used.

Even when the geometry of the scene surface is merely known approximately, photo-realistic rendering using a PP light field is possible if the following equations for the maximum distance $(\Delta u_{\max}, \Delta v_{\max})^{\mathrm{T}}$ of neighboring (camera) grid positions in the uv-plane hold [Hei04, Cha00]:

$$\Delta u_{\max} = \frac{z_{\min}^2 - \Delta z^2}{2F_{\mathrm{x}}\eta_{\max}\Delta z}\,, \qquad \Delta v_{\max} = \frac{z_{\min}^2 - \Delta z^2}{2F_{\mathrm{y}}\eta_{\max}\Delta z}\,, \tag{3.26}$$

where $z_{\min}$ is the minimum distance of all surface points to the uv-plane, Δz is the assumed maximum error of the computed value for $z_{\min}$, and $\eta_{\max}$ is the maximum frequency in the image. If the equations above are not fulfilled, so-called *ghosting artefacts* occur during the rendering process (see Figure 3.6).

Finally, it should be mentioned that the visualization by PP light fields is only capable of



(a) Interpolation with depth uncertainty **(b)** Ghosting artefacts

Figure 3.6: Ghosting artefacts occur if the scene geometry is not known exactly. Then, wrong light rays (color information) are used to interpolate the color value for a pixel $\boldsymbol{q}$ (see (a)). A typical ghosting artefact of a reconstructed (rendered) image from a light field is a multiple-occurring edge (cf. the places marked with white arrows in (b)). Generally, ghosting artefacts reduce the sharpness of the reconstructed image.

reconstructing/rendering views lying "inside" the recorded views (interpolation). Extrapolation is not possible. More general approaches of light field visualization, capable of unstructured camera positions and extrapolation, are described in the next section.

3.2.2 Free Form Light Fields

The basic idea of free form light field visualization is to use the input data directly. In our case, each captured image represents a bundle of samples of the plenoptic function. No particular discretization step is applied.

In [Hei04] three approaches are described for free form light field visualization. Local depth information is assumed to be available. All methods are based on *mapping via planes*, i. e., how to map an image onto a 3-D plane and vice versa. Before summarizing the three approaches, the theory of mapping via planes is introduced.

Let a 3-D plane be defined by one point $\boldsymbol{x}_0$ on the plane and two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ spanning the plane. Each 3-D point $\boldsymbol{w}$ on the plane can then be represented by

$$\boldsymbol{w} = \alpha \cdot \boldsymbol{x}_1 + \beta \cdot \boldsymbol{x}_2 + \boldsymbol{x}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_0] \begin{pmatrix} \alpha \\ \beta \\ 1 \end{pmatrix} , \tag{3.27}$$

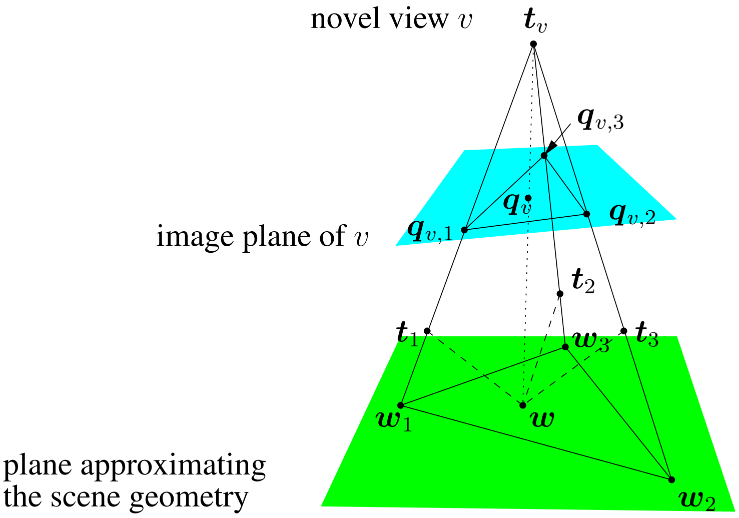


Figure 3.7: Free form light field visualization (adapted from [Hei04]). The virtual viewing ray corresponding to pixel $\boldsymbol{q}_v$ (dotted line) is interpolated from three recorded images, where $\boldsymbol{t}_1$,$\boldsymbol{t}_2$, and $\boldsymbol{t}_3$ are the corresponding viewpoints. The contributing light rays $\boldsymbol{q}_{v,1}, \boldsymbol{q}_{v,2}$ and $\boldsymbol{q}_{v,3}$ (dashed lines) are obtained by a mapping via the plane that approximates the scene geometry.

where α and β are the coordinates of $\boldsymbol{w}$ in the local coordinate system of the plane. A mapping between the local plane coordinate system of the 3-D plane and camera image plane coordinates is obtained by inserting equation (3.27) into equation (3.9), page 37:

$$\underline{\boldsymbol{q}} \sim \underbrace{\boldsymbol{K}\boldsymbol{R}^{\mathrm{T}} \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_0 - \boldsymbol{t}\right]}_{=:\boldsymbol{H}} \begin{pmatrix} \alpha \\ \beta \\ 1 \end{pmatrix} . \tag{3.28}$$

The 3×3 homography matrix $\boldsymbol{H}$ fully describes the mapping. In order to re-project the image of a camera onto the plane followed by a projection into another camera, a pixel $\underline{\boldsymbol{q}}_1$ of the first camera is multiplied by the inverse of the homography matrix $\boldsymbol{H}_1$ of the first camera. Then the obtained local coordinates are projected into the second camera using $\boldsymbol{H}_2$, yielding the corresponding pixel $\underline{\boldsymbol{q}}_2$:

$$\underline{\boldsymbol{q}}_2 = \boldsymbol{H}_2 {\boldsymbol{H}_1}^{-1} \underline{\boldsymbol{q}}_1 \, . \tag{3.29}$$

Projective mappings, in general, are provided by graphics hardware in real-time.

Single plane approach [Koc99a, Koc99b, Hei04]: The geometry of the scene is approximated by a single plane and the camera centers do not need to be located at regular grid positions. Figure 3.7 illustrates the approach: for a novel view v that has to be reconstructed, all projection centers of recording camera positions are projected into the image plane of v (using equation (3.9)). For each pixel $\underline{\boldsymbol{q}}_v$ the three neighboring projections $\boldsymbol{q}_{v,1}, \boldsymbol{q}_{v,2}$ and $\boldsymbol{q}_{v,3}$ are obtained applying a Delaunay triangulation [Lee80] of the projected centers first. The light rays of each of the three corresponding cameras can be determined by mapping via the plane that approximates the scene geometry:

$$\underline{\boldsymbol{q}}_k = \boldsymbol{H}_k {\boldsymbol{H}_v}^{-1} \underline{\boldsymbol{q}}_v, \ k = 1, 2, 3 \,. \tag{3.30}$$

${\boldsymbol{H}_v}^{-1}$ is the mapping of $\underline{\boldsymbol{q}}_v$ onto the geometry plane and $\boldsymbol{H}_k$ re-projects the point into the camera k. Since $\underline{\boldsymbol{q}}_k$ in general will not be a discrete value, the corresponding color value is obtained by bilinear interpolation. The weights of the three contributing light rays are defined according to the distance of the projected camera centers to $\boldsymbol{q}_v$ and in such a way that their sum is always 1. If $\boldsymbol{q}_v = \boldsymbol{q}_{v,i}$, then the weight for camera i is 1 and the others are zero. Between the triangle corners the weights are interpolated linearly. The whole image is built as a mosaic of triangles. All required operations are provided by graphics hardware, which allows very fast rendering.

Adaptive geometry approach [Hei99, Koc01, Hei04]: This approach employs local depth information. The single plane approach is extended by calculating the plane that approximates scene geometry *for each triangle* rather than for the whole scene. This improves accuracy but requires more computation time. The plane used for mapping is redefined for each triangle. The points $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, and $\boldsymbol{w}_3$ define the plane. They are the intersections of the scene geometry with the line through $\boldsymbol{t}_i$ and $\boldsymbol{t}_v$ for $i = 1, 2, 3$ (cf. Figure 3.7):

$$\boldsymbol{w}_i = d_t(\boldsymbol{H}_i {\boldsymbol{H}_v}^{-1} \underline{\boldsymbol{q}}_v) \frac{\boldsymbol{t}_i - \boldsymbol{t}_v}{\|\boldsymbol{t}_i - \boldsymbol{t}_v\|} + \boldsymbol{t}_i \,, \tag{3.31}$$

where $d_t(\boldsymbol{H}_i {\boldsymbol{H}_v}^{-1} \underline{\boldsymbol{q}}_v)$ is the distance between $\boldsymbol{t}_i$ and the scene geometry in the direction $\boldsymbol{t}_i - \boldsymbol{t}_v$. A small number of triangles (projected camera centers) results in a coarse approximation of the scene geometry. The approximation can be refined by subdividing a triangle into four subtriangles. The subdivision process may be applied recursively until the required approximation quality is achieved. An implementation of this approach can be found in [Sch01b], exploiting graphics hardware allows real-time rendering.

Extrapolation approach [Hei04]: The previous two approaches implicitly assume that a novel view of the scene can be composed of several triangles which are built by projected camera centers. For those cases in which this constraint is violated, e. g., if the camera is moved along a straight line over the scene, an extension of the adaptive geometry approach is needed. It should be capable of extrapolating views to guarantee at least rendering of novel views which are close to the recorded views.

The principle of extrapolation can be illustrated if it is assumed that only a single source view should be used for rendering a novel view. The color value of a pixel of the novel view is obtained by computing the intersection point of the corresponding light ray with the scene surface and projecting this point into the source view. The surface of the scene is thereby defined by the camera parameters $\boldsymbol{K}$, $\boldsymbol{R}$, $\boldsymbol{t}$ and the depth information $d_t(\boldsymbol{q}_s)$ for each pixel $\boldsymbol{q}_s$ of the source view. If the search range on the light ray is restricted between a minimal and a maximal depth plane, the calculation can be done very fast: the line between the intersection points is projected into the source image and only the depth values for the projected line have to be searched. If two depth map entries are found where the depth on the light ray is between those two values, the intersection point is found. The final value can be interpolated according to the depth differences of the two points.

If no intersection point exists, i. e., the light ray passes outside the known scene surface, and therefore no valid color value can be calculated, a fixed color value, e. g., black has to be used.

This basic extrapolation principle is accelerated and extended to more than one source view: Instead of extrapolating each pixel of the novel view a regular triangulation grid is chosen where four points of a square define two triangles. [Hei04] uses a grid of 20×20 points, i. e., 722 triangles. For each grid point the intersection with the scene surface is computed for each source view. Each triangle of the novel view is drawn by overlapping N_v triangles from different source views (e. g., $N_\text{v} = 5$). The triangles are mapped from the source views into the novel view by the plane defined by the particular triple of 3-D points. All contributing triangles are overlayed, i. e., weighted and added. The weight for a triangle point is defined relative to the angle between the novel light ray and the source light ray for this point (the smaller the angle the higher the weight). Although the use of graphics hardware would be possible for this approach, only the straight forward software-based implementation is currently available.

Unstructured lumigraph rendering [Büh01]: This approach meets the following objectives while providing rendering in real-time:

- Use of geometric information: If knowledge of the geometry of the scene is available, it is used to increase the quality of rendering.
- Unstructured input: Arbitrary camera movements are possible and a resampling step is not used (e. g., needed to “convert” a light field to a PP light field, see Section 3.3). This includes that forward camera motion is handled well.

- Epipole consistency: If a desired light ray passes through the projection center of a source camera, it is reconstructed directly from the source image, provided that the light ray is inside the field of view of the camera.
- Minimal angular deviation: Source image light rays with similar angles to the desired light ray are used.
- Continuity: Reconstructed neighboring points have similar color values to avoid artefacts.
- Resolution sensitivity: Image pixels are not measured by a single light ray (point on the scene surface), but instead by an integral over a set of rays (area on the scene surface). This is taken into account during the rendering. It is especially important if light rays from cameras with varying distance or different focal length are combined.
- Equivalent ray consistency: Through an empty region of space, the light ray along a given line-of-sight is reconstructed consistently.

At first, a “camera blending field” is generated. It describes how each source camera is weighted to reconstruct a given pixel. The computation of the field is based on the specified objectives. Three penalties are defined to compute the weight of a camera: angular penalty π_{ang}, resolution penalty π_{res}, and field-of-view penalty π_{fov}. Additionally, a k-nearest neighbor approach is applied: only those cameras with the k smallest penalties are used for interpolation (in [Büh01] $k = 4$ was used). Then the weight $w(i)$ of camera i is defined as

$$w(i) = 1 - \pi(i)/\pi_{\max} \,, \tag{3.32}$$

where

$$\pi(i) = \alpha \cdot \pi_{\text{ang}} + \beta \cdot \pi_{\text{res}} + \gamma \cdot \pi_{\text{fov}} \,, \tag{3.33}$$

$\pi_{\max}$ is the largest of the k smallest penalties, and the scalars α, β, and γ control the relative importance of the different penalties for the overall penalty $\pi(i)$. Finally, all weights $w(i)$ are normalized to sum to unity.

For $\pi_{\text{ang}}(i)$ the angle between the desired light ray and the light ray from the surface point through the projection center of camera i is used (the larger the angle, the larger the penalty, the lesser the weight). For π_{res} the distances of the projection centers of the novel view $\boldsymbol{t}_n$ and the source view $\boldsymbol{t}_s$ to the scene point $\boldsymbol{w}$ are used:

$$\pi_{\text{res}}(i) = \max(0, \|\boldsymbol{w} - \boldsymbol{t}_s\| - \|\boldsymbol{w} - \boldsymbol{t}_n\|) \,, \tag{3.34}$$

i. e., the shorter the distance to the scene point of the novel view compared to the source view, the larger the penalty (undersampling is punished). The penalty $\pi_{\text{fov}}(i)$ is defined as 0 for light rays within the field-of-view of camera i and as ∞ otherwise.

The strategy for real-time rendering is to evaluate the camera blending field at a sparse set of points and to interpolate the values in between. The samples for the blending field are selected as follows: All vertices of the geometric information, given as 3-D points/vertices, are projected into the blending field and used as sample points. Next, the projections of every source camera center into the novel view are added to the set of sample points. Finally, a regular grid of sample points is defined to obtain a dense set of samples. Applying a Delaunay triangulation a triangular mesh is obtained and used for interpolation. For each vertex of the triangular mesh the blending weights are computed and used for rendering. The spacing of the regular grid can be defined arbitrarily where a smaller grid spacing leads to slower rendering because the blending weights for more vertices have to be computed.

3.2.3 Inclusion of Confidence Maps

Pixels (light rays) marked with low confidence are either not used at all during the rendering process (confidence zero) or with reduced weight. The currently implemented rendering tools allow the confidence value to be either zero or one, i. e., $c(\boldsymbol{q}) \in \{0, 1\}$.

3.3 Light Field Reconstruction

In general light fields are reconstructed from image sequences captured by a camera. The goal of light field reconstruction is the computation of the data required for a light field: the intrinsic and extrinsic camera parameters for each captured image and, if possible, depth and confidence maps. Two different kinds of approaches are described in this section. The first kind uses mechanical calibration to compute the extrinsic camera parameters. The second one allows for the reconstruction of a light field based only on the image sequence.

3.3.1 Mechanical Calibration

This kind of approach assumes the intrinsic camera parameters to be constant. They can then be estimated in advance using a camera calibration technique (see Section 4.1.2, page 61). The computation of the extrinsic camera parameters (pose) is done by using additional apparatus:

- Different types of pose determination systems exist: robot arms as well as magnetic and optical tracking systems. Using one of these the computation of the extrinsic camera parameters is possible (e. g., see [Vog04a, Sch02b, Sch01a, Sal01, Sch01a]). Optical and magnetic tracking systems require a so-called *target* to be attached to the tracked object/camera. Robot arms require the camera to be attached to the arm of the robot. In one of the first publications about light fields by Levoy and Hanrahan in 1996 [Lev96] a special pose determination system was built: a computer-controlled planar camera gantry. It allows digitizing images of an object on a regular grid (suitable for PP light fields). The object is placed on a rotating tripod which allows capturing images of the object from all sides ($360°$). The camera is equipped with pan and tilt motors. Given the angles of the pan and tilt motors the extrinsic camera parameters can be computed.

- A specially designed environment which provides calibration markers for determining the extrinsic camera parameters can be used. Markers are easily detectable world points of which the 3-D coordinates are known. Given the intrinsic camera parameters three markers are sufficient to compute the extrinsic parameters for an image in which these markers were detected [Har94]. In order to increase the accuracy of the result usually as many markers as possible are detected and used. In [Gor96] this kind of approach was used. It allows capturing the image sequence by a hand-held camera without restrictions to the pose. In [Gor96] a rough estimate of the shape (depth map) is calculated with an octree construction algorithm [Sze93] that requires a segmentation of each image into object/background. The idea is to start with a voxel that contains the whole object. Voxels at a coarse level of the octree are then projected into the image, and only if the voxel falls on the silhouette of the object it is marked for further subdivision. At the end a collection of voxels describing a volume that contains the object is obtained.

- A fixed camera array can be used for capturing images [Wil02]. At the beginning the pose of each camera is determined, e. g., by a camera calibration technique, and a light field is obtained by simultaneously capturing an image from all cameras. The number of images contained in the light field equals the number of cameras and therefore restricts the spatial resolution. The main challenge of this approach is the hardware setup for simultaneously capturing and transferring the image data.

3.3.2 Structure-From-Motion

It is known that extrinsic camera parameters as well as surface geometry can theoretically be determined by only using the captured image sequence [Ull79]. Algorithms solving the problem of determining the geometry of the scene from point correspondences are referred to as *structure-from-motion* algorithms. An overview over many structure-from-motion algorithms is found in [Har03]. The method developed in [Hei04] is summarized in the following paragraphs. It consists of five steps:

1. **Extraction of 2-D point correspondences.** Due to the static scene assumption for light fields, 3-D scene points are fixed and detectable by their 2-D projections into the captured camera images. For each visible scene point and its 2-D projection into one view a corresponding 2-D projection exists in another view if the 3-D scene point is still visible in that view. The computation of 2-D point correspondences from frame to frame is also called *point tracking*. The differential point tracking method of Tomasi and Kanade [Tom91] with the extensions made by Shi [Shi94] is employed. Points that can be tracked well are selected for tracking, according to the minimal eigenvalue of the so-called *structure matrix*

   $$\boldsymbol{G} = \begin{pmatrix} f_x^2(\boldsymbol{q}) & f_x(\boldsymbol{q})f_y(\boldsymbol{q}) \\ f_x(\boldsymbol{q})f_y(\boldsymbol{q}) & f_y^2(\boldsymbol{q}) \end{pmatrix}, \tag{3.35}$$

   where $f_x(\boldsymbol{q})$ and $f_y(\boldsymbol{q})$ are the first derivatives of the gray-value image f at pixel $\boldsymbol{q}$ in x- and y-direction. Around each tracked point a feature window is defined, i. e., feature windows are tracked rather than single points. Gray-value pixels are used by this point tracking method. Since color images are usually captured, either the color image has to be converted into a gray-value image or only one of the three color channels is used. In this approach the green-channel of the color image is used for point tracking. In [Hei04] it was shown that an extension to color images does not lead to better tracking results if the parameters for tracking are chosen well (tracking window size ≥ 7).

   All following steps are based on the knowledge of 2-D point correspondences.

2. **Outlier detection.** Outliers occur due to different “problems” in real images, e. g., the Lambertian assumption that a surface point results in identical colors when viewed from arbitrary viewpoints is violated because of specular effects and mirroring, or because the scene contains occluding or self-occluding contours. For achieving good results with the factorization method of the next step, the number of outliers should be as small as possible. The *trifocal tensor* motion constraint together with an LMedS technique [Rou87]

eliminates outliers if at least six points are visible in three images. With the trifocal tensor the camera parameters/projection matrices and scene points can be calculated from those six points. Projecting the computed scene points $\widehat{\boldsymbol{w}}_j$ into the images using the computed projection matrices $\widehat{\boldsymbol{P}}_i$, the *back-projection error*

$$\epsilon_{\mathrm{BPE}} = \sum_i \sum_j \|\boldsymbol{q}_{i,j} - \widehat{\boldsymbol{q}}_{i,j}\| \quad \text{with} \quad \underline{\widehat{\boldsymbol{q}}}_{i,j} = \widehat{\boldsymbol{P}}_i \underline{\widehat{\boldsymbol{w}}}_j \tag{3.36}$$

is a measure for the correctness of the contributing points where $\boldsymbol{q}_{i,j}$ is the j-th 2-D point in the i-th image obtained by point tracking and $\widehat{\boldsymbol{q}}_{i,j}$ is obtained by projection and eliminating the homogeneous component: $\boldsymbol{q}_{i,j}$ and $\widehat{\boldsymbol{q}}_{i,j}$ should be the same point.

3. **Factorization of initial sequence.** This is the core of this structure-from-motion approach. Camera parameters and scene geometry are calculated from 2-D point correspondences. Let N_{w} scene points $\boldsymbol{w}_j$, $1 \leq j \leq N_{\mathrm{w}}$, be visible in N_{f} captured frames and $\boldsymbol{q}_{i,j}$ denote the j-th scene point projected into the i-th frame. The basic idea of the factorization method is the decomposition of a measurement matrix $\boldsymbol{\Gamma}$ (containing all 2-D point correspondences) into a motion matrix $\boldsymbol{\Psi}$ (containing all projection matrices) and a shape matrix $\boldsymbol{\Phi}$ (containing all 3-D scene points):

$$\underbrace{\begin{pmatrix} \underline{\boldsymbol{q}}_{1,1} & \underline{\boldsymbol{q}}_{1,2} & \cdots & \underline{\boldsymbol{q}}_{1,N_{\mathrm{w}}} \\ \underline{\boldsymbol{q}}_{2,1} & \underline{\boldsymbol{q}}_{2,2} & \cdots & \underline{\boldsymbol{q}}_{2,N_{\mathrm{w}}} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\boldsymbol{q}}_{N_{\mathrm{f}},1} & \underline{\boldsymbol{q}}_{N_{\mathrm{f}},2} & \cdots & \underline{\boldsymbol{q}}_{N_{\mathrm{f}},N_{\mathrm{w}}} \end{pmatrix}}_{\boldsymbol{\Gamma}} = \underbrace{\begin{pmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \\ \vdots \\ \boldsymbol{P}_{N_{\mathrm{f}}} \end{pmatrix}}_{\boldsymbol{\Psi}} \underbrace{(\underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2, \ldots, \underline{\boldsymbol{w}}_{N_{\mathrm{w}}})}_{\boldsymbol{\Phi}} \tag{3.37}$$

The projection matrices $\boldsymbol{P}_k$, $1 \leq k \leq N_{\mathrm{f}}$, contain the intrinsic and extrinsic camera parameters. Since in general not all scene points are visible throughout the whole captured image sequence a subset of frames and 2-D correspondences has to be chosen where all scene points are visible in all frames. After defining a minimum number of scene points the largest subset of captured frames in which at least the chosen number of scene points are visible in all frames (initial sequence) can easily be computed.

The decomposition is done by a singular value decomposition [Tre97] of $\boldsymbol{\Gamma}$. However, the obtained solution is not unique since for any non-singular 4×4 matrix $\boldsymbol{D}$ the solution $\boldsymbol{\Gamma} = (\boldsymbol{\Psi}\boldsymbol{D})(\boldsymbol{D}^{-1}\boldsymbol{\Phi})$ is also valid. For simplicity the procedure for estimating $\boldsymbol{D}$ up to an unknown scalar (*self-calibration*) is omitted as well as a detailed description of further techniques applied to increase the robustness of the factorization (see [Hei04] for details).

4. **Extension to the whole sequence.** The aim is the reconstruction of the whole sequence. Up to now, camera parameters (and scene points) are only known for the initial sequence. The idea for obtaining the camera parameters of the remaining frames is to use the already reconstructed scene points as a calibration pattern for a new frame. All data required for applying a standard calibration algorithm [Tsa87, Zha00] are available: 2-D projections of 3-D world points. The 2-D projections are known from point tracking, the 3-D world points are computed by triangulating all 2-D projections of a scene point by solving a linear equation system and subsequent non-linear maximum-likelihood optimization of the back-projection error (cf. equation (3.36)). In this approach, instead of using a standard calibration algorithm that includes a linear estimation of the camera parameters followed by a non-linear minimization of the back-projection error, the non-linear minimization is applied directly. The initialization is defined by the already calibrated neighboring frame with the assumption that a continuous image sequence was recorded and therefore the difference will be very small. Furthermore, the intrinsic camera parameters are assumed to be constant, which restricts the minimization to the extrinsic camera parameters and increases robustness.

   New frames are added alternately at the beginning and at the end of the new frame. The experiments in [Hei04] demonstrated that a minimum of 20 points are required for a reliable estimation. If less points are visible the algorithm therefore stops.

5. **Reconstructing scene geometry.** Information about scene geometry can be used by many light field visualization approaches to increase the quality of the rendered images. A rough representation of the scene geometry is given by the reconstructed 3-D scene points. It can be used to generate dense depth maps for each image by first setting the depth value for the projection of each reconstructed scene point and then interpolating all missing points from the depth values of the three nearest projected points, weighted by the distance to each point. The number of known scene points can be increased if the reconstructed camera parameters are used to find new point correspondences. The idea is to restrict the search range for a point correspondence to a line (the *epipolar line*), which increases the probability of finding a point correspondence and therefore increases their number.

3.4 Dynamic Light Fields

Many natural scenes are dynamic rather than static, i. e., something in the scene is moving. For dynamic light fields the *static scene* assumption is not valid (cf. Section 3.1, page 35). Objects in the scene can be deformable and are allowed to move. Dynamic light fields therefore store samples of $\boldsymbol{\psi}_{\mathbf{6}}(\theta, \phi, \tau, \boldsymbol{p})$ instead of samples of $\boldsymbol{\psi}_{\mathbf{5}}(\theta, \phi, \boldsymbol{p})$. A dynamic light field and an extended dynamic light field are defined according to the definitions of a light field and an extended light field.

A *dynamic light field* consists of all captured images of a scene, a time τ for each image, and the projection parameters. The information contained in a dynamic light field allows the determination of the plenoptic function $\boldsymbol{\psi}_{\mathbf{6}}(\theta, \phi, \tau, \boldsymbol{p})$ for the bundle of light rays corresponding to the pixels of each captured image. A *DC dynamic light field* is a dynamic light field that additionally contains a depth map and a confidence map for each image.

When capturing image sequences for light field reconstruction with a camera, the frame number is used as time value τ. All techniques described in Section 3.2 are based on static light fields, but they can all be extended if information about the dynamic changes in the scene is available:

1. Determine all subsets of captured frames that correspond to the same static scene. For instance, for a periodical object movement with known period, corresponding frames can be determined. Another example are discrete movements, i. e., the object remains on the same position for a known time and then moves to the next position, ideally in zero time, e. g., a rabbit hopping over a meadow, or a chess game. The number of frames corresponding to the same static scene is also known if the points in time of the movements are known.

2. During rendering only plenoptic samples of frames with the *same* time value are used.

The described method can also be regarded as discretization of the time dimension. For each defined point in time where the scene is known to be static a static light field is reconstructed. Combining several static light fields, rendering new views depending on the pose of the camera as well as on a specified point in time is possible. If the new view is rendered from plenoptic samples with two or more different time values, interpolation is also done in the time domain.

This technique is applied in dynamic light field publications, e. g. see [Li98, Büh01, Wil02, Gol02, Sch04b]. Only the approach used for light field reconstruction differs. In the earliest approach [Li98] simulated data were used. [Büh01] and [Sch04b] use structure-from-motion algorithms for the reconstruction of real scenes, and in [Wil02] and [Gol02] the Stanford *Light Field Video Camera* provides several simultaneously captured images of real scenes. The prototype provides six images, the goal being 128.

An alternative way for dynamic light field reconstruction and visualization of scenes containing rigid moving objects was presented recently in [Sch05]: based on 2-D point tracking, points lying on rigid moving objects are separated automatically from static 2-D points. This allows reconstructing a static light field of the scene (background) with structure-from-motion techniques. For visualization of the *dynamic* light field, the objects are segmented in the image by using the detected 2-D points. Temporally neighboring images are then employed for rendering the object at a specific point in time τ, where the confidence value is set to 1 for object pixels if the object was visible at time point τ, to 0 if the object was not visible at time point τ, and to 0.5 for (static) background pixels.

Chapter 4

Image Enhancement

The last two chapters described the state of the art in computer assisted endoscopic surgery, light field reconstruction, and light field visualization. This is the first of three chapters describing the methods developed in this thesis. Experiments and evaluations of these methods can be found in Chapter 7. Sections 4.1 to 4.4 present real-time image enhancement methods. Section 4.5 explains the usage of image zoom and rotation. Section 4.6 describes a technique for image enhancement based on light fields.

Before explaining the methods, the system which allows applying these methods during minimally invasive operations is presented (see Figure 4.1). The main component is a typical video-endoscopic system [Ric05]. It includes a rack, an endoscopic camera, a light source, a carbon dioxide insufflator and a video monitor for displaying the image of the endoscope. In order to provide real-time computer assisted image enhancement methods, the system is extended by a 3.2 GHz PC (Pentium 4) with 3 GB memory containing a S-VHS frame grabber card and a second monitor. This setup allows grabbing the image from the endoscopic camera, processing it with the PC and displaying it on the second monitor at the same time as the original image.

Unfortunately, the camera was not designed to provide high quality (digital) images with as little noise as possible, but for displaying an image on a conventional TV monitor. The single-chip Charge-Coupled Device (CCD) camera provides S-Video PAL output: 25 *interlaced* color images per second, size 768×576 pixels (columns × rows). For each pixel the red, green, and blue color components are quantized using eight bits. An interlaced image consists of two *half images* with only half the number of rows (size 768×288 pixels). Two half images together result in one interlaced image $\boldsymbol{f}$. At the beginning of TV it was easier (faster) to provide 50 interlaced images instead of 25 non-interlaced (progressive scan) images. Therefore, TV monitors also use interlaced images, which is the reason why even modern video cameras provide this

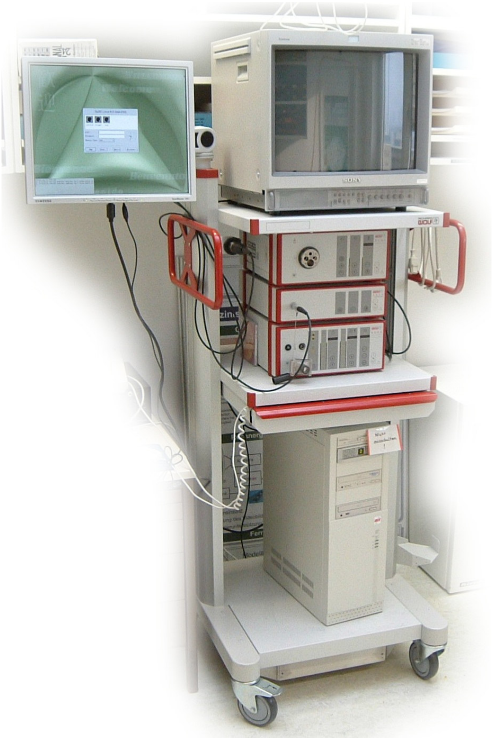


Figure 4.1: The real-time endoscopic image enhancement system: a typical video-endoscopy system on a rack is extended by a PC and a second monitor (on the left-hand side of the originally contained monitor, fixed with a positioning arm) to display the original and the processed image at the same time.

format. A typical video-endoscopic system displays the image of the camera/endoscope on a TV monitor. The human observer does not notice the difference between interlaced and progressive scan images. However, the difference has to be taken into account for digital image processing (cf. Section 5.1.1, page 84).

The question of describing the quality of the images remains. Different criteria can be employed to define the quality of an image captured by a digital camera. Despite the number of color channels and the color depth, i. e., the number of bits used to quantize each color channel, the *sensor noise* is a good measure with respect to digital image processing. Capturing the same static scene with the same illumination conditions from the same camera pose should lead to the same color value for each pixel. However, this is not true. The difference is called sensor noise, since the sensor of the frame grabber is the reason for the difference when assuming that everything else remains constant. The sensor noise is measured per color channel and defined as the standard deviation $\sigma(\boldsymbol{q})$ for each pixel $\boldsymbol{q}$ [Tru98]. Let $\boldsymbol{f}_0, \ldots, \boldsymbol{f}_{N-1}$ be N captured images, then

$$\sigma(\boldsymbol{q}) = \sqrt{\frac{1}{N-1}\sum_{k=0}^{N-1}\left(\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{f}_k(\boldsymbol{q})\right)^2} \tag{4.1}$$

with

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \frac{1}{N}\sum_{k=0}^{N-1}\boldsymbol{f}_k(\boldsymbol{q})\,. \tag{4.2}$$

Since $\sigma(\boldsymbol{q})$ defines the sensor noise only at pixel $\boldsymbol{q}$, the mean value of several or all pixel positions is normally used. Let $\mathcal{P}$ be a set of pixel positions. The mean standard deviation for the pixel positions in $\mathcal{P}$ is given as

$$\sigma_{\mathcal{P}} = \frac{1}{|\mathcal{P}|}\sum_{\boldsymbol{q}\in\mathcal{P}}\sigma(\boldsymbol{q})\,. \tag{4.3}$$

As will be shown in Section 7.2 the sensor noise of the endoscopic camera is approximately three times larger compared to a standard consumer video camera.

Examples of the methods described in the following three sections, namely distortion correction, color normalization, and temporal filtering are illustrated in Figure 4.2

4.1 Distortion Correction

For endoscopic operations, optical lenses with small focal lengths are used to enlarge the visible area and gain clarity. Since lenses with small focal lengths and perfect projection properties cannot be manufactured, the image is distorted, e. g., straight lines become bent (cf. Figure 4.3). Additionally, inaccuracies during the manufacturing process of the optics is another reason for image distortion. The distortion in general increases towards the borders of the image. The pinhole camera model described in Section 3.1.1 does not model image distortions.

There are two reasons for applying distortion correction. Firstly, the real world is not distorted. Since the goal is to enhance endoscopic live-images, the distortion during the projection of the real world onto the image plane should be corrected. In the case of endoscopic operations: a straight anatomical structure should also be straight in the projected image and not bent. Additionally, the spatial location of pixels in distorted images is non-linear, especially towards the borders of the image, leading to a false impression of distance and range. Secondly, the theory of light field reconstruction and visualization requires undistorted images since it is based on the pinhole camera model. If distorted images were used, the color value of pixels, especially those at the border of captured images, would not correspond to the light ray computed by applying

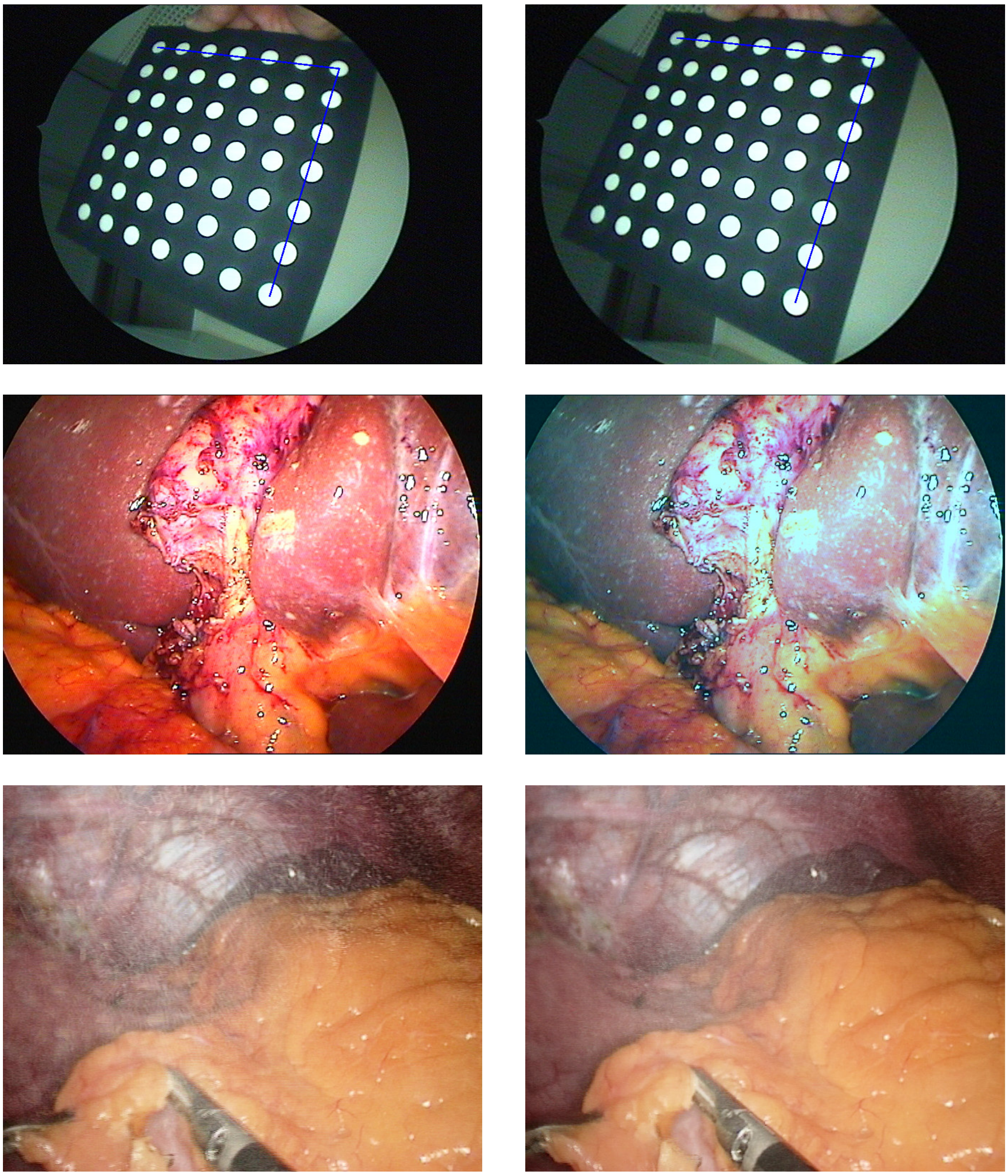
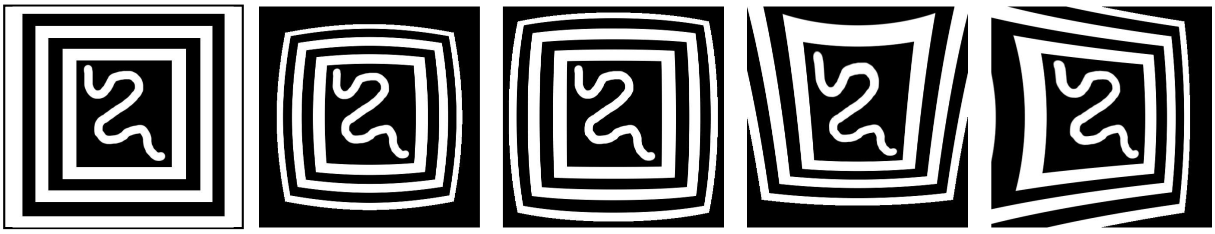


Figure 4.2: Examples of the developed methods for endoscopic image enhancement: distortion correction (top row), color normalization (middle row), and temporal filtering (last row). The enhanced image is always displayed to the right of the original image. The result of distortion correction becomes visible when regarding the two additionally plotted lines and the right edge of the calibration pattern (top row): withouth distortion, the edge should be straight and the lines should cut the intermediate circles into two equal parts.



(a) Original **(b)** Radial dist. **(c)** Radial dist. **(d)** Tang. dist. **(e)** Tang. dist.

Figure 4.3: Examples of image distortion types: (a) original image (b) image with radial distortion ($\kappa_1 = -1$) (c) image with radial distortion ($\kappa_2 = -1$) (d) image with tangential distortion ($p_1 = -0.2$) (e) image with tangential distortion ($p_2 = -0.2$).

the pinhole camera model. Therefore, each image sequence used for light field reconstruction is undistorted before it is further processed (see Chapter 5).

4.1.1 Distortion Model

Lens distortion can be modeled in two different ways: either the distorted point $(x_{\mathrm{ds}}, y_{\mathrm{ds}})^{\mathrm{T}}$ is obtained by adding a distortion term $(\delta_x, \delta_y)^{\mathrm{T}}$ to the undistorted point $(x_{\mathrm{us}}, y_{\mathrm{us}})^{\mathrm{T}}$ (cf. equation (4.4) and [Tsa87, Zha96]), or $(x_{\mathrm{us}}, y_{\mathrm{us}})^{\mathrm{T}}$ is obtained by adding a distortion *correction* term $\left(\tilde{\delta}_x, \tilde{\delta}_y\right)^{\mathrm{T}}$ to $(x_{\mathrm{ds}}, y_{\mathrm{ds}})^{\mathrm{T}}$ (cf. equation (4.5) and [Zha98, Zha99]):

$$\begin{pmatrix} x_{\mathrm{ds}} \\ y_{\mathrm{ds}} \end{pmatrix} = \begin{pmatrix} x_{\mathrm{us}} \\ y_{\mathrm{us}} \end{pmatrix} + \begin{pmatrix} \delta_x \\ \delta_y \end{pmatrix}, \tag{4.4}$$

$$\begin{pmatrix} x_{\mathrm{us}} \\ y_{\mathrm{us}} \end{pmatrix} = \begin{pmatrix} x_{\mathrm{ds}} \\ y_{\mathrm{ds}} \end{pmatrix} + \begin{pmatrix} \tilde{\delta}_x \\ \tilde{\delta}_y \end{pmatrix}. \tag{4.5}$$

The distortion is modeled in the sensor coordinate system, not in the image coordinate system. This is indicated by the index “s”.

For distortion correction, equation (4.4) has to be used: the discrete color value $\boldsymbol{f}(x_{\mathrm{us}}, y_{\mathrm{us}})$ is obtained by computing the distorted point $(x_{\mathrm{ds}}, y_{\mathrm{ds}})^{\mathrm{T}}$ and setting $\boldsymbol{f}(x_{\mathrm{us}}, y_{\mathrm{us}}) = \boldsymbol{f}(x_{\mathrm{ds}}, y_{\mathrm{ds}})$, where the color value for $\boldsymbol{f}(x_{\mathrm{ds}}, y_{\mathrm{ds}})$ is interpolated from the four discrete neighbor pixels.

Two different types of distortions (cf. Figure 4.3), radial and decentering or tangential distortion are modeled.

$$\begin{pmatrix} \delta_x \\ \delta_y \end{pmatrix} = \underbrace{\begin{pmatrix} x_{\text{us}}\left(\kappa_1 r^2 + \kappa_2 r^4\right) \\ y_{\text{us}}\left(\kappa_1 r^2 + \kappa_2 r^4\right) \end{pmatrix}}_{\text{radial distortion}} + \underbrace{\begin{pmatrix} 2p_1 x_{\text{us}} y_{\text{us}} + p_2\left(r^2 + 2{x_{\text{us}}}^2\right) \\ 2p_2 x_{\text{us}} y_{\text{us}} + p_1\left(r^2 + 2{y_{\text{us}}}^2\right) \end{pmatrix}}_{\text{tangential distortion}}, \tag{4.6}$$

where $r = \sqrt{{x_{\text{us}}}^2 + {y_{\text{us}}}^2}$. Radial distortions occur symmetrically around the principal point because the lenses are rotated during grinding. Therefore, only even powers of r occur in the radial distortion term. Tangential distortions are due to the alignment of the lenses and do not have such an extreme symmetric form.

According to the proposed model, lens distortion is fully described by the four coefficients: κ_1 and κ_2 (radial distortion), and p_1 and p_2 (tangential distortion).

In general, points are specified in image coordinates/pixels rather than in sensor coordinates. Applying distortion correction therefore involves conversion from image to sensor coordinates and vice versa. From a distorted point $(x_{\text{d}}, y_{\text{d}})^{\text{T}}$ specified in image coordinates, the corresponding point $(x_{\text{ds}}, y_{\text{ds}})^{\text{T}}$ in sensor coordinates is computed. Applying the distortion correction leads to the undistorted point $(x_{\text{us}}, y_{\text{us}})^{\text{T}}$ in sensor coordinates, which has to be transformed back into image coordinates, leading to the undistorted point $(x_{\text{u}}, y_{\text{u}})^{\text{T}}$. More formally, these three steps can be written as:

$$(x_{\text{d}}, y_{\text{d}})^{\text{T}} \overset{(1)}{\Longrightarrow} (x_{\text{ds}}, y_{\text{ds}})^{\text{T}} \overset{(2)}{\Longrightarrow} (x_{\text{us}}, y_{\text{us}})^{\text{T}} \overset{(3)}{\Longrightarrow} (x_{\text{u}}, y_{\text{u}})^{\text{T}}. \tag{4.7}$$

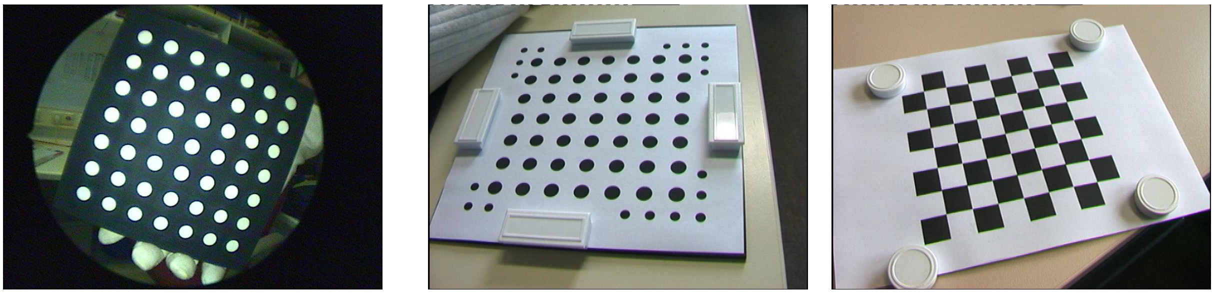
Step (2) was already explained. Steps (1) and (3) are defined by the following equations:

$$\begin{pmatrix} x_{\text{ds}} \\ y_{\text{ds}} \end{pmatrix} = \begin{pmatrix} dx\left(x_{\text{d}} - C_{\text{x}}\right) \\ dy\left(y_{\text{d}} - C_{\text{y}}\right) \end{pmatrix} \quad \begin{pmatrix} x_{\text{u}} \\ y_{\text{u}} \end{pmatrix} = \begin{pmatrix} \frac{x_{\text{us}}}{dx} \\ \frac{y_{\text{us}}}{dy} \end{pmatrix} + \begin{pmatrix} C_{\text{x}} \\ C_{\text{y}} \end{pmatrix}, \tag{4.8}$$

where $(C_{\text{x}}, C_{\text{y}})^{\text{T}}$ is the principal point, specified in pixels, dx and dy are the size of a pixel in x- and y-direction, specified in mm/pixel.

According to [Hei04, Tru98, Zha96, Tsa87], radial distortions represent the main part of divergence to the pinhole camera model and tangential distortions are negligible. Furthermore, if the distortions at the border of the image are not larger than five pixels, one radial coefficient is sufficient. For endoscopic images distortions up to 50 pixels at the border of the image occur. Therefore, the distortion model described in this section is applied for distortion correction of endoscopic images. In all experiments, both radial distortion coefficients were remarkably larger than zero and the tangential coefficients were very close to zero (cf. Table 7.4, page 150).

Note that the distortion coefficients are intrinsic camera parameters. They are required in



(a) 7 × 7 pattern with circles

(b) 7 × 7 pattern with circles and marked edges

(c) chess board pattern

Figure 4.4: Different types of 2-D calibration patterns: manufactured (a) or printed patterns attached to a planar surface (b,c), squares (c) or circles (a,b), symmetric (a,c) and asymmetric patterns (b). A 3-D calibration pattern can be built out of two 2-D patterns, which are arranged perpendicular to each other.

order to correctly model the projection of world points to undistorted image points. An algorithm for estimating the parameters of the model is described in the following section.

4.1.2 Camera Calibration

This section explains the process of determining the intrinsic and extrinsic camera parameters using a calibration pattern and Zhang's camera calibration algorithm [Zha00] in detail. The intrinsic camera parameters are required for applying the distortion correction algorithm as laid out in the last section. Extrinsic camera parameters will be required for the hand-eye calibration algorithms described in Sections 5.4.2 and 5.5.2, and for many experiments of Section 7.3.

All camera calibration algorithms are based on world-image point correspondences. These are usually obtained using a calibration pattern (see Figure 4.4). A calibration pattern is manufactured or printed so that known points of the pattern are easily detectable in an image captured by a digital camera. If one is only interested in the intrinsic camera parameters, a symmetric calibration pattern is sufficient. If extrinsic camera parameters are to be determined it is proposed to use an asymmetric calibration pattern as shown in Figure 4.4(b). The advantage of an asymmetric pattern is that the world-image point correspondences can be calculated automatically whereas this is only possible with restrictions for the movement of the camera for a symmetric calibration pattern. An asymmetric calibration pattern is suited as well as a symmetric calibration pattern for determining intrinsic camera parameters. The calibration procedure used here requires either symmetric or asymmetric calibration patterns with circles in a regular 2-D grid.

How world-image point correspondences are determined is usually not addressed in publications about camera calibration algorithms. In many cases this is done manually since only a small number of images are used. For the algorithms and experiments described here, camera calibration should be completely automatic. If the extrinsic camera parameters of a short image sequence with 50 to 100 images have to be determined, manual assignment of world-image point correspondences for each image would be a very tedious and time consuming task. The following algorithm therefore includes an automatic method for the assignment of world-image point correspondences (steps 2 and 3):

1. **Capturing images of a calibration pattern:** Zhang's algorithm [Zha00] works best when five to ten images with different orientations of the calibration pattern are captured. There is even a degenerate configuration for this algorithm if the world points are all on parallel planes (e. g., if the camera pose is fixed and the pattern is only moved on a table).

2. **Determining 2-D calibration points:** The projections of the 3-D world points of the calibration patterns have to be identified in the image, i. e., the pixel coordinates for each point have to be computed. At first, each color image is converted into a gray-value image and then binarized by applying a threshold. Since the calibration pattern is designed so that the interesting points are black on a white background or vice versa, the threshold can easily be defined, e. g., by analyzing the histogram of the image, and has only to be changed for special illumination situations. A morphological 3×3 erosion operation removes very small pixel regions (noise). The contours of the remaining pixel regions are then extracted using the algorithm described in [Suz85] and an ellipse is fitted to each contour [Fit95]. The centers of the fitted ellipses are used as 2-D calibration points, which provides sub-pixel accuracy.

   Before performing ellipse-fitting, all contours not belonging to points on the calibration pattern have to be removed. It is assumed that the circles of the calibration pattern are the only circular structures in the image. Then, non-circular contours are detected and omitted from further processing by the following circularity criterion C for contours:

   $$C = \frac{(\text{circumference})^2}{\text{area}} = \frac{(2\pi r)^2}{\pi r^2} = \frac{4\pi^2 r^2}{\pi r^2} = 4\pi \approx 12.6\,. \tag{4.9}$$

   C is constant for circles with arbitrary radius r. Since the fraction circumference/area is minimal for circles, C is minimal for circles. For non-circular contours, C will be larger than 4π. Considering the fact that ellipses are only approximately circular, a structure is defined as non-circular if $C \geq 15$. In extreme cases, e. g., for extreme projective distortions, a larger threshold may be applied. Additionally, invalid contours are removed by defining a valid range for the contour area and number of contour points.

3. **Assigning 3-D world points to 2-D calibration points:** The basic idea of this step is to perform a 2-D projective mapping (homography) to simplify the assignment of 3-D world points to 2-D calibration points. It is assumed that a planar calibration pattern was used, i. e., without loss of generality $z = 0$ for all 3-D world points. Let $\boldsymbol{p}_i$ denote the i-th column of the projection matrix $\boldsymbol{P}$. The projection equation (3.10) can then be simplified:

$$\underline{\boldsymbol{q}} \sim \boldsymbol{P}\underline{\boldsymbol{w}} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4] \begin{pmatrix} w_x \\ w_y \\ 0 \\ 1 \end{pmatrix} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_4] \begin{pmatrix} w_x \\ w_y \\ 1 \end{pmatrix} = \boldsymbol{H}\underline{\boldsymbol{w}}' , \qquad (4.10)$$

where the homography $\boldsymbol{H} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_4] \in \mathrm{I\!R}^{3\times 3}$ defines the projective 2-D mapping from $\underline{\boldsymbol{w}}' = (w_x, w_y, 1)^{\mathrm{T}}$ to $\underline{\boldsymbol{q}} = (q_x, q_y, q_h)^{\mathrm{T}}$. Let $\overline{\boldsymbol{h}}_i$ be the i-th row of $\boldsymbol{H}$. Equation (4.10) can then be rewritten as

$$\begin{pmatrix} q_x \\ q_y \\ q_h \end{pmatrix} \sim \begin{pmatrix} \overline{\boldsymbol{h}}_1\underline{\boldsymbol{w}}' \\ \overline{\boldsymbol{h}}_2\underline{\boldsymbol{w}}' \\ \overline{\boldsymbol{h}}_3\underline{\boldsymbol{w}}' \end{pmatrix} \qquad (4.11)$$

and the transformation of $\underline{\boldsymbol{q}}$ to the Euclidean point $\boldsymbol{q}$ results in

$$\begin{pmatrix} q_x \\ q_y \end{pmatrix} = \frac{1}{q_h} \begin{pmatrix} q_x \\ q_y \end{pmatrix} = \frac{1}{\overline{\boldsymbol{h}}_3\underline{\boldsymbol{w}}'} \begin{pmatrix} \overline{\boldsymbol{h}}_1\underline{\boldsymbol{w}}' \\ \overline{\boldsymbol{h}}_2\underline{\boldsymbol{w}}' \end{pmatrix} , \qquad (4.12)$$

which can be rewritten as

$$\begin{bmatrix} \underline{\boldsymbol{w}}'^{\mathrm{T}} & \boldsymbol{0}_3^{\mathrm{T}} & -q_x\underline{\boldsymbol{w}}'^{\mathrm{T}} \\ \boldsymbol{0}_3^{\mathrm{T}} & \underline{\boldsymbol{w}}'^{\mathrm{T}} & -q_y\underline{\boldsymbol{w}}'^{\mathrm{T}} \end{bmatrix} \begin{pmatrix} \overline{\boldsymbol{h}}_1^{\mathrm{T}} \\ \overline{\boldsymbol{h}}_2^{\mathrm{T}} \\ \overline{\boldsymbol{h}}_3^{\mathrm{T}} \end{pmatrix} = \boldsymbol{0}_2 , \qquad (4.13)$$

where all elements of $\boldsymbol{0}_3 \in \mathrm{I\!R}^3$ and $\boldsymbol{0}_2 \in \mathrm{I\!R}^2$ are zero. At least four point correspondences are required to estimate $\boldsymbol{H}$, where not more than two points may lie on a line [Har03]. Figure 4.5 visualizes the method for determining the four point correspondences:

   (a) Asymmetric pattern: each corner is marked by a number of smaller circles (3, 4, 5, and 6). Small circles are identified by their area (the radius of the small circles is

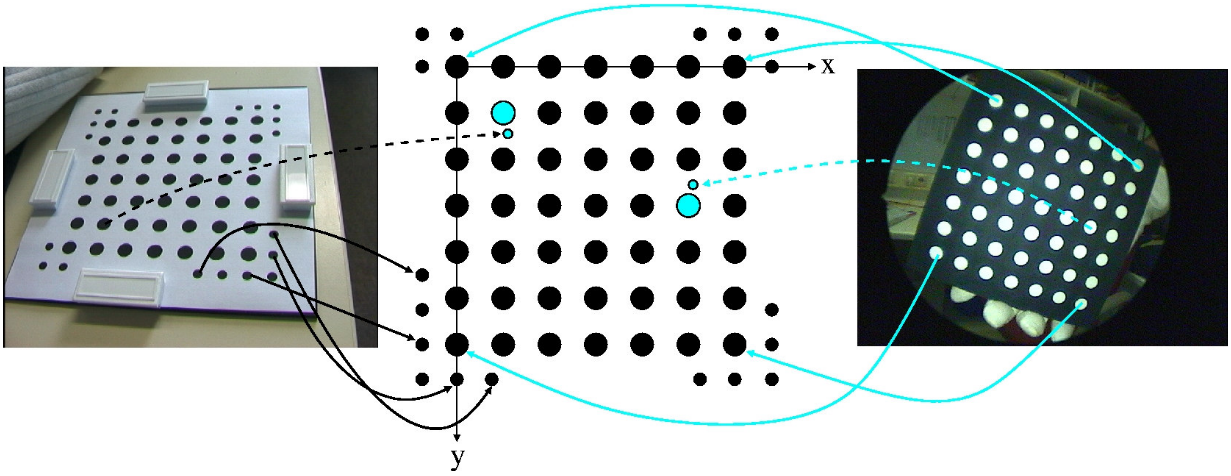


Figure 4.5: Assignment of 2-D calibration points to 3-D world points, shown for an asymmetric calibration pattern (left image) and a symmetric calibration pattern (right image). Firstly, a 2-D projective mapping (homography) from image coordinates to calibration pattern world coordinates is computed (assuming $z = 0$ for all world points) by selecting four point correspondences (solid lines). Then each calibration point of the captured image is transformed by the computed homography into the coordinate system of the calibration pattern, and the nearest world point on the calibration pattern is assigned to it (dashed lines).

defined as $3/5$ of the radius of the large circles, therefore the area is reduced by a factor of $9/25 = 0.36$). The number of circles identifies the corner. For the corners with 5 or 6 small circles, these circles are sufficient to determine the four projection points. For the corners with 3 and 4 small circles, the nearest larger circle has to be used as well. It is determined as the nearest larger circle to the center of mass of the smaller circles. If more than one corner is visible the four points are selected from the small circles of the visible corners to increase the accuracy of the projective mapping.

(b) Symmetric pattern: the four corner points are used. It is assumed that they can be detected in the image (which is not always true, especially not for large lens distortions and parallel axes of calibration pattern and image, and of course also not if the calibration pattern is only partially visible).

For each point one equation such as (4.13) is obtained and the four equations can be written in one matrix equation $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{0}_8$ with $\boldsymbol{x} = \left(\overline{\boldsymbol{h}}_1, \overline{\boldsymbol{h}}_2, \overline{\boldsymbol{h}}_3\right)^{\mathrm{T}}$ and $\boldsymbol{L} \in \mathbb{R}^{8\times 9}$. The solution $\boldsymbol{x}$ is the null-space of $\boldsymbol{L}$, which is obtained by a singular value decomposition of $\boldsymbol{L}$ (see Appendix B). If $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$, $\boldsymbol{x}$ is defined as the row vector in $\boldsymbol{V}^{\mathrm{T}}$ corresponding to the smallest singular value (for sorted SVDs this is the last row vector).

Finally, each 2-D pixel of the image is transformed by the computed homography $\boldsymbol{H}$, and the nearest calibration point is assigned as 3-D world point with $z = 0$ (cf. Figure 4.5, dashed lines).

4. **Estimation of an initial solution for the camera parameters [Zha00]:** Zhang uses the following equation for the projection of a homogeneous 3-D point $\underline{\boldsymbol{w}}$ into the image:

$$\underline{\boldsymbol{q}} \sim \widetilde{\boldsymbol{K}} \left[\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{t}}\right] \underline{\boldsymbol{w}} = \widetilde{\boldsymbol{K}} \left[\widetilde{\boldsymbol{r}}_{\mathrm{x}}, \widetilde{\boldsymbol{r}}_{\mathrm{y}}, \widetilde{\boldsymbol{r}}_{\mathrm{z}}, \widetilde{\boldsymbol{t}}\right] \begin{pmatrix} w_x \\ w_y \\ w_z \\ 1 \end{pmatrix} . \tag{4.14}$$

This is slightly different from the projection equation (3.10), page 37, which is used in this thesis. $\widetilde{\boldsymbol{K}}$ contains one more intrinsic camera parameter than $\boldsymbol{K}$: the image skew (element k_{12}). Both equations are equivalent when assuming no image skew and setting $\widetilde{\boldsymbol{R}} = \boldsymbol{R}^{\mathrm{T}}$ and $\widetilde{\boldsymbol{t}} = -\boldsymbol{R}^{\mathrm{T}}\boldsymbol{t}$. Assuming $w_z = 0$ for all world points and substituting $\sim$ by multiplication with an unknown scalar s yields

$$s\underline{\boldsymbol{q}} = \underbrace{\widetilde{\boldsymbol{K}} \left[\widetilde{\boldsymbol{r}}_{\mathrm{x}}, \widetilde{\boldsymbol{r}}_{\mathrm{y}}, \widetilde{\boldsymbol{t}}\right]}_{=:\boldsymbol{H}} \begin{pmatrix} w_x \\ w_y \\ 1 \end{pmatrix} = \boldsymbol{H}\underline{\boldsymbol{w}}' , \tag{4.15}$$

where the homography $\boldsymbol{H}$ relates the world point $\underline{\boldsymbol{w}}' = (w_x, w_y, 1)^{\mathrm{T}}$ to the calibration point $\underline{\boldsymbol{q}}$ (cf. equation (4.10)). Given an initial guess for $\boldsymbol{H}$, the back-projection error

$$\sum_i \|\boldsymbol{q}_i - \widehat{\boldsymbol{q}}_i\|^2 \quad \text{with} \quad \widehat{\underline{\boldsymbol{q}}}_i = \boldsymbol{H}\underline{\boldsymbol{w}}' \tag{4.16}$$

is minimized non-linearly using the Levenberg-Marquardt algorithm [Den83]. The initial guess is computed using the homography estimation of the previous step (equations (4.10) to (4.13)).

Now, the parameters $\widetilde{\boldsymbol{K}}$, $\widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{t}}$ are obtained by decomposing the computed homography $\boldsymbol{H}$ into $\widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{t}}$ using constraints on the intrinsic parameters. Let $\boldsymbol{h}_i$ denote the i-th column of $\boldsymbol{H}$. The following two constraints on the intrinsic parameters are defined:

$$\begin{aligned} \boldsymbol{h}_1{}^{\mathrm{T}}\widetilde{\boldsymbol{K}}^{-\mathrm{T}}\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_2 &= 0 , & (4.17) \\ \boldsymbol{h}_1{}^{\mathrm{T}}\widetilde{\boldsymbol{K}}^{-\mathrm{T}}\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_1 &= \boldsymbol{h}_2{}^{\mathrm{T}}\widetilde{\boldsymbol{K}}^{-\mathrm{T}}\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_2 , & (4.18) \end{aligned}$$

where

$$\widetilde{\boldsymbol{K}}^{-\mathrm{T}} = \left(\widetilde{\boldsymbol{K}}^{-1}\right)^{\mathrm{T}} = \left(\widetilde{\boldsymbol{K}}^{\mathrm{T}}\right)^{-1} . \tag{4.19}$$

The equations can be verified using the fact that $\widetilde{\boldsymbol{r}}_{\mathrm{x}}{}^{\mathrm{T}} \cdot \widetilde{\boldsymbol{r}}_{\mathrm{y}} = 0$ ($\widetilde{\boldsymbol{R}}$ is orthonormal) and $\boldsymbol{h}_1 = \widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{r}}_{\mathrm{x}}$ and $\boldsymbol{h}_2 = \widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{r}}_{\mathrm{y}}$.

Using

$$\widetilde{\boldsymbol{K}} = \begin{pmatrix} F_{\mathrm{x}} & \gamma & C_{\mathrm{x}} \\ 0 & F_{\mathrm{y}} & C_{\mathrm{y}} \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{K}}^{-1} = \begin{pmatrix} \frac{1}{F_{\mathrm{x}}} & -\frac{\gamma}{F_{\mathrm{x}}F_{\mathrm{y}}} & \frac{C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}}}{F_{\mathrm{x}}F_{\mathrm{y}}} \\ 0 & \frac{1}{F_{\mathrm{y}}} & -\frac{C_{\mathrm{y}}}{F_{\mathrm{y}}} \\ 0 & 0 & 1 \end{pmatrix} , \tag{4.20}$$

where γ is the image skew, a new matrix $\boldsymbol{B}$ is defined as

$$\begin{aligned} \boldsymbol{B} &= \widetilde{\boldsymbol{K}}^{-\mathrm{T}}\widetilde{\boldsymbol{K}}^{-1} \\ &= \begin{pmatrix} \frac{1}{F_{\mathrm{x}}{}^2} & -\frac{\gamma}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}} & \frac{C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}}}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}} \\ -\frac{\gamma}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}} & \frac{\gamma^2}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}{}^2} + \frac{1}{F_{\mathrm{y}}{}^2} & -\frac{\gamma(C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}})}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}{}^2} - \frac{C_{\mathrm{y}}}{F_{\mathrm{y}}{}^2} \\ \frac{C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}}}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}} & -\frac{\gamma(C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}})}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}{}^2} - \frac{C_{\mathrm{y}}}{F_{\mathrm{y}}{}^2} & \frac{(C_{\mathrm{y}}\gamma - C_{\mathrm{x}}F_{\mathrm{y}})^2}{F_{\mathrm{x}}{}^2F_{\mathrm{y}}{}^2} + \frac{C_{\mathrm{y}}{}^2}{F_{\mathrm{y}}{}^2} + 1 \end{pmatrix} . \end{aligned} \tag{4.21}$$

$\boldsymbol{B}$ is symmetric and can therefore be represented by a 6-D vector

$$\boldsymbol{b} = (b_{11}\ b_{12}\ b_{22}\ b_{13}\ b_{23}\ b_{33})^{\mathrm{T}} . \tag{4.22}$$

Let the i-th column of $\boldsymbol{H}$ be $\boldsymbol{h}_i = (h_{1i}, h_{2i}, h_{3i})^{\mathrm{T}}$. This allows defining

$$\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{B}\boldsymbol{h}_j = \boldsymbol{a}_{ij}^{\mathrm{T}}\boldsymbol{b} \tag{4.23}$$

with

$$\boldsymbol{a}_{ij}^{\mathrm{T}} = (h_{1i}h_{1j}, h_{1i}h_{2j} + h_{2i}h_{1j}, h_{2i}h_{2j}, h_{3i}h_{1j} + h_{1i}h_{3j}, h_{3i}h_{2j} + h_{2i}h_{3j}, h_{3i}h_{3j}) . \tag{4.24}$$

Now, the two constraints of equations (4.17) and (4.18) can be rewritten as two homogeneous equations in $\boldsymbol{b}$:

$$\begin{bmatrix} \boldsymbol{a}_{12}^{\mathrm{T}} \\ (\boldsymbol{a}_{11} - \boldsymbol{a}_{22})^{\mathrm{T}} \end{bmatrix} \boldsymbol{b} = \boldsymbol{0}_2 . \tag{4.25}$$

For n images, n equations can be stacked together, yielding

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{0}_{2n}\,, \tag{4.26}$$

where $\boldsymbol{A} \in \mathbb{R}^{2n\times 6}$. If the image skew is assumed to be zero ($\gamma = 0$), as done in this thesis, the additional equation $(0\ 1\ 0\ 0\ 0\ 0)\boldsymbol{b} = 0$ has to be added to equation (4.26). A solution for equation (4.26) is obtained by singular value decomposition: $\boldsymbol{b}$ is defined as the last row vector of $\boldsymbol{V}^{\mathrm{T}}$, corresponding to the smallest singular value, with $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$.

The matrix $\boldsymbol{B}$ is estimated up to an arbitrary scale factor $s \neq 0$, i. e., $\boldsymbol{B} = s\widetilde{\boldsymbol{K}}^{-\mathrm{T}}\widetilde{\boldsymbol{K}}^{-1}$. Nevertheless, the intrinsic parameters can be extracted uniquely (see [Zha00] with $b_{12} = \gamma = 0$):

$$\begin{aligned}
s &= b_{33} - (b_{13}^2 + C_{\mathrm{y}} b_{11} b_{23})/b_{11} && (4.27)\\
F_{\mathrm{x}} &= \sqrt{s/b_{11}} && (4.28)\\
F_{\mathrm{y}} &= \sqrt{s b_{11}/(b_{11} b_{22})} && (4.29)\\
C_{\mathrm{x}} &= b_{13} {F_{\mathrm{x}}}^2/s && (4.30)\\
C_{\mathrm{y}} &= -b_{11} b_{23}/(b_{11} b_{22})\,. && (4.31)
\end{aligned}$$

Equations (4.27) to (4.31) can be verified by substituting $b_{11}, b_{22}, b_{13}, b_{23}$, and b_{33} with the values shown in equation (4.21).

Once $\widetilde{\boldsymbol{K}}$ is known, the extrinsic camera parameters for each image are computed directly using equation (4.15):

$$\widetilde{\boldsymbol{r}}_{\mathrm{x}} = s\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_1, \;\; \widetilde{\boldsymbol{r}}_{\mathrm{y}} = s\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_2, \;\; \widetilde{\boldsymbol{r}}_{\mathrm{z}} = \widetilde{\boldsymbol{r}}_{\mathrm{x}} \times \widetilde{\boldsymbol{r}}_{\mathrm{y}}, \text{ and } \widetilde{\boldsymbol{t}} = s\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_3\,, \tag{4.32}$$

where × denotes the vector product and $s = 1/\|\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_1\| = 1/\|\widetilde{\boldsymbol{K}}^{-1}\boldsymbol{h}_2\|$. In general, the computed rotation matrix $\widetilde{\boldsymbol{R}}$ does not satisfy the properties of a rotation matrix because of noise in the data. An approximation that is best in the sense of the Frobenius norm can then be obtained by singular value decomposition (see Appendix B).

5. **Optimizing the solution non-linearly [Zha00]:** The solution of the last step was obtained by minimizing an algebraic distance which is not physically meaningful. Therefore, the solution can be refined by minimizing the squared back-projection error over N_{w} world points projected into N_{f} images with the following functional, which also introduces image

distortion:

$$\sum_{i=1}^{N_\mathrm{f}} \sum_{j=1}^{N_\mathrm{w}} \|\boldsymbol{q}_{i,j} - \mathrm{proj}(\boldsymbol{w}_j, \widetilde{\boldsymbol{K}}, \kappa_1, \kappa_2, p_1, p_2, \widetilde{\boldsymbol{R}}_i, \widetilde{\boldsymbol{t}}_i)\|^2 \tag{4.33}$$

where $\mathrm{proj}(\boldsymbol{w}_j, \widetilde{\boldsymbol{K}}, \kappa_1, \kappa_2, p_1, p_2, \widetilde{\boldsymbol{R}}_i, \widetilde{\boldsymbol{t}}_i)$ is the projection of world point $\boldsymbol{w}_j$ into image i, followed by computing the distorted point using equations (4.4) and (4.6), and $\boldsymbol{q}_{i,j}$ is the j-th 2-D calibration point in the i-th image. The distortion coefficients are initialized with zero. The optimization is done with the Levenberg-Marquardt implementation of `MINPACK` [Mor77].

The result of this optimization is the final estimation of the intrinsic and extrinsic camera parameters for each captured image.

4.2 Color Normalization

The illumination of a scene varies under real-world conditions. This leads to different color values for images of the same scene. Computer vision systems have to cope with this effect and some kind of illumination correction or color normalization is often used to increase the power of the system. The idea is to transform all pixels of a captured image in such a way that the color values of the resulting image are almost independent of the illumination variations. Therefore, illumination correction is also denoted as color normalization. Feature tracking [Fus99, Jin01, Grä03], object localization [Pau98], and object recognition [Fin98] are prominent examples where color normalization can be used to increase the power of computer vision algorithms.

For endoscopic image sequences illumination changes are even more problematic. The light source is located directly beneath the lens of the endoscope optics and therefore moves while the endoscope is moving. Additionally, another effect changes the colors of the image: bleeding due to tissue cuts with imbibition of the tissue with hemoglobin leads to a reddish coloring of the image. As a result, the identification of different tissue types is more difficult. Hence, the goal of color normalization for endoscopic images is to transform each image so that different tissue types can be separated more clearly.

The correction by normalization should lead to as natural an image as possible. Color calibration patterns like the MacBeth Color Checker [McC76] could be used to determine the sensor characteristics under normed illumination conditions. This would provide at least the same colors for images captured of the same scene with different types of cameras [Fin95]. However, it does not solve the problem of illumination changes and bleeding. Here, the *color cluster rotation*

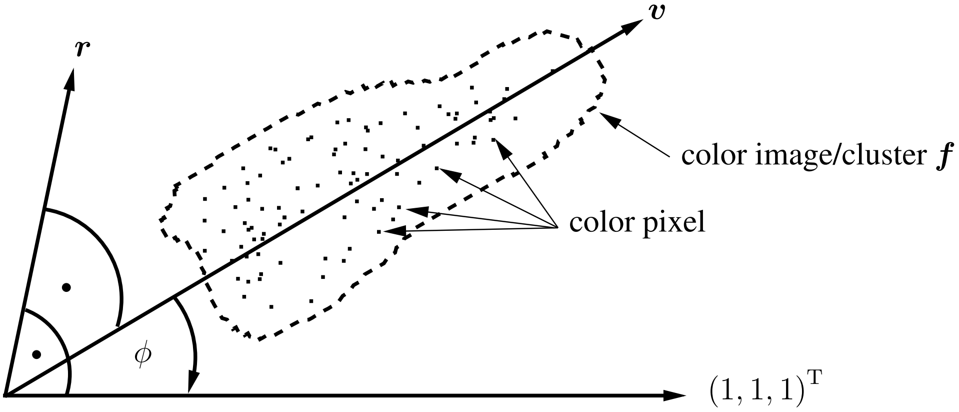


Figure 4.6: Color cluster rotation [Pau98]: The transformation is defined in such a way that the principal orientation $\boldsymbol{v}$ of the color image/cluster $\boldsymbol{f}$ is rotated onto the gray-axis $(1,1,1)^{\mathrm{T}}$ of the *RGB* color space. The rotation is defined by the rotation axis $\boldsymbol{r}$ and the rotation angle ϕ, where $\boldsymbol{r} \perp \boldsymbol{v}$ and $\boldsymbol{r} \perp (1,1,1)^{\mathrm{T}}$.

algorithm [Pau98] is used (see [Vog01a, Vog03a]). The transformation of the color pixels is done directly in the *RGB* color space provided by the camera. Since real-time image processing is required the conversion into another color space before applying the transformation (and requiring a re-transformation afterwards), e. g., done in [Oja84], is not appropriate.

The color cluster rotation algorithm normalizes the distribution of color pixels based solely on the image data. The normalization is performed by a transformation of each color pixel with a 3×3 rotation matrix $\boldsymbol{R}_{\mathrm{C}}$. The basic idea is to define the transformation so that the principal orientation of the color cluster corresponds to the gray-axis of the *RGB* color space (see Figure 4.6). Additionally, the mean value of the color vectors should lie on the gray-axis. The following three steps describe the algorithm in detail.

1. **Principal orientation of the color cluster.** Let $\boldsymbol{f}$ be a color image with N_{r} rows and N_{c} columns, where $\boldsymbol{f}(x,y)$ is the color value of the x-th column and y-th row. The center $\boldsymbol{\mu}$ of all color pixels is computed by

$$\boldsymbol{\mu} = \frac{1}{N_{\mathrm{r}}N_{\mathrm{c}}} \sum_{y=0}^{N_{\mathrm{r}}-1} \sum_{x=0}^{N_{\mathrm{c}}-1} \boldsymbol{f}(x,y)\,. \tag{4.34}$$

Let $\boldsymbol{v} = (v_x, v_y, v_z)^{\mathrm{T}}$ denote the eigenvector belonging to the largest eigenvalue of the covariance matrix

$$\boldsymbol{C} = \frac{1}{N_{\mathrm{r}}N_{\mathrm{c}}} \sum_{y=0}^{N_{\mathrm{r}}-1} \sum_{x=0}^{N_{\mathrm{c}}-1} (\boldsymbol{f}(x,y) - \boldsymbol{\mu})(\boldsymbol{f}(x,y) - \boldsymbol{\mu})^{\mathrm{T}}\,. \tag{4.35}$$

Then $\boldsymbol{v}$ defines the principal orientation of the cluster of color vectors of the image $\boldsymbol{f}$.

2. **Rotation matrix.** After rotation by $\boldsymbol{R}_{\mathrm{C}}$, the principal orientation $\boldsymbol{v}$ should be $\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$, i. e., $\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}} = \boldsymbol{R}_{\mathrm{C}}\boldsymbol{v}$. The rotation is characterized by an axis $\boldsymbol{r} \in \mathbb{R}^3$ and an angle ϕ. In order to obtain the desired rotation, $\boldsymbol{r}$ has to be orthogonal to $\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$ and $\boldsymbol{v}$:

$$\boldsymbol{r} = \frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}} \times \boldsymbol{v}\,. \tag{4.36}$$

ϕ is then defined as the angle between $\boldsymbol{v}$ and $\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$:

$$\phi = \arccos\left(\boldsymbol{v}^{\mathrm{T}} \cdot \frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}\right)\,. \tag{4.37}$$

The rotation matrix $\boldsymbol{R}_{\mathrm{C}}$ is computed from $\boldsymbol{r}$ and ϕ using Rodrigues' formula [Fau93]:

$$\boldsymbol{R}_{\mathrm{C}}(\boldsymbol{r},\phi) = \boldsymbol{I}_{3\times 3} + \sin\phi\ [\boldsymbol{r}]_{\times} + (1-\cos\phi)[\boldsymbol{r}]_{\times}^2 \tag{4.38}$$

where

$$[\boldsymbol{r}]_{\times} = \left[\begin{pmatrix} r_x \\ r_y \\ r_z \end{pmatrix}\right]_{\times} = \begin{pmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{pmatrix}\,. \tag{4.39}$$

3. **Transformation.** Each color value $\boldsymbol{f}(x,y)$, $y = 0,\ldots,N_{\mathrm{r}}-1$, $x = 0,\ldots,N_{\mathrm{c}}-1$, is transformed to $\boldsymbol{f}'(x,y)$ according to the following equation:

$$\boldsymbol{f}'(x,y) = \boldsymbol{R}_{\mathrm{C}}(\boldsymbol{r},\phi)\cdot(\boldsymbol{f}(x,y)-\boldsymbol{\mu}) + \boldsymbol{\mu}'\,, \tag{4.40}$$

where $\boldsymbol{r}$ and ϕ are given by equations (4.36) and (4.37) and $\boldsymbol{\mu}'$ is the center of the transformed color cluster. Either $\boldsymbol{\mu}'$ is fixed to a specific value to obtain the same mean brightness for all images, e. g., $\boldsymbol{\mu}' = (100,100,100)^{\mathrm{T}}$, or it is computed dependent on $\boldsymbol{f}$:

$$\boldsymbol{\mu}' = \frac{\|\boldsymbol{\mu}\|}{\cos\phi\sqrt{3}}(1,1,1)^{\mathrm{T}}\,. \tag{4.41}$$

This means $\boldsymbol{\mu}'$ is the projection of $\boldsymbol{\mu}$ onto $\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$.

As each color channel is represented by eight bits, i. e., possible values are 0 to 255, overflows above 255 and underflows below 0 are clipped to 255 and 0, respectively.

The experiments described in [Pau98] were done with single images. Applying color normalization to an image sequence requires the computation of $\boldsymbol{R}_{\mathrm{C}}$ for each image. With respect to the goal of real-time image enhancement, two acceleration possibilities concerning the computation of $\boldsymbol{R}_{\mathrm{C}}$ are proposed:

- Assuming that the illumination changes are continuous, the computation of $\boldsymbol{R}_{\mathrm{C}}$ is only performed for every k-th image. A suitable value for k has to be determined experimentally.
- Assuming that a good estimate of the principal direction is also possible with a subset of pixels, various subsets could be defined. A straight forward way is to select only every k-th pixel.

The computation of $\boldsymbol{R}_{\mathrm{C}}$ only requires about 29 % of the computation time of the whole algorithm (see Table 7.1, page 148). However, this is the only way to accelerate the algorithm's computation time. Once the rotation matrix is determined, the transformation has to be computed for each pixel without any possibilities for acceleration, apart from using a fast implementation, e. g., the Open Computer Vision Library (OpenCV) [Ope05].

4.3 Temporal Filtering

The aim of temporal filtering of an endoscopic image sequence is the reduction of degradations such as small flying particles or fast moving smoke. If these degradations are defined as temporal noise, i. e., if it is assumed that the degradations are only visible at a certain pixel position for a short period of time, a temporal color median filter is a good method to reduce this temporal noise in the image sequence. This assumption can be made because flying particles move very fast and the camera stands still because the surgeon needs a steady image during the whole period while performing the operation. It is especially during the cutting of tissue, where the camera is always kept as still as possible, that the aforementioned degradations appear.

There are two methods for median filtering of a color image:

1. Each color channel is filtered separately by a gray-value median filter. The pixels contained in the filter mask are sorted according to their gray-value.

2. An ordering criterion is defined for color pixels, for example,

$$\boldsymbol{f}(\boldsymbol{q}_1) < \boldsymbol{f}(\boldsymbol{q}_2) \Leftrightarrow \|\boldsymbol{f}(\boldsymbol{q}_1)\| < \|\boldsymbol{f}(\boldsymbol{q}_2)\| \,. \tag{4.42}$$

   The pixels contained in the filter mask are sorted according to the defined criterion. The median of the sorted sequence is used as the result. This kind of filter will be denoted as *vector median filter*.

The disadvantage of the vector median filter is the computation time due to the sorting process: for each pixel a set with the size of the filter mask has to be sorted; for each comparison of two color vectors three multiplications and two additions have to be computed. The disadvantage of the first method is that the resulting image can (and usually does) contain new color values. As the difference between the two methods on real images is small (in the experiments the mean value of the pixel difference was less than 1 gray-value, see Section 7.7, page 155), the first method is used which also enables the usage of optimized image processing libraries as described in the following paragraphs.

Temporal filtering as described above is easy to implement. However, no temporal filters can be found in currently available image processing libraries. Since endoscopic image enhancement should be performed in real-time, optimized image processing libraries such as the Intel Image Processing Library (IPL) or the Intel Integrated Performance Primitives (IPP, successor of IPL) [Int05a], which make use of the MMX registers of the CPU allow for the application of simple filters such as median or Gauß in a few milliseconds (cf. Table 7.6, page 153, and [Vog01b]). With MMX registers, mathematical operations like addition or multiplication can be computed for eight bytes simultaneously on a common SIMD (Single Instruction Multiple Data) CPU. This leads to a considerable speedup compared to a conventional implementation.

In the following, a simple but efficient technique is described, which allows to use spatial filters for the implementation of temporal filters. It was published in [Vog01b]. The basic idea is to re-order temporal data into a spatial data structure, filter the spatial data structure and finally extract the data for the result. There are two possibilities for the process of fusion (re-ordering), filtering and extraction. Figure 4.7 displays the first possibility, Figure 4.8 displays the second. Both figures show the case of a temporal filter of size 3, implemented with the spatial version of the filter.

Let $\boldsymbol{f}_1, \boldsymbol{f}_2$ and $\boldsymbol{f}_3$ be three consecutive color images. In the first case (Figure 4.7), all three

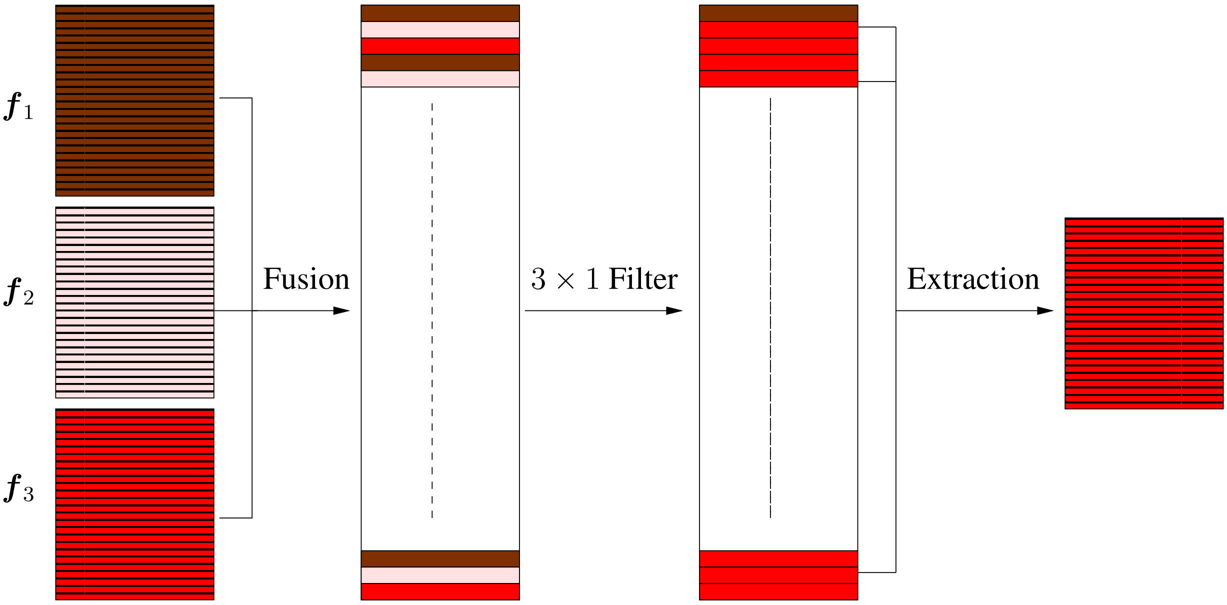


Figure 4.7: Implementation of a temporal filter by using a spatial filter: method 1. The resulting image is obtained by fusion of the original images f_1, f_2, and f_3 into one large image, filtering this image with a spatial filter, and extracting the corresponding rows. Displayed is a temporal filter of size 3, which uses a spatial filter of size 3×1.

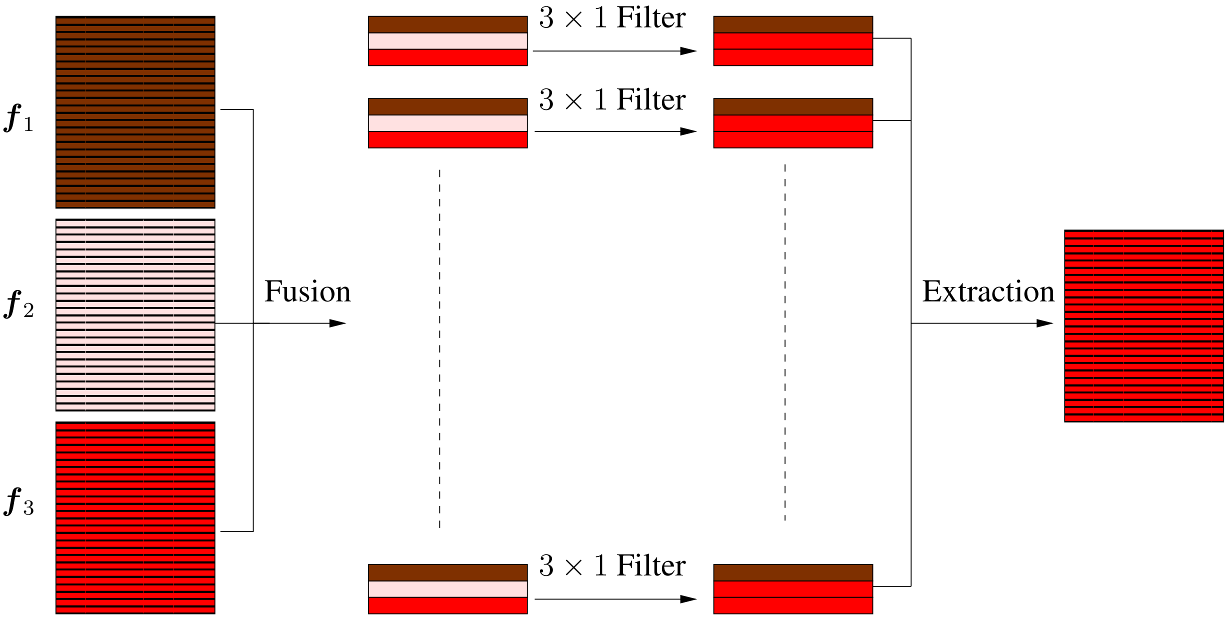


Figure 4.8: Implementation of a temporal filter by using a spatial filter: method 2. The resulting image is obtained by fusion of the original images f_1, f_2, and f_3 into N_r small images, filtering these N_r images, and extracting the corresponding rows (N_r is the number of rows of the original images). Displayed is a temporal filter of size 3, which uses a spatial filter of size 3×1.

images are fused into one large image $\boldsymbol{f}_\mathrm{L}$ with $3 \cdot N_\mathrm{r}$ rows:

$$\boldsymbol{f}_\mathrm{L}(x,y) = \begin{cases} \boldsymbol{f}_1(x, \lfloor y/3 \rfloor), & \text{if } y \bmod 3 = 0 \\ \boldsymbol{f}_2(x, \lfloor y/3 \rfloor), & \text{if } y \bmod 3 = 1 \\ \boldsymbol{f}_3(x, \lfloor y/3 \rfloor), & \text{if } y \bmod 3 = 2 \end{cases}, \quad y = 0, \ldots, 3 \cdot N_\mathrm{r} - 1, \;\; x = 0, \ldots, N_\mathrm{c} - 1\,. \tag{4.43}$$

This means the first row of $\boldsymbol{f}_1$, followed by the first row of $\boldsymbol{f}_2$ followed by the first row of $\boldsymbol{f}_3$, followed by the second row of $\boldsymbol{f}_1$, and so on, are fused into $\boldsymbol{f}_\mathrm{L}$. Then the 3×1 spatial filter is applied to $\boldsymbol{f}_\mathrm{L}$. Afterwards the rows $1, 4, 7, \ldots, 3 \cdot N_\mathrm{r} - 2$ are extracted from $\boldsymbol{f}_\mathrm{L}$, leading to the temporal filtered image. This fusion-filtering-extraction process can easily be extended to larger sizes of the temporal filter. The only restriction is that the temporal filter size has to be odd.

In the second case (Figure 4.8), the r-th rows of the three images are fused into N_r images $\boldsymbol{f}'_0, \ldots, \boldsymbol{f}'_{N_\mathrm{r}-1}$ of size $3 \times N_\mathrm{c}$, where

$$\boldsymbol{f}'_r(x,i) = \boldsymbol{f}_{i+1}(x,r), \;\; r = 0, \ldots, N_\mathrm{r} - 1, \;\; i = 0, 1, 2, \;\; x = 0, \ldots, N_\mathrm{c} - 1\,. \tag{4.44}$$

Each of these images is filtered separately by the 3×1 spatial filter. Afterwards the middle row of each image (the second row) is extracted, leading to the temporally filtered image. Again, the only restriction is that the filter size has to be odd.

The whole algorithm is fast since the temporal filtering is achieved by copying data and applying optimized spatial filters (see Table 7.1, page 148). It is therefore suitable for real-time temporal filtering during minimally invasive operations.

The computational cost of both methods for temporal filtering is the same (the same number of rows has to be filtered). If the borders of the images are excluded from filtering the second method is faster. The experiments in Section 7.2.3, page 152, were performed with a temporal color median filter implemented with the second method.

For rank-ordering filters both methods for temporal filtering can easily be extended to a temporal *and* spatial filter (3-D filter) of size $y \times x \times t$ (rows × columns × time). The only restriction is that t and y are odd. Instead of fusing one line out of each source image to a new image, y lines are used, leading to N_r images with $y \cdot t$ rows and N_c columns. These images are filtered by the $y \times x$ spatial filter and the middle row is extracted, leading to the time and space filtered result image. This process has a very large overhead because each source image row is copied into y images. Naturally, the implementation of temporal filtering based on spatial filtering, as described above, can also be used for all seperable filters (such as Gauß) in order to implement a 3-D filter.

The subjective impression of several surgeons was that *spatial* filters like color median, Gauß, and symmetric nearest neighbor [Reh98], only reduce the sharpness of the image but not the mentioned degradations (cf. [Vog01a, Sch02a]). Therefore, *spatial* filtering and 3-D filtering with $y, x > 1$ is no alternative to reduce these kinds of degradations.

4.4 Filtering the Region of Interest

The main reason why filtering is only applied in a region of interest (ROI) is the reduction of the computation time since a smaller amount of pixels has to be filtered. The speedup is proportional to the reduced size of the ROI. Additionally, the possibility of ROI-filtering allows the surgeon to define which methods he requires to be applied to the whole image (e. g., distortion correction) and which to be applied only to the ROI (e. g., temporal filtering). ROI-filtering is not difficult to implement and was therefore added to the real-time image enhancement system.

4.5 Image Geometry Transformations

The image geometry transformations described in this section are *zoom* and *orientation/rotation*. Nowadays it is easy to provide digital zoom and rotation of an image in real-time. Currently the resolution of an endoscopic image is 768×576 pixels (PAL). This restricts the range of zooming that is possible without reducing the subjectively perceived image quality. Nevertheless, it is to be expected that digital endoscopic cameras with higher resolutions will be employed in the future. Digital zooming will then be more useful than it is today and a system for real-time endoscopic image enhancement should provide this possibility. The developed system allows selecting the zoom factor either by subsequent small zoom operations or by defining a rectangular area inside the image. The larger of the two sides of the rectangle with respect to the corresponding image axis is then used to calculate the zoom factor and a translatory movement is performed in such a way that the upper left corner of the rectangle corresponds to the upper left corner of the image.

During the course of a minimally invasive operation, the operation site is frequently examined from various directions. So-called *side view* endoscopes are often used, e. g., in laparoscopic cholecystectomies, where the angle of the optical axis is changed with respect to the cylinder of the optics (see Figure 4.9). This allows looking *behind* objects although the entry point into the body of the patient is fixed (see Figure 4.10). In order to look behind an object from the left and from the right, the endoscope has to be rotated by $180°$ to benefit from the side view

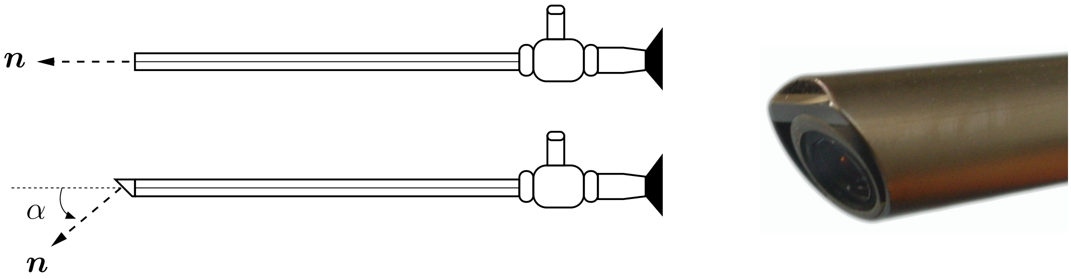


Figure 4.9: Side view endoscope (bottom left) compared to a conventional endoscope (top left): the angle α of the viewing direction $\boldsymbol{n}$ with respect to the cylinder of the endoscope is changed. For conventional endoscopes $\alpha = 0°$. Common angles for side view endoscopes are $\alpha = 30°$ and $\alpha = 45°$. The tip of a $30°$ side view endoscope is shown on the right.

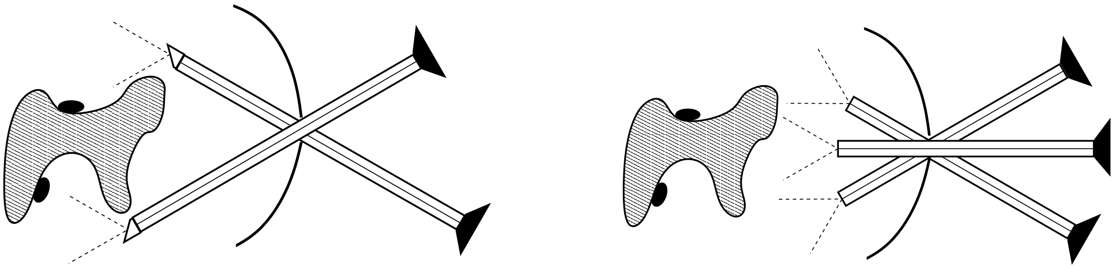


Figure 4.10: The advantage of a side view endoscope (left) compared to a conventional endoscope (right). A side view endoscope allows looking *behind* objects (e. g., to see the black ellipses) although the endoscope has a fixed entry point (trocar) into the body of the patient.

angle. However, this leads to images where the horizon is rotated by $180°$. If the surgeon wants to keep the horizon steady he has to rotate the camera head into the opposite direction. The setup of camera head and endoscope optics enables this rotation explicitely. Practically the horizon is kept steady by fixing the camera head in one hand while rotating only the endoscope optics with the other hand.

If the orientation of the camera is known, the rotation of the image to keep the horizon steady can be performed by the PC and the camera head can remain fixed onto the endoscope optics. Instead of applying a time-consuming structure-from-motion algorithm to compute the orientation of the camera (cf. Section 3.3.2 and [Kop01]), here a pose determination system as described in Sections 5.4 and 5.5 is used which provides the orientation of the camera in real-time. This also allows displaying the rotated image in real-time. The surgeon can define an arbitrary horizon, e. g., once at the beginning of the operation, and the displayed image is rotated automatically. A new horizon can be defined during the operation whenever necessary. The rotation is performed as follows: The y-axis of the horizon camera coordinate system defines the horizon (normal vector of the xz-plane). It is projected into the xy-plane of the current camera coordinate system. The current image is then rotated by the angle ϕ between the projection and $(0,1)^{\mathrm{T}}$. Let the horizon be given as rotation matrix $\boldsymbol{R} = [\boldsymbol{r}_{\mathrm{x}}, \boldsymbol{r}_{\mathrm{y}}, \boldsymbol{r}_{\mathrm{z}}]$ and the orientation of the current camera as $\boldsymbol{R}' = [\boldsymbol{r}_{\mathrm{x}}', \boldsymbol{r}_{\mathrm{y}}', \boldsymbol{r}_{\mathrm{z}}']$. Then the projection ${}^{\mathrm{p}}\boldsymbol{y}$ of $\boldsymbol{r}_{\mathrm{y}}$ into the xy-plane of the current camera coordinate system is given as

$$ {}^{\mathrm{p}}\boldsymbol{y} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{R}'^{\mathrm{T}} \boldsymbol{r}_{\mathrm{y}} \tag{4.45} $$

and

$$ \phi = -\arccos\left(\frac{1}{\|{}^{\mathrm{p}}\boldsymbol{y}\|} \cdot {}^{\mathrm{p}}\boldsymbol{y}^{\mathrm{T}} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) . \tag{4.46} $$

Finally, if $(1,0)\ {}^{\mathrm{p}}\boldsymbol{y} \geq 0$, the image has to be rotated by $-\phi$. A singularity is given if the vectors $\boldsymbol{r}_{\mathrm{z}}$ and $\boldsymbol{r}_{\mathrm{z}}'$ as well as $\boldsymbol{r}_{\mathrm{y}}$ and $\boldsymbol{r}_{\mathrm{y}}'$ are orthonormal: in this case the angle between $\boldsymbol{r}_{\mathrm{y}}$ and $\boldsymbol{r}_{\mathrm{y}}'$ cannot be changed by a rotation of the image. For use during minimally invasive operations the rotation is performed only if both angles, between $\boldsymbol{r}_{\mathrm{y}}$ and $\boldsymbol{r}_{\mathrm{y}}'$ and between $\boldsymbol{r}_{\mathrm{z}}$ and $\boldsymbol{r}_{\mathrm{z}}'$, are less than $80°$ or more than $100°$ ($90° \pm 10°$).

The reconstruction of light fields requires the camera head to be fixed onto the endoscope optics (see Chapter 5, page 83). Therefore, automatic image rotation according to a predefined horizon provides the correct image for the surgeon while allowing the reconstruction of light fields.

4.6 Image Enhancement by Light Fields

The additional information provided by a light field can be used to reduce or even remove arbitrary degradations in images. With respect to endoscopy, relevant degradations are highlights, small flying particles, smoke, and blood drops or other soilings on the camera lens. The first prerequisite is the existence of a static light field of a scene. The second prerequisite is that the degradation does not remain at the same position with respect to the scene while the camera moves. The third prerequisite is the ability of detecting the degradation(s) that should be removed. Then, degradations in the images rendered from the light field, in the images used to reconstruct the light field, and in new images of the scene can be reduced or removed. The distorted pixels correspond to surface points of the scene. Since the degradation moves it is very likely that these surface points were seen without degradations from another point of view. With the help of the light field and the information about the location of the distorted pixels in each image, the camera positions from which the surface points were seen without degradation can be

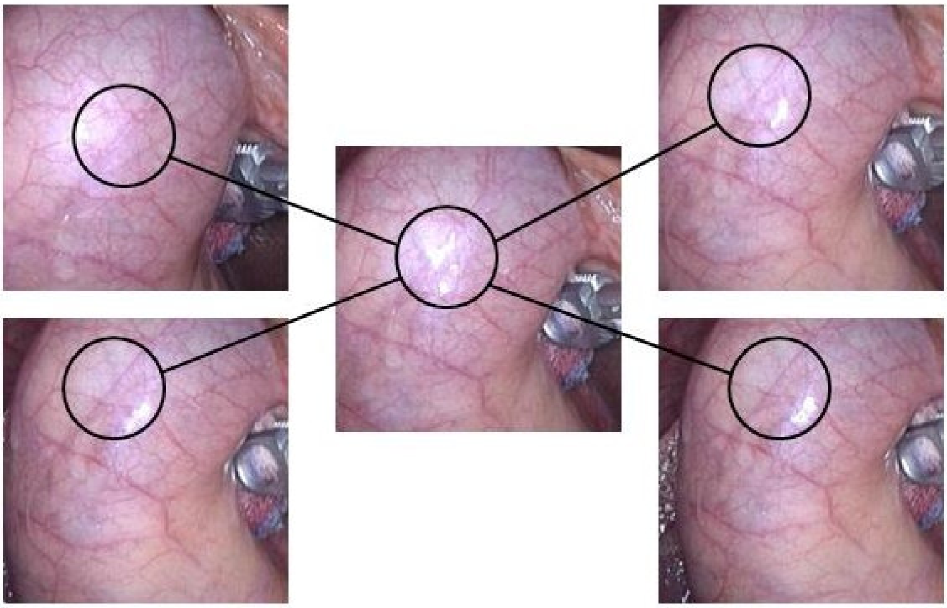


Figure 4.11: Degradation reduction by light fields: the degradation (highlight) of the image in the middle is substituted by color values from the other four images which do not show the degradation at this position. The images show a gall bladder.

determined and used to substitute the distorted pixels by “real” color values. Figure 4.11 shows the idea of the algorithm which was published in [Vog02b].

The second prerequisite is fulfilled for the mentioned degradations. The first is fulfilled by the reconstruction of a light field. In order to fulfill the third prerequisite a degradation detection algorithm has to be developed. An alternative exists only for blood drops and soilings on the camera lens: these degradations could be marked once by hand since they do not change their position (in the image) over time.

The algorithm is now described exemplarily for one considerable degradation: highlights. When color images of natural scenes are captured and displayed, highlights due to specular reflection may considerably incommode the observer. This is particularly the case when medical images are recorded and humid tissue is subject to inspection. The problem even increases for endoscopic images when light source and viewing direction are almost identical; thereby, surfaces orthogonal to the viewing direction are often over-imposed to such an extent, that the physicians can only guess the tissue at that position. Since the light source is located directly beneath the camera lens, the highlights move when the endoscope is moved. It is assumed that a light field was reconstructed from a number of images where at least in some of them highlights were visible.

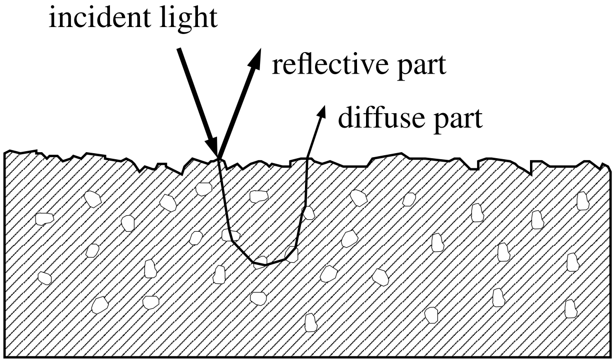


Figure 4.12: The di-chromatic reflectance model: partly the incident light is reflected directly, partly it passes through the surface and leaves the object without a specific direction (diffuse).

First, pixels distorted by highlights have to be detected. Methods for highlight detection are described in Section 4.6.1. Section 4.6.2 then describes the substitution of the detected pixels using a light field.

4.6.1 Highlight Detection

This section contains three algorithms for highlight detection. The first two are based on the di-chromatic reflectance model [Sha85]. Although human tissue does not fit the model of di-electric inhomogeneous material, algorithms based on it have been applied to detect (and remove) highlights for biological material [Pal99, Stö00]. According to the di-chromatic reflectance model, the specular distribution of the reflecting light Y is composed of a specular part L_S and a diffuse part L_D (cf. Figure 4.12):

$$\begin{aligned} Y(\boldsymbol{\theta},\lambda) &= c_\mathrm{S}(\boldsymbol{\theta})\cdot L_\mathrm{S}(\lambda) + c_\mathrm{D}(\boldsymbol{\theta})\cdot L_\mathrm{D}(\lambda) & (4.47)\\ L_\mathrm{S}(\lambda) &= P_\mathrm{S}(\lambda)\cdot E(\lambda) & (4.48)\\ L_\mathrm{D}(\lambda) &= P_\mathrm{D}(\lambda)\cdot E(\lambda) & (4.49)\end{aligned}$$

The two color components L_S and L_D depend solely on the wavelength λ of the incident light. $P_\mathrm{S}(\lambda)$ and $P_\mathrm{D}(\lambda)$ are reflectance properties, $E(\lambda)$ denotes the spectrum of the light source. The weight factors c_S and c_D depend on the capture properties $\boldsymbol{\theta}$: the spatial relation between camera/eye, light source, and object. A light ray hitting the object is partly reflected on the surface. The nature of the object determines the diffuse reflection property of the object and consequently its color.

Color gradients [Gev00]: Based on the *RGB* values, two new color spaces are defined in such a way that highlight edges are only visible in one of these color spaces [Gev99]. The properties of the color spaces are derived and verified based on the di-chromatic reflectance model. The edges of highlights can then be easily separated from normal (object) edges. The disadvantage of this approach is that only edges are detected. Some post-processing, e. g., region filling, is necessary to detect the whole region of the highlight.

Calculation of the color of the light source [Pal99]: Since $Y(\boldsymbol{\theta}, \lambda)$ is a linear combination of L_S and L_D, these two variables define a one-dimensional subspace (line) in the normalized *RG* color space. This subspace describes possible observations that are consistent with the model. $P_\mathrm{S}(\lambda)$ is assumed to be constant. Consequently the spectrum of the light source and the spectrum of the specular reflection is the same. For objects with different diffuse reflections parts, viewed under the same illumination conditions, the subspaces for the objects intersect due to the common variable L_S. Subspaces of different objects allow the computation of an intersection point which enables determining the color of the light source. Based on the assumption that the color of the light source is the same as the color of the reflection, highlights can be detected (and removed).

Thresholds in *HSV* color space [Vog02a, Vog02b, Vog02c]: Assuming that no over-imposure is present in the images, highlights can be detected in the *HSV* color space by thresholds on the saturation *S* and value *V*. The obtained highlight mask can be dilated with a 3×3 mask, possibly several times, to obtain closed highlight regions. This would also detect other white colored anatomical structures, but as such structures do not exist, this is no problem for endoscopic images.

4.6.2 Highlight Substitution

The idea for substituting highlight degradations with the help of a light field is straight forward: the confidence value for highlight pixels is set to zero. Then, light rays corresponding to highlight pixels are not used during the rendering of new views of the scene. An example of highlight substitution is illustrated in Figure 4.13. In order to substitute highlights in the images that were used to reconstruct the light field, or in new images of the scene of which the camera parameters are known (e. g., by a pose determination system), only highlight regions are rendered from the light field.

Note that this substitution is not restricted to highlights. All detectable degradations, e. g., those mentioned at the beginning of Section 4.6, page 77, can be substituted applying this method

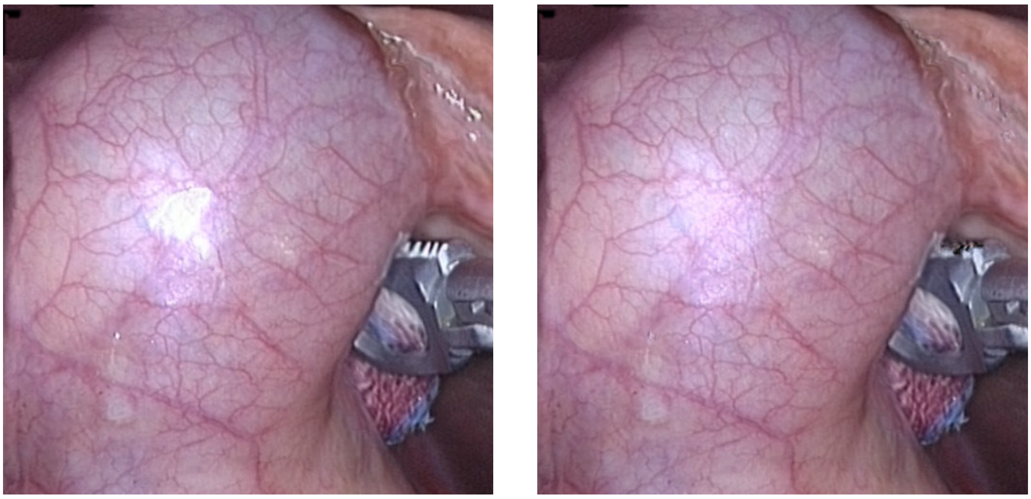


Figure 4.13: Example of highlight substitution based on a light field: the original image rendered from the light field (left) and the same image with substituted highlights are shown.

in such a way.

4.7 Summary

This chapter described a system for real-time endoscopic image enhancement. Distortion correction based on camera calibration with a calibration pattern is applied to correct distortions that are mainly due to lenses with small focal length (Section 4.1). The aim of color normalization is to provide illumination independent images in which different tissue types can also be separated in difficult situations, e. g., if the tissue is covered by blood (Section 4.2). During the cutting of tissue small flying particles are generated. These degradations are reduced by temporal color median filtering (Section 4.3). Image geometry transformations allow zooming and rotating the image (Section 4.5). Especially for side view endoscopes the rotation of an image according to a predefined horizon allows keeping the horizon steady for almost arbitrary movements of the endoscope. For the first three methods a region of interest can be defined to reduce computation time (Section 4.4). All algorithms can be computed in real-time, i. e., with ≥ 25 frames per second, even without defining a region of interest. Finally, a powerful method for substituting arbitrary degradations was described (Section 4.6). Three prerequisites are necessary to apply this technique: a static light field of a scene has to exist, the degradation does not remain at the same position with respect to the scene while the camera moves, and the degradation can be detected in the image.

Apart from image degradations, two other problems occur in endoscopic surgery: loss of stereoscopic depth perception and limited vision. The following chapter deals with these problems.

Chapter 5

Light Fields of the Operation Area

Loss of stereoscopic depth perception and *limited vision* are two of the problems that arise in endoscopic surgery. The method proposed in this thesis to reduce these two problems is to support the surgeon by reconstructing a light field of the operation site during the operation. The light field then allows the operation site to be viewed in 3-D and to extend the visible field by virtually decreasing the focal length or moving the endoscope backwards.

It has been shown that a light field can be reconstructed based only on the video images by structure-from-motion techniques (see Section 3.3.2, page 50). Some modifications of the algorithm of Section 3.3.2 are necessary in order to apply it to *endoscopic* image sequences. However, the quality of the reconstruction result depends mainly on the accuracy of 2-D point tracking, which is prone to errors for endoscopic image sequences due to the bad image quality. Furthermore, the algorithm is very time-consuming. Therefore, two new methods for light field reconstruction are presented. What both methods have in common is that the endoscope's pose is determined by additional apparatus in real-time, either by a robot arm or by an optical tracking system. The new methods have two main advantages: they reduce the necessary computation time and yield reliable extrinsic camera parameters independent of the image quality. This chapter describes all three methods for light field reconstruction of the operation site in endoscopic surgery:

- using structure-from-motion techniques in Section 5.3,
- using the robot arm AESOP in Section 5.4, and
- using the optical tracking system *smARTtrack1* in Section 5.5.

The main difference between the three approaches for light field reconstruction is the method for computing the extrinsic camera parameters. Therefore, Sections 5.3 to 5.5 focus on the computation of these parameters, whereas the remaining sections describe methods common to all three algorithms. Section 5.1 addresses preprocessing and Section 5.2 discusses the computation of intrinsic camera parameters. Since the visualization of light fields can be improved if information about the scene geometry is available, Section 5.6 deals with the *computation of scene geometry* in terms of 3-D surface points and their representation as depth maps for usage in DC light fields. The employed visualization techniques are summarized in Section 5.7. Finally, the three approaches are compared in Section 5.8.

The state of the art of reconstructing *dynamic* light fields is to reconstruct several static light fields for points in time where the scene is known to be static (cf. Section 3.4, page 53). This corresponds well to the successive steps of a minimally invasive operation, e. g., those of a cholecystectomy (cf. Section 2.2, page 15). This is also the kind of dynamics that the physicians are interested in. Therefore, dynamic light fields are generated in this way, i. e., by reconstructing several static light fields with one of the techniques described in this chapter. It is thus possible to view the temporal changes during an operation in 3-D. Additionally, past operation steps are still available and can be viewed.

5.1 Preprocessing

The developed system for acquiring images during minimally invasive operations has already been presented at the beginning of Chapter 4. This system is also employed for light field reconstruction. Three preprocessing steps are performed directly after grabbing the image: de-interlacing, distortion correction, and cropping.

5.1.1 De-Interlacing

The S-Video output of the endoscope camera provides 25 interlaced PAL color images of size 768×576 pixels (columns $\times$ rows) per second. This means that actually 50 half images are captured, where each image contains only half the number of rows (size 768×288 pixels). Two half images $\boldsymbol{f}_{\mathrm{h1}}$ and $\boldsymbol{f}_{\mathrm{h2}}$ together result in one interlaced image $\boldsymbol{f}$:

$$\boldsymbol{f}(x,y) = \begin{cases} \boldsymbol{f}_{\mathrm{h1}}(x,k), & \text{with } k = y/2 \text{ if } y \text{ is even} \\ \boldsymbol{f}_{\mathrm{h2}}(x,k), & \text{with } k = \lfloor y/2 \rfloor \text{ if } y \text{ is odd} \end{cases}, \quad y = 0,\ldots,575, \;\; x = 0,\ldots,767\,, \tag{5.1}$$

where $(x,y)^{\mathrm{T}}$ is the pixel in row y and column x. Since the two half images for each interlaced image are captured at different points in time, inconsistencies depending on the speed of the camera, object movement, and shutter time occur, i. e., consecutive rows may be slightly shifted. Images without interlacing artefacts are only obtained when a static scene is captured with fixed camera pose. When viewing 25 interlaced images per second, the inconsistencies are generally not noticed by a human observer. However, for the reconstruction of light fields it is assumed that each pixel of a captured image corresponds to *one* camera position. Two types of de-interlacing techniques can be applied to fulfill the requirement. In both cases only one of the two half images is used for further processing. Without loss of generality it is assumed that the second half image $\boldsymbol{f}_{\mathrm{h}2}$ is further processed:

- **Subsampling:** The capturing of a half image can also be interpreted as subsampling in the vertical direction since every other row is omitted. Consequently a correct image can be obtained by also subsampling the horizontal direction, leading to an image of size 384×288 pixels. This is achieved by convolving the image with a Gaussian filter with $\sigma = 1/\sqrt{2}$ and then using only every other pixel of each row. The advantage of this method is that the number of pixels for further processing is reduced to $1/4$. The disadvantage is that not all available information is maintained since the resolution in horizontal direction could be two times larger.

- **Interpolation:** The omitted rows are filled by linear interpolation, i. e.,

$$\boldsymbol{f}(x,y) = \begin{cases} \boldsymbol{f}_{\mathrm{h}2}(x,0), & \text{if } y = 0 \\ \left(\boldsymbol{f}_{\mathrm{h}2}(x,k-1) + \boldsymbol{f}_{\mathrm{h}2}(x,k)\right)/2, & \text{with } k = y/2 \text{ if } y \text{ is even and } y > 0 \\ \boldsymbol{f}_{\mathrm{h}2}(x,k), & \text{with } k = \lfloor y/2 \rfloor \text{ if } y \text{ is odd} \end{cases}, \tag{5.2}$$

  for $y = 0, \ldots, 575$ and $x = 0, \ldots, 767$. The first row ($y = 0$) is copied from the first row of $\boldsymbol{f}_{\mathrm{h}2}$ since an interpolation cannot be performed in this case. The intermediate rows are interpolated from the corresponding rows of $\boldsymbol{f}_{\mathrm{h}2}$ before and after the current row. The resulting image has the full horizontal resolution but every other row contains interpolated pixels. The *linear interpolation* method was chosen because it is simple and fast. Naturally, more complex techniques like *cubic splines* could also be employed, but they increase computation time. A survey of interpolation methods including an analysis of their computation complexity is given in [Leh99].

Note that the use of a robot arm for endoscope positioning (cf. Section 5.4, page 94) offers the possibility of increasing image quality, namely resolution: de-interlacing can be avoided when the robot is halted during the capturing of each image and a static scene is assumed. The disadvantage is then that it takes approximately two seconds to capture one image in contrast to the conventional 40 msec per image.

5.1.2 Distortion Correction

In order to determine the intrinsic camera parameters that are necessary for distortion correction, namely, $\boldsymbol{K}, \kappa_1, \kappa_2, p_1$, and p_2, the camera calibration algorithm of Section 4.1.2 is employed. Two calibration patterns are used (cf. Figure 4.4, page 61): a manufactured symmetric 7×7 pattern of white circles on black background and a more sophisticated asymmetric 7×7 pattern of black circles on white background that can simply be printed. The advantage of the latter is that it is scalable, i. e., it can be printed in different sizes, and in contrast to the symmetric pattern arbitrary rotations can be handled even when the pattern is only partly visible.

Ten images are captured with the endoscope from different directions with different orientations. The calibration pattern should be fully visible and cover the whole image. This ensures that the maximum amount of calibration points can be detected and that these points cover the whole image. The latter is important for estimating the distortion parameters correctly. Let N_{w} world points $\boldsymbol{w}_j, 1 \leq j \leq N_{\mathrm{w}}$, of the calibration pattern be visible in N_{f} captured images. Since the algorithm minimizes ϵ_{BPE}, i. e., the error of back-projecting all world points into each image (cf. equation (3.36), page 51), the mean back-projection error

$$\overline{\epsilon}_{\mathrm{BPE}} = \frac{1}{N_{\mathrm{f}} N_{\mathrm{w}}} \sum_{i=1}^{N_{\mathrm{f}}} \sum_{j=1}^{N_{\mathrm{w}}} \|\boldsymbol{q}_{i,j} - \widehat{\boldsymbol{q}}_{i,j}\| \tag{5.3}$$

is a measure for the accuracy of the estimated parameters, where $\boldsymbol{q}_{i,j}$ is the j-th calibration point of the i-th image and $\widehat{\boldsymbol{q}}_{i,j}$ is obtained by projecting the world point $\boldsymbol{w}_j$ into image i by using the pinhole camera model and the estimated intrinsic parameters. This error is specified in pixels and indicates the mean distance between estimated model and real data, i. e., the distance between projected world points and detected 2-D calibration points.

The theory of light fields is based on the pinhole camera model for perspective projection of a world point into the camera image. Since image distortion is not modeled, each captured image has to be undistorted before it can be used for light field reconstruction. This is performed by applying the algorithm presented in Section 4.1 using the calibrated intrinsic camera parameters.

5.1.3 Cropping

Algorithms and software for the visualization (rendering) of light fields are not developed in this thesis.

The software that is currently used for light field rendering exhaustively uses graphics hardware [Vog05a]. Therefore, it works best when the size of the employed images is quadratic and to the power of two. Thus, a quadratic image with side length $2^9 = 512$ pixels is generated by cropping. Due to the black border of endoscopic images (cf. Figure 5.1a), only very little usable information is lost by cropping the image[1]. It is assumed that the images are de-interlaced with the interpolation method. The cropped image is defined by the image coordinates of the new top-left and bottom-right pixel. In this case the two coordinates are $(128, 32)^{\mathrm{T}}$ and $(639, 543)^{\mathrm{T}}$. This means two times 128 columns and 32 rows of the original image are not used.

After cropping the image, the intrinsic parameters have to be transformed. Focal length and distortion parameters do not change when an image is cropped but the principal point $(C_{\mathrm{x}}, C_{\mathrm{y}})^{\mathrm{T}}$ does. The new principal point $(C_{\mathrm{x}}', C_{\mathrm{y}}')^{\mathrm{T}}$ is computed by

$$\begin{pmatrix} C_{\mathrm{x}}' \\ C_{\mathrm{y}}' \end{pmatrix} = \begin{pmatrix} C_{\mathrm{x}} \\ C_{\mathrm{y}} \end{pmatrix} - \begin{pmatrix} 128 \\ 32 \end{pmatrix} . \tag{5.4}$$

The cropping of images which are de-interlaced with the subsampling method can be derived analogously.

After performing the preprocessing steps described in the last three sections (Sections 5.1.1 to 5.1.3), the intrinsic camera parameters are known, the images are de-interlaced and undistorted, and the size of the images is 512×512 pixels. Figure 5.1 illustrates the single steps.

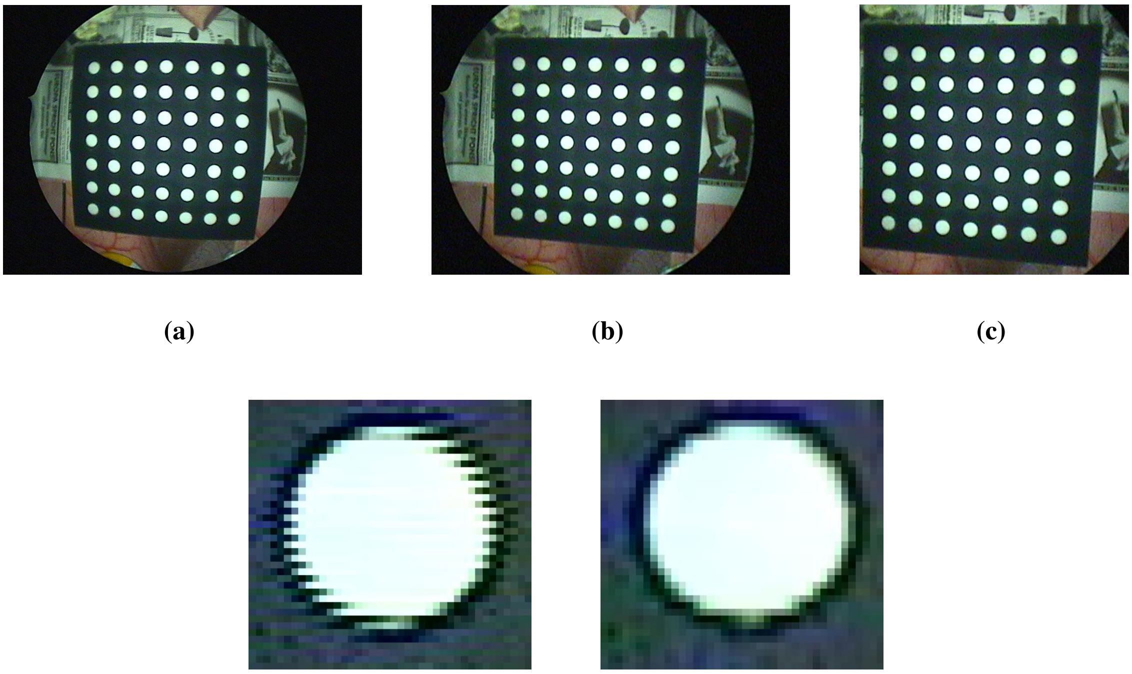
5.2 Intrinsic Parameters

The following assumption is made for all three light field reconstruction approaches: *the intrinsic camera parameters do not change during an endoscopic surgery.*

The cameras that are employed in laparoscopic and thoracoscopic operations do not provide zooming, i. e., the focal length cannot be changed. The remaining intrinsic camera parameters, namely, principal point and distortion parameters, are changed when the camera head is rotated with respect to the endoscope optics. This rotation occurs when side view optics that allow

---

[1] Resampling is not used here to achieve quadratic image size, because resampling a PAL image of size 768×576 pixels to a quadratic 512×512 pixels image changes the pixel aspect ratio dx/dy. The resulting image is compressed in the x-direction which makes it difficult to use for light field rendering.



(a) (b) (c)

(d) (e)

Figure 5.1: Illustration of preprocessing steps for light field reconstruction. The scene contains a calibration pattern which was captured with a 5 mm Storz side view endoscope. After calibrating the endoscope the original image (a) is de-interlaced and undistorted (b) and finally cropped (c). The size of the final image is 512×512 pixels. The two images of the bottom row show the process of de-interlacing in more detail: a magnified region of the original image (d) and the de-interlaced counterpart (e).

looking behind objects are used: the conventional way of keeping the horizon steady is to fix the camera head in one hand while rotating the endoscope with the other hand (see Figure 5.2 and Section 4.5, page 75). The assumption made here means that the surgeon *must not* rotate the camera head. The following observations justify this restriction:

- Prohibiting the described rotation is no drawback for the two methods based on a pose determination system, because the algorithm of Section 4.5, page 75, can be applied in real-time to keep the horizon steady. This method has the additional advantage that only one hand is necessary, i. e., the surgeon could theoretically move the camera himself while manipulating with a surgical instrument with his other hand. This is not possible when a conventional video-endoscopic system is employed.
- If the surgeon were allowed to rotate the camera head, it would not be possible to determine the correct camera pose by one of the pose determination systems employed here. In order

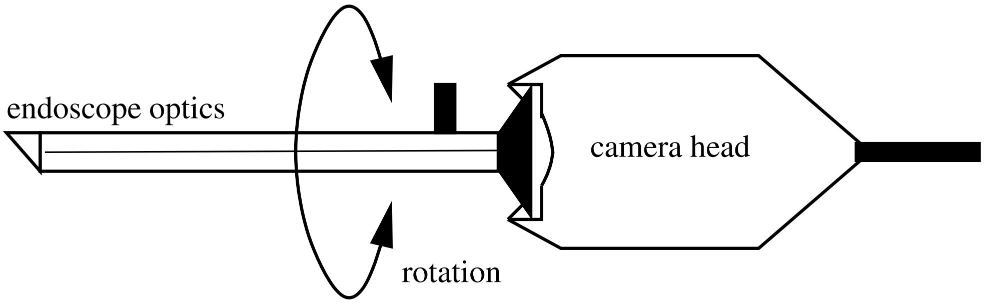


Figure 5.2: The setup of camera head and endoscope optics allows rotating the endoscope optics with respect to the camera head. This is exploited by the surgeon when side view optics are employed that allow looking behind objects. Then the horizon is kept steady by fixing the camera head in one hand while rotating only the endoscope optics with the other hand.

to obtain the correct *camera pose*, the pose of *camera head* and *endoscope optics* would have to be determined, which is impossible when using AESOP and an unsolved problem for optical tracking: the shape of the camera head and the fact that it is wrapped in a sterile foil complicate the attachment of a target, which is necessary for optical tracking.

- For the reconstruction of light fields based on point correspondences the assumption of constant intrinsic camera parameters is *theoretically* not necessary. However, in practice it has to be made in order to achieve usable reconstruction results (cf. Section 5.3, page 90).
- The assumption allows the intrinsic camera parameters to be determined once at the beginning of an operation by a camera calibration algorithm (cf. Section 4.1.2, page 61). The whole calibration process including image acquisition can be accomplished in approximately one minute.
- It is reasonable to assume that necessary distortion correction of the images (cf. Section 5.1.2) is more accurate when the intrinsic camera parameters only have to be determined once by camera calibration rather than estimated for each image anew.

Formally, the assumption of constant intrinsic parameters can be specified as follows: let $\boldsymbol{K}_i$ be the calibration matrix of the i-th image and $\kappa_{1,i}, \kappa_{2,i}, p_{1,i}$, and $p_{2,i}$ the corresponding distortion parameters. Then

$$\boldsymbol{K}_i = \boldsymbol{K}, \quad \begin{pmatrix} \kappa_{1,i} \\ \kappa_{2,i} \end{pmatrix} = \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix}, \quad \begin{pmatrix} p_{1,i} \\ p_{2,i} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \quad \forall i, \tag{5.5}$$

where $\boldsymbol{K}, \kappa_1, \kappa_2, p_1$, and p_2 are the constant intrinsic camera parameters.

5.3 Extrinsic Parameters from Point Correspondences

Based only on the video sequence, the reconstruction of a light field is possible by applying structure-from-motion techniques, where the extrinsic *and* intrinsic camera parameters are estimated. Since image distortion is not modeled by this approach, only the focal lengths and the principal point are estimated.

The algorithm of Section 3.3.2, page 50, is applied. Some modifications are necessary as the application of the algorithm to endoscopic image sequences is very difficult and only possible if certain prerequisites are fulfilled for the most part:

- no movement inside the scene,
- smooth camera movement during recording of the image sequence,
- structure in the scene for point tracking, and
- good illumination conditions.

The first modification concerns the number of parameters that have to be estimated by the algorithm. It is reasonable to restrict the estimation to the extrinsic camera parameters since the mentioned prerequisites cannot be fulfilled completely during endoscopic surgery: small movements occur due to breathing of the patient and heartbeat; the illumination conditions are poor and also not static, since the light source is located at the tip of the endoscope and moves together with the endoscope. Instead of ten parameters of the complete pinhole camera model only the six extrinsic parameters have to be estimated, which allows obtaining useful results with respect to the difficult conditions during endoscopic surgery.

The second modification concerns the extraction of 2-D point correspondences by point tracking. Here, a recently published approach is employed [Zin04]. It extends the originally used Shi-Tomasi-Kanade point tracker [Tom91, Shi94]. Since [Zin04] will also be used for computing scene geometry with pose determination systems, it is described in more detail.

Instead of employing the green-channel of the color image for point tracking as done in the algorithm of Section 3.3.2, page 50, the gray-value image corresponding to the captured color image is used. At first, points that are well suited for tracking have to be detected. Rather than selecting single points, feature windows around each point are selected. The selection is still based on the eigenvalues of the structure matrix (cf. equation (3.35), page 50), but using a smaller window size for feature detection than for feature tracking is possible. Interesting points, e. g., corners, may be located at the edge of the feature window, which could reduce tracking performance. By using smaller windows for detection than for tracking, the interesting point will be located well inside the larger tracking window even if it lies at the border of the detection window. In general, not all detected points can be tracked from one image to the next, e. g., points vanish due to the camera movement. Since it is often desired to keep the number of tracked points approximately constant, new points have to be selected from time to time.

The basic principle of the Shi-Tomasi-Kanade tracker is to iteratively minimize the sum of squared differences (SSD) of the intensities of the tracking windows with a gradient descent method. Points are tracked in *consecutive* images. Occlusions and false correspondences are detected by measuring the dissimilarity of a tracking window between the *first* and the *current* image. Affine distortions of the tracking window are taken into account by estimating its affine transformation before computing the SSD. The corresponding point is discarded if the SSD of the tracking window exceeds a predefined threshold. The singular values of the affine transformation matrix represent the scale of the tracking window along the principal axes of the transformed window. This allows rejecting points with extremely distorted tracking windows but valid area.

A coarse-to-fine strategy with a Gaussian image pyramid increases the basin of convergence of the gradient descent algorithm and allows handling larger displacements. Linear motion prediction additionally increases convergence. Since feature selection is only performed in the original image, but the features are tracked in all hierarchy levels, the tracking window does not have to be larger than the selection window: tracking in the next higher hierarchy level with the same tracking window size corresponds to tracking in the current hierarchy level with a window of double size.

The inverse compositional approach for affine motion estimation [Bak01] is employed and combined with a linear model for illumination compensation [Jin01]. The conventional error function for point tracking, i. e., the formula for the SSD is

$$\mathrm{SSD} = \sum_{\boldsymbol{q}} \Big(f(\boldsymbol{q}) - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a} + \Delta\boldsymbol{a})) \Big)^2 , \tag{5.6}$$

where $f(\boldsymbol{q})$ and $f_i(\boldsymbol{q})$ denote the gray-values of pixel $\boldsymbol{q}$ in the first and the i-th image. The affine transformation is represented by the parameterized warp function

$$\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a}) = \begin{pmatrix} 1 + a_1 & a_2 \\ a_3 & 1 + a_4 \end{pmatrix} \boldsymbol{q} + \begin{pmatrix} a_5 \\ a_6 \end{pmatrix} , \tag{5.7}$$

where the six motion parameters are contained in $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6)^{\mathrm{T}}$. The motion parameters of the previous image f_{i-1} are assumed to be known (initialized for $f_0 = f$ with $\boldsymbol{a} =$

$\mathbf{0}$). The affine update $\varDelta \boldsymbol{a}$ has to be computed. The inverse compositional algorithm minimizes

$$\mathrm{SSD_I} = \sum_{\boldsymbol{q}} \Big(f(\boldsymbol{g}(\boldsymbol{q}, \varDelta \boldsymbol{a})) - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 \tag{5.8}$$

instead of minimizing SSD. The role of f and f_i are reversed, the affine update $\varDelta \boldsymbol{a}$ is now estimated in the first image f rather than in the current image f_i. The proof of equivalence is given in [Bak01]. A first-order Taylor approximation around $\boldsymbol{g}(\boldsymbol{q}, \mathbf{0})$ yields

$$\begin{aligned} \mathrm{SSD_I} &\approx \sum_{\boldsymbol{q}} \Big(f(\boldsymbol{g}(\boldsymbol{q}, \mathbf{0})) + \nabla f(\boldsymbol{q}) \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{a}}(\boldsymbol{q}, \mathbf{0}) \varDelta \boldsymbol{a} - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 \\ &= \sum_{\boldsymbol{q}} \Big(f(\boldsymbol{q}) + \nabla f(\boldsymbol{q}) \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{a}}(\boldsymbol{q}, \mathbf{0}) \varDelta \boldsymbol{a} - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 . \end{aligned} \tag{5.9}$$

Introducing the vector

$$\boldsymbol{h}(\boldsymbol{q}) = (q_x f_x(\boldsymbol{q}), q_y f_x(\boldsymbol{q}), q_x f_y(\boldsymbol{q}), q_y f_y(\boldsymbol{q}), f_x(\boldsymbol{q}), f_y(\boldsymbol{q}))^{\mathrm{T}} \tag{5.10}$$

with $\boldsymbol{q} = (q_x, q_y)^{\mathrm{T}}$, equation (5.9) can be written as

$$\mathrm{SSD_I} \approx \sum_{\boldsymbol{q}} \Big(\boldsymbol{h}(\boldsymbol{q})^{\mathrm{T}} \varDelta \boldsymbol{a} + f(\boldsymbol{q}) - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 . \tag{5.11}$$

The solution for this least-squares problem is

$$\varDelta \boldsymbol{a} = - \left(\sum_{\boldsymbol{q}} \Big(\boldsymbol{h}(\boldsymbol{q}) \boldsymbol{h}(\boldsymbol{q})^{\mathrm{T}} \Big)^{-1} \right) \left(\sum_{\boldsymbol{q}} \Big(\boldsymbol{h}(\boldsymbol{q}) \big(f(\boldsymbol{q}) - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \big) \Big) \right) . \tag{5.12}$$

The new rule for updating the motion parameters is then given by

$$\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a}_{\text{new}}) = \boldsymbol{g}(\boldsymbol{g}(\boldsymbol{q}, \varDelta \boldsymbol{a})^{-1}, \boldsymbol{a}) = \begin{pmatrix} 1 + a_1 & a_2 \\ a_3 & 1 + a_4 \end{pmatrix} \boldsymbol{g}(\boldsymbol{q}, \varDelta \boldsymbol{a})^{-1} + \begin{pmatrix} a_5 \\ a_6 \end{pmatrix} \tag{5.13}$$

where

$$\boldsymbol{g}(\boldsymbol{q}, \varDelta \boldsymbol{a})^{-1} = \begin{pmatrix} 1 + \varDelta a_1 & \varDelta a_2 \\ \varDelta a_3 & 1 + \varDelta a_4 \end{pmatrix}^{-1} \left(\boldsymbol{q} - \begin{pmatrix} \varDelta a_5 \\ \varDelta a_6 \end{pmatrix} \right) . \tag{5.14}$$

A linear illumination compensation model adjusts the intensity $f(\boldsymbol{q})$ by

$$\alpha f(\boldsymbol{q}) + \beta \,, \tag{5.15}$$

where α adjusts contrast and β brightness. Applying this model to equation (5.11) results in

$$\mathrm{SSD}_{\mathrm{IC}} = \sum_{\boldsymbol{q}} \Big(\alpha f(\boldsymbol{g}(\boldsymbol{q}, \varDelta\boldsymbol{a})) + \beta - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 \,. \tag{5.16}$$

The corresponding Taylor expansion yields

$$\mathrm{SSD}_{\mathrm{IC}} \approx \sum_{\boldsymbol{q}} \Big(\alpha f(\boldsymbol{q})) + \alpha \nabla f(\boldsymbol{q}) \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{a}}(\boldsymbol{q}, 0) \varDelta\boldsymbol{a} + \beta - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 \,. \tag{5.17}$$

Extending the two vectors $\varDelta\boldsymbol{a}$ and $\boldsymbol{h}(x)$ to

$$\tilde{\boldsymbol{h}}(\boldsymbol{q}) = (q_x f_x(\boldsymbol{q}), q_y f_x(\boldsymbol{q}), q_x f_y(\boldsymbol{q}), q_y f_y(\boldsymbol{q}), f_x(\boldsymbol{q}), f_y(\boldsymbol{q}), f(\boldsymbol{q}), 1)^{\mathrm{T}} \tag{5.18}$$

and

$$\varDelta\tilde{\boldsymbol{a}} = (\alpha \varDelta a_1, \alpha \varDelta a_2, \alpha \varDelta a_3, \alpha \varDelta a_4, \alpha \varDelta a_5, \alpha \varDelta a_6, \alpha, \beta)^{\mathrm{T}} \,, \tag{5.19}$$

equation (5.17) can be written as

$$\mathrm{SSD}_{\mathrm{IC}} \approx \sum_{\boldsymbol{q}} \Big(\tilde{\boldsymbol{h}}(\boldsymbol{q})^{\mathrm{T}} \varDelta\tilde{\boldsymbol{a}} - f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \Big)^2 \,. \tag{5.20}$$

The solution is then given by

$$\varDelta\tilde{\boldsymbol{a}} = \left(\sum_{\boldsymbol{q}} \left(\tilde{\boldsymbol{h}}(\boldsymbol{q}) \tilde{\boldsymbol{h}}(\boldsymbol{q})^{\mathrm{T}} \right)^{-1} \right) \left(\sum_{\boldsymbol{q}} \tilde{\boldsymbol{h}}(\boldsymbol{q}) f_i(\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{a})) \right) \,. \tag{5.21}$$

After estimating the *translation* of a feature window based on the previous image and rejecting windows according to $\mathrm{SSD}_{\mathrm{IC}}$, the estimated *affine* motion of the tracking window with respect to the first image is additionally used for preventing feature drift. Since tracking windows will generally not be identical in two consecutive images, e. g., due to image noise, perspective distortions by camera movement, and intensity changes, image-to-image translation estimation cannot be absolutely accurate. When the errors from image to image accumulate, the feature window drifts from its true position. Therefore, the translation component $(a_5, a_6)^{\mathrm{T}}$ of the *affine* motion determines the final feature position with higher accuracy and thus prevents feature drift.

The third modification concerns the factorization method. Instead of assuming a perspective camera model, the factorization is done based on a weak-perspective camera model. This is more robust while maintaining sufficient accuracy. The result of the factorization, i. e., the estimated camera parameters and 3-D world points, are optimized non-linearly utilizing the intrinsic camera parameters determined by camera calibration.

The described modifications allow for the computation of the extrinsic camera parameters for an endoscopic image sequence.

5.4 Extrinsic Parameters from a Robot Arm

The Department of Surgery of the University of Erlangen-Nuremberg routinely performs minimally invasive operations of the abdomen with the voice-controlled endoscope positioning robot AESOP 3000 [Com05] (cf. Figure 2.4, page 20). This robot arm is also employed here. The model number “3000” is omitted in the following. The features of AESOP have already been described in Section 2.3, page 18.

This section describes techniques for determining the camera pose using AESOP. In the case of endoscopic surgery the camera pose denotes the camera coordinate system that is located in the projection center of the endoscope. Therefore, camera pose and endoscope pose are used synonymously in the following. The endoscope pose corresponding to the i-th image is represented by the rotation matrix $\boldsymbol{R}_i$ and the translation vector $\boldsymbol{t}_i$. Without loss of generality it is assumed that the origin of the world coordinate system corresponds to the robot’s base coordinate system. Then

$$^{\mathrm{c}}\boldsymbol{w}_i = \boldsymbol{R}_i{}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{t}_i) \tag{5.22}$$

transforms a 3-D point $\boldsymbol{w}$ specified in base/world coordinates to a 3-D point $^{\mathrm{c}}\boldsymbol{w}_i$ specified in the camera coordinate system that corresponds to the i-th endoscope pose (cf. equation (3.6), page 36). Multiplication by $\boldsymbol{R}_i$ from the left and addition of $\boldsymbol{t}_i$ yields the inverse transformation, i. e., from endoscope coordinates to world coordinates:

$$\boldsymbol{w} = \boldsymbol{R}_i\,{}^{\mathrm{c}}\boldsymbol{w}_i + \boldsymbol{t}_i \,. \tag{5.23}$$

The latter equation is generally used in the case of robot arms since the objective is to determine world coordinates of points given in camera coordinates. Using homogeneous coordinates, equation (5.23) can be written as

$$\underline{\boldsymbol{w}} \sim \underbrace{\begin{pmatrix} \boldsymbol{R}_i & \boldsymbol{t}_i \\ \boldsymbol{0}_3{}^{\mathrm{T}} & 1 \end{pmatrix}}_{=:\boldsymbol{T}_i} {}^{\mathrm{c}}\underline{\boldsymbol{w}} = \boldsymbol{T}_i\, {}^{\mathrm{c}}\underline{\boldsymbol{w}}\,. \tag{5.24}$$

In the case of a robot arm, each 4×4 transformation matrix $\boldsymbol{T}_i$ can be separated into two transformations: $\boldsymbol{T}_i = \boldsymbol{T}_{\mathrm{B,H}}\boldsymbol{T}_{\mathrm{H,E}}$. The first transformation $\boldsymbol{T}_{\mathrm{B,H}}$ relates the base/world coordinate system (base frame) and the hand of the robot (hand frame). It therefore specifies the pose of the hand in space with respect to the base of the robot. The second transformation $\boldsymbol{T}_{\mathrm{H,E}}$ relates the poses of endoscope and robot hand. It is also known as *hand-eye transformation*. Both transformations together result in the required pose of the endoscope in world coordinates by first specifying the endoscope's pose in hand coordinates by applying $\boldsymbol{T}_{\mathrm{H,E}}$ and then transforming this pose to world coordinates by $\boldsymbol{T}_{\mathrm{B,H}}$. Note that only $\boldsymbol{T}_{\mathrm{B,H}}$ changes when the robot moves.

At first the computation of $\boldsymbol{T}_{\mathrm{B,H}}$ based on robot kinematics is shown in the next section. After that, Section 5.4.2 describes a method for obtaining $\boldsymbol{T}_{\mathrm{H,E}}$.

5.4.1 AESOP's Kinematics

This section describes the computation of $\boldsymbol{T}_{\mathrm{B,H}}$ based on robot kinematics. *Kinematics* is the relationships between the positions, velocities, and accelerations of the links of a robot arm [McK91]. Only the relationships between positions are relevant here. AESOP is a serial link manipulator: the hand is connected to the base by links, with each link connected to the next by a joint. AESOP has seven joints, i. e., seven degrees of freedom: three active and two passive rotary (revolute) joints, one rotary joint that has to be set by the user, and one active translational (sliding) joint. If a coordinate system is attached to each link, the relationship between two links i and j, i. e., the transformation of a point given in the j-th coordinate system to a point specified in the i-th coordinate system, can be expressed by a rotation matrix $\boldsymbol{R}_{i,j} \in \mathbb{R}^{3\times 3}$ and a translation vector $\boldsymbol{t}_{i,j} \in \mathbb{R}^3$. Again a homogeneous 4×4 transformation matrix is used:

$$\boldsymbol{T}_{i,j} = \begin{pmatrix} \boldsymbol{R}_{i,j} & \boldsymbol{t}_{i,j} \\ \boldsymbol{0}_3{}^{\mathrm{T}} & 1 \end{pmatrix}\,. \tag{5.25}$$

Robot arms are generally manufactured such that the rotation from one link to the next can simply be specified by a rotation about one of the coordinate axes. AESOP is also manufactured in this way. Thus, only three types of rotation matrices are necessary, corresponding to rotation about the x, y, and z-axis [McK91]:

$$\boldsymbol{R}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix} \tag{5.26}$$

$$\boldsymbol{R}_y(\beta) = \begin{pmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{pmatrix} \tag{5.27}$$

$$\boldsymbol{R}_z(\gamma) = \begin{pmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix} . \tag{5.28}$$

This means $\boldsymbol{R}_{i,j} \in \{\boldsymbol{R}_x(\alpha), \boldsymbol{R}_y(\beta), \boldsymbol{R}_z(\gamma)\}$. Using homogeneous transformation matrices, $\boldsymbol{T}_{\mathrm{B,H}}$ can be expressed as

$$\boldsymbol{T}_{\mathrm{B,H}} = \boldsymbol{T}_{\mathrm{B},1}\boldsymbol{T}_{1,2}\boldsymbol{T}_{2,3} \cdot \ldots \cdot \boldsymbol{T}_{n-1,n} \,, \tag{5.29}$$

where n is the number of links and $\boldsymbol{T}_{\mathrm{B},1}$ is the transformation from the first link to the robot base, $\boldsymbol{T}_{1,2}$ is the transformation from the second to the first link, and so on. This equation is also called *forward kinematic transform* [McK91]. Figure 5.3 shows the kinematics of AESOP as provided by the manufacturer Computer Motion Inc. The following forward kinematic transform corresponds to the provided kinematics:

$$\begin{aligned} \boldsymbol{T}_{\mathrm{B,H}} = & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & l_1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & l_2 \\ & \boldsymbol{R}_z(\alpha_1) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & l_3 \\ & \boldsymbol{R}_z(\alpha_2) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_x(90^\circ) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \\ & \begin{pmatrix} & & & l_4 \\ & \boldsymbol{R}_z(\alpha_3) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_z(\alpha_4) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & l_5 \\ & \boldsymbol{R}_x(\alpha_5) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_y(\alpha_6) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ = & \begin{pmatrix} \boldsymbol{R}_{\mathrm{B,H}} & \boldsymbol{t}_{\mathrm{B,H}} \\ \boldsymbol{0}_3{}^{\mathrm{T}} & 1 \end{pmatrix} . \end{aligned} \tag{5.30}$$

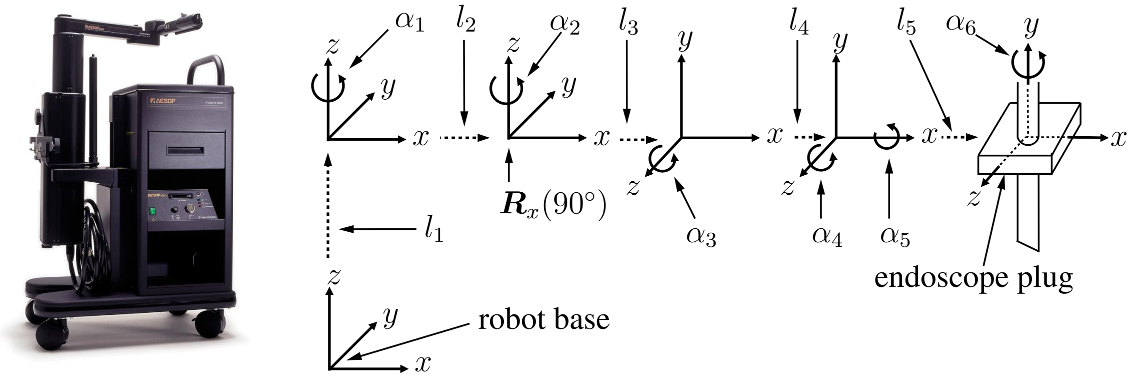


Figure 5.3: On the left, the robot arm AESOP 3000 on its transportation cart is shown. Its kinematics is visualized on the right. AESOP has seven joints, i. e., seven degrees of freedom: three active ($\alpha_1, \alpha_2, \alpha_6$) and two passive rotary joints (α_4, α_5), one rotary joint that has to be set by the user (α_3), and one active translational joint (l_1). Several translations between joints are fixed: $l_2 = 384.2\,\text{mm}$, $l_3 = 81.8\,\text{mm}$, $l_4 = 304.8\,\text{mm}$, and $l_5 = 16.8\,\text{mm}$. The rotation $\boldsymbol{R}_x(90°)$ changes the orientation of the coordinate system: after the rotation is performed, the y-axis points to the direction of the former z-axis and the z-axis points to the negative direction of the former y-axis.

The columns of $\boldsymbol{R}_{\text{B,H}}$ correspond to the coordinate system axes of AESOP's hand, i. e., the endoscope plug, and $\boldsymbol{t}_{\text{B,H}}$ is the position of the hand. $\boldsymbol{T}_{\text{B,H}}$ can be computed if all parameters of AESOP's kinematics are known, namely $l_1, \ldots, l_5$, and $\alpha_1, \ldots, \alpha_6$. The controller of AESOP permits reading out the seven parameters that change during robot movements: $\alpha_1, \ldots, \alpha_6$ and l_1. The data are obtained through a serial interface. For light field reconstruction the seven parameters are read out before and after grabbing an image. If AESOP is moved continuously, the seven parameters before and after grabbing an image will usually be slightly different. Since the exact parameters corresponding to the grabbed image are not known, the mean values are employed for further computations. Now, the pose of AESOP's hand can be computed for each captured image.

Before each endoscopic surgery, the endoscope is mounted onto the hand. The pose of the endoscope with respect to AESOP's hand is fixed but has to be computed once for each setup. The problem of determining the transformation $\boldsymbol{T}_{\text{H,E}}$ from a robot hand to a camera that is mounted onto the hand is also known as *hand-eye calibration*. In the following section, a hand-eye calibration technique for AESOP is presented.

5.4.2 Hand-Eye Calibration of AESOP

This hand-eye calibration method extends AESOP's kinematics in order to describe the transformation $\boldsymbol{T}_{\text{H,E}}$. Four additional parameters are introduced (see Figure 5.4): the length of the

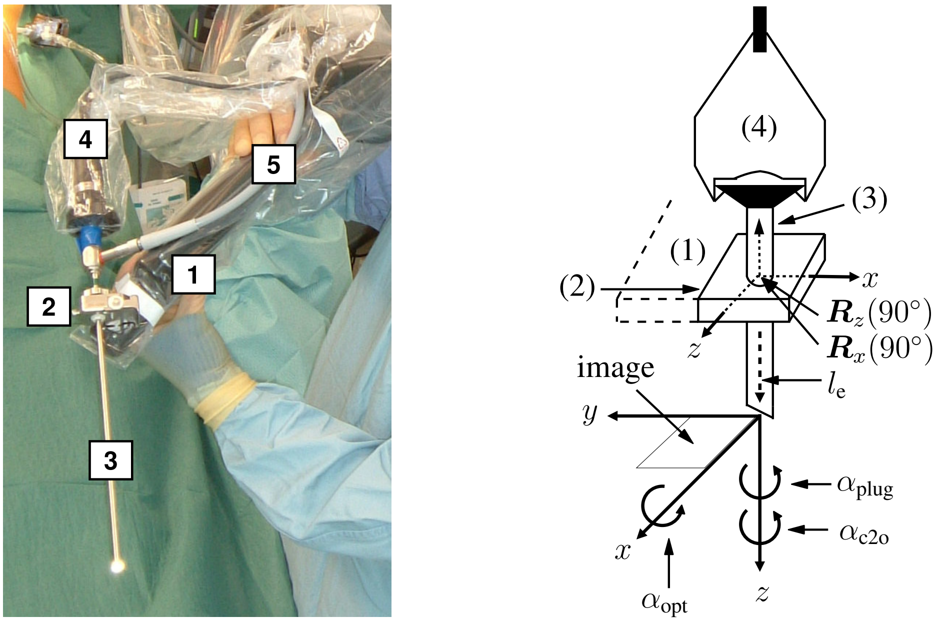


Figure 5.4: The left side shows the setup of AESOP, endoscope, and camera head. The figure to the right shows the corresponding extended kinematics. AESOP's hand (2), i. e., the plug for the endoscope, is connected to the proximal end of the robot arm (1). The camera head (4) is mounted onto the endoscope (3). Four additional parameters are introduced which are sufficient for obtaining the pose of the endoscope: the length of the endoscope l_e, the angle inside the endoscope plug α_plug, the angle of the side view optics α_opt, and the angle between camera head and optics α_c2o. Additionally, two 90° rotations are introduced: the first rotates 90° about the x-axis and results in the z-axis pointing to the viewing direction of the endoscope; the second rotation about the z-axis by 90° is optional and was introduced such that the x-axis points to a desired direction for $\alpha_6 = 0$.

endoscope l_e, the angle inside the endoscope plug α_plug, the angle of the side view optics α_opt, and the angle between camera head and optics α_c2o. Based on the hand coordinate system, the pose of the endoscope is obtained as follows. The hand coordinate system is transformed by a translatory movement of the length of the endoscope l_e. Following this, two 90° rotations are performed. The first rotates by 90° about the x-axis and results in the z-axis pointing to the viewing direction of the endoscope. The second rotation about the z-axis by 90° is optional and was introduced so that the x-axis points to a desired direction for $\alpha_6 = 0$ (cf. Figure 5.3). The next rotation about the z-axis by α_plug takes into account the fact that the endoscope can be rotated arbitrarily inside the endoscope plug before it is fixed. Since side view optics are employed, the angle of the optics α_opt is modeled by changing the viewing direction by a rotation about the x-axis by α_opt. Finally, the possibility of rotating the camera head is represented by a rotation

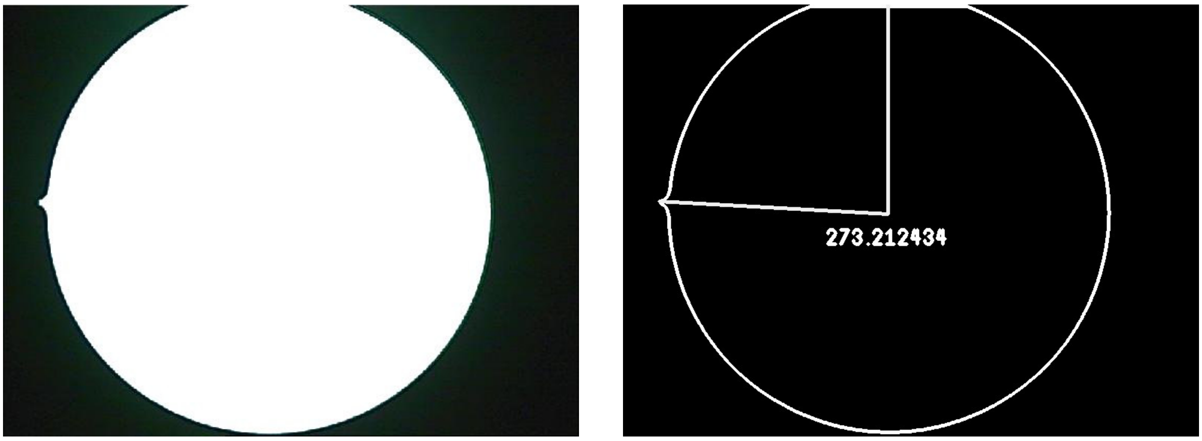


Figure 5.5: Notch detection: an image of a homogeneous white surface is captured (left). The contour and its center are computed. The notch is defined as that point on the contour with the largest distance to the middle point (right). The angle between the "up" direction and the vector from the center of the contour to the notch is computed. In this example it is 273.2°.

about the z-axis by α_{c2o}. Altogether, these transformations allow determining $\boldsymbol{T}_{\text{H,E}}$:

$$\boldsymbol{T}_{\text{H,E}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -l_{\text{e}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_x(90^\circ) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_z(90^\circ) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_z(\alpha_{\text{plug}}) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot$$

$$\begin{pmatrix} & & & l_3 \\ & \boldsymbol{R}_x(-\alpha_{\text{opt}}) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} & & & 0 \\ & \boldsymbol{R}_z(\alpha_{\text{c2o}}) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{R}_{\text{H,E}} & \boldsymbol{t}_{\text{H,E}} \\ \boldsymbol{0}_3{}^{\text{T}} & 1 \end{pmatrix} . \tag{5.31}$$

The minus sign of $\boldsymbol{R}_x(-\alpha_{\text{opt}})$ is due to the construction of the endoscope: based on the attached coordinate system (see Figure 5.4) the optical axis of a side view endoscope with angle α_{opt} is rotated about the x-axis towards the negative mathematical rotation direction, i. e., by $-\alpha_{\text{opt}}$.

For each new operation setup only the angle of the optics is known. For the operations regarded here it is always 30 °. The length of the endoscope is measured by hand. The angle α_{c2o} is obtained by detecting a notch at the border of the optics (see Figure 5.5). This notch indicates the "up"-direction of the endoscope and facilitates the rotation of the camera head with respect to the optics during the operation. In a first step an image of a homogeneous white surface, e. g., a sheet of paper, is captured. The color image is converted to a gray-value image. A binary image is then obtained by applying a threshold value of 200. The contour and its center are computed as described in [Suz85]. The notch is defined as the contour point with the largest distance to the center. The angle between the "up" direction, i. e., the negative y-axis of the image coordinate system, and the vector from the center of the contour to the notch is computed. This angle corresponds to a *rotation of the optics* when the camera head is kept fixed. However, for the extended kinematics it was assumed that the endoscope is mounted into the plug and the *camera head is rotated* by α_{c2o}. Thus, α_{c2o} is the negative value of the computed angle.

Now only one parameter needs to be determined: α_{plug}. The idea for its computation is to calculate the relative movement between two endoscope poses by using a calibration pattern and to adjust α_{plug} in such a way that the relative movement calculated by the extended kinematics equals the one computed by camera calibration. Let $\boldsymbol{R}_{\mathrm{C1}}$ and $\boldsymbol{R}_{\mathrm{C2}}$ be the rotation matrices obtained by camera calibration, $\boldsymbol{t}_{\mathrm{C1}}$ and $\boldsymbol{t}_{\mathrm{C2}}$ the corresponding translation vectors, and $\boldsymbol{R}_{\mathrm{A1}}, \boldsymbol{R}_{\mathrm{A2}}, \boldsymbol{t}_{\mathrm{A1}}$, and $\boldsymbol{t}_{\mathrm{A2}}$ the parameters computed by AESOP's extended kinematics with $\alpha_{\mathrm{plug}} = 0$. The relative positions of the second camera to the first, denoted by $\boldsymbol{t}_{\mathrm{C1,2}}$ and $\boldsymbol{t}_{\mathrm{A1,2}}$ are calculated as follows. The translation vector of the second endoscope pose is transformed into the coordinate system of the first endoscope pose:

$$\begin{aligned} \boldsymbol{t}_{\mathrm{C1,2}} &= {\boldsymbol{R}_{\mathrm{C1}}}^{\mathrm{T}}(\boldsymbol{t}_{\mathrm{C2}} - \boldsymbol{t}_{\mathrm{C1}}) && (5.32) \\ \boldsymbol{t}_{\mathrm{A1,2}} &= {\boldsymbol{R}_{\mathrm{A1}}}^{\mathrm{T}}(\boldsymbol{t}_{\mathrm{A2}} - \boldsymbol{t}_{\mathrm{A1}}) && (5.33) \end{aligned}$$

The 3-D angle $\sphericalangle(\boldsymbol{t}_{\mathrm{C1,2}}, \boldsymbol{t}_{\mathrm{A1,2}})$ is used as a one-dimensional similarity measure which is optimized. Because α_{plug} is in the range $[0, 360]^{\circ}$, a discrete search over the range is sufficient to determine the value that minimizes $\sphericalangle(\boldsymbol{t}_{\mathrm{C1,2}}, \boldsymbol{t}_{\mathrm{A1,2}})$.

The main drawbacks of this method are that the length of the endoscope has to be measured by hand and that two separate steps are required to obtain $\boldsymbol{T}_{\mathrm{H,E}}$. However, it has to be noted that the application of conventional hand-eye calibration techniques like the one employed for the optical tracking system smARTtrack1 (cf. Section 5.5.2, page 106) would be complicated since several images of a calibration pattern have to be captured, from different hand/eye poses, where AESOP can only be moved when the endoscope is fixed at a certain point (usually the keyhole). Since the calibration has to be performed under sterile conditions in the operating room, a sterile patient model would have to be constructed, into which the calibration pattern could be inserted. But even if such a patient model were available, the movements of AESOP are restricted due to the fixation point at the keyhole, and thus not well suited for hand-eye calibration, where hand/eye poses with as different a rotation axis as possible are necessary to

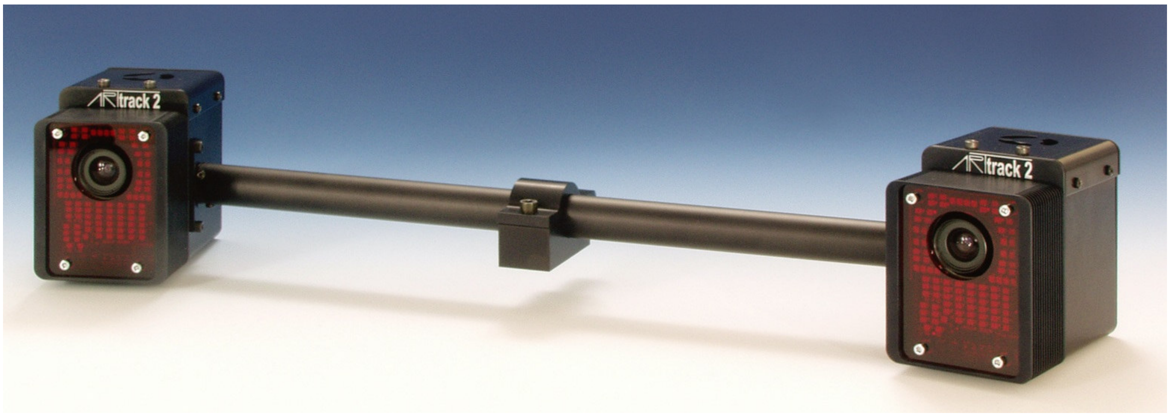


Figure 5.6: The infrared optical tracking system smARTtrack1 with two ARTtrack2 cameras (image by courtesy of Advanced Realtime Tracking GmbH).

obtain accurate results (cf. Section 5.5.2, page 108).

Regarding the arguments of the last paragraph, the question arises as to how the two images of a calibration pattern that are necessary for the method presented in this section to compute α_{plug} are captured. This problem is simpler since only two images have to be captured and a simple movement is sufficient. The two images are acquired as follows: a calibration pattern is placed on a sterile table; the surgeon fixates the endoscope by hand; the first image is captured and AESOP is moved a small distance upwards for capturing the second image.

5.5 Extrinsic Parameters from an Optical Tracking System

The infrared optical tracking system smARTtrack1 [Adv05] is employed. It consists of two or more cameras, a PC, and one or more targets that are tracked (cf. Figures 5.6 and 5.7).

A target is built from markers that can easily be identified in the captured images. The use of infrared light simplifies marker identification. The cameras are calibrated with a calibration kit (cf. Figure 5.7): the world coordinates are defined by a rectangular target while the transformations between the cameras are calibrated using a wand with two markers that is rotated in front of the cameras for a few seconds.

The pose of a target is obtained by computing the 3-D position of each visible marker. The knowledge of the geometry of a target then allows computing its pose. The geometry of an unknown target is obtained by another calibration step: the target is moved for a few seconds in front of the cameras, and since only this target is visible, the computed relative positions of all visible markers define the geometry of the target.

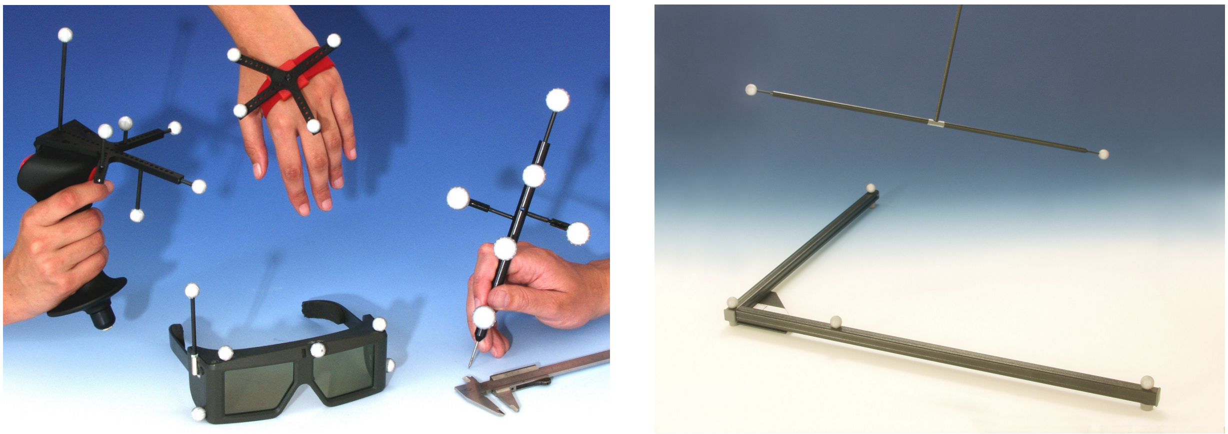


Figure 5.7: Some commonly used targets with passive markers (left) and the calibration kit (right) for smARTtrack1 (images by courtesy of Advanced Realtime Tracking GmbH).

smARTtrack1 is capable of providing the poses of up to 15 targets with 60 Hz. The pose information is provided via Ethernet. The maximum allowed distance of the target from the cameras is 4 m. The resolution of the cameras is 658×496 pixels. The use of more than two cameras increases the accuracy and enlarges the measurement volume, i. e., the space where the target is visible by at least two cameras.

The system used here consists of two cameras mounted on a tripod, as shown in Figure 5.6. Only one target is used at a time. Lenses with a focal length of 8 mm are employed, resulting in 34° horizontal and 26° vertical aperture angle. According to A.R.T. the accuracy for this setup in terms of root mean square errors (RMSE) is:

- 0.19 mm position error in x- and y-direction, parallel to the image plane of the cameras,
- 0.36 mm position error in z-direction, and
- 0.14° rotation error.

In order to obtain accurate pose information for each captured *endoscopic* image, the A.R.T. PC has to be synchronized with the grabbing PC (see Figure 5.8). For this purpose the video signal (S-VHS) is used as synchronization signal. S-VHS is a 50 Hz signal since 50 half images are transferred. The S-VHS output of the endoscope-camera is connected to the synchronization card of the A.R.T. PC, which transfers the signal to the smARTtrack1 cameras. Thus smARTtrack1 captures its images for pose computation at the same time when the endoscope-camera captures one half-image. After receiving the synchronization signal it takes 15 to 18 milliseconds until smARTtrack1 provides the pose information via Ethernet. As it takes 20 milliseconds

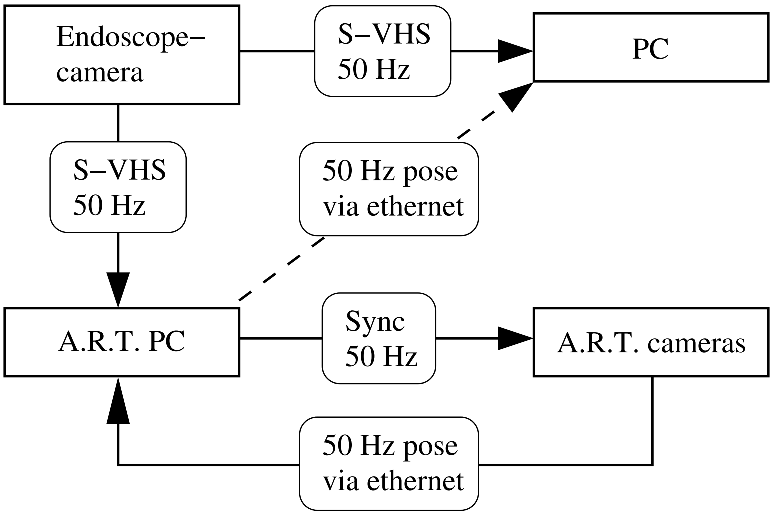


Figure 5.8: Synchronization of smARTtrack1 with the PC of the video-endoscopic system: the S-VHS signal of the endoscope-camera serves as synchronization signal. The signal is transfered to the smARTtrack1 cameras via the A.R.T. PC. The cameras grab images according to the synchronization signal and perform the pose computation (they contain small PCs). The pose information is provided via Ethernet by the A.R.T. PC.

to transfer one half image, the pose information will be available shortly before the transfer of the half image is completed. After finishing the transfer of the half image the last obtained pose information is assigned to this half image.

A suitable target is necessary in order to obtain the pose of the endoscope by smARTtrack1, and the hand-eye transformation from target to endoscope has to be determined. Section 5.5.1 deals with the target design and Section 5.5.2 describes the hand-eye calibration process.

5.5.1 Target Design

A suitable target for endoscope tracking was not available. Therefore, three targets were developed in cooperation with A.R.T. This section describes the design and construction of these targets.

A fundamental question is whether active or passive markers should be used. Here, passive markers in terms of spheres with a retro-reflective surface are used (cf. Figure 5.7). Using active markers would be more complicated since usually infrared LEDs are employed which have to be supplied with power. Additionally, the manufacturing of an active target is expensive and its design cannot be changed afterwards. Passive markers are therefore better suited for research as they are cheaper and can easily be re-arranged to build a new target.

A u-shaped adapter is screwed with its cover around the endoscope (cf. Figures 5.9 and 5.10)

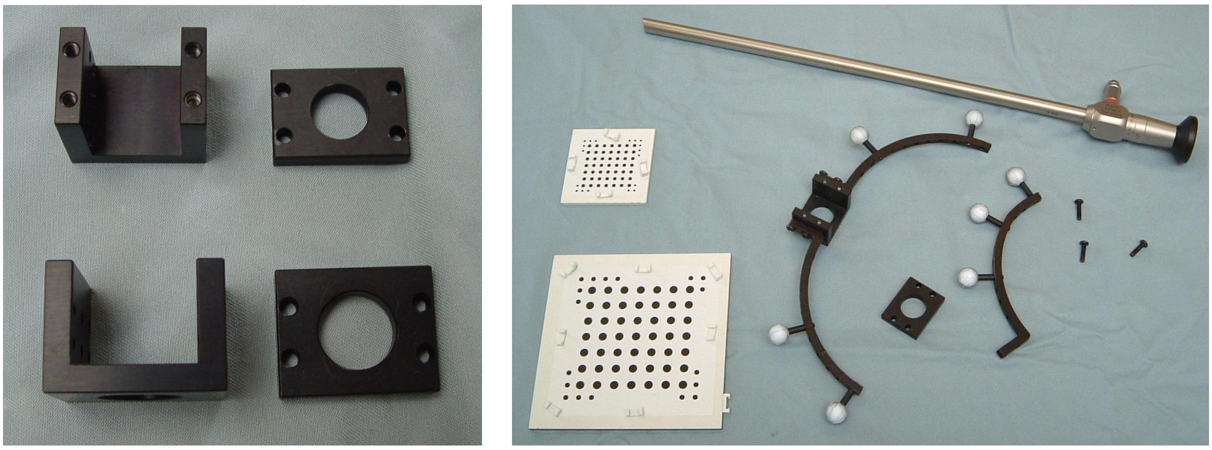


Figure 5.9: The target adapter (left) and the *epee* target taken apart (right). The u-shaped target adapter is screwed with its cover around the endoscope. The cover contains a hole for the light conductor. On the right, two connecting arcs, each with two markers, are already screwed onto the u-shaped adapter. Beside it to the right lie the cover, the middle arc with three markers, and the screws. Additionally, a 10 mm Storz endoscope and two calibration patterns are shown (circle distance 10 mm and 5 mm).

in order to fixate the target to the endoscope. The cover contains a hole for the light conductor. The adapter allows screwing on the metallic or plastic connecting pieces on which the markers are mounted.

Markers with 12 and 14 mm diameter are employed. The design of the target has to be such that the surgeon is not hampered when using the endoscope during a minimally invasive operation. This restricts the possible target designs and optimal designs known from theoretical examinations [Wes04] cannot be realized. Three target designs are presented: *epee*, *double decker* and *double decker 2z* (see Figures 5.10 to 5.12).

The *epee* target consists of three arcs (see Figures 5.9 and 5.10). Two markers are screwed onto the two lateral arcs and three markers are screwed onto the middle arc. The idea of this target is to distribute the markers in all directions without hampering the surgeon. The main drawback of this design is that sometimes merging and occluded markers occur. Markers are either occluded by the endoscope and target itself or by the light conductor (cf. Figure 5.10). Merging markers occur when it is not possible for smARTtrack1 to separate two markers because they overlap in the camera image. This leads to errors in pose determination because if the merging markers are not recognized, the error of the estimated 3-D position for this marker is larger than normal. The design of the *epee* target, i. e., the distribution of markers in all directions, leads to a higher likelihood of merging markers. Therefore, the next target was designed such that the likelihood for merging and occluded markers is decreased.

The basic idea of the *double decker* target is to design the target such that all markers lie

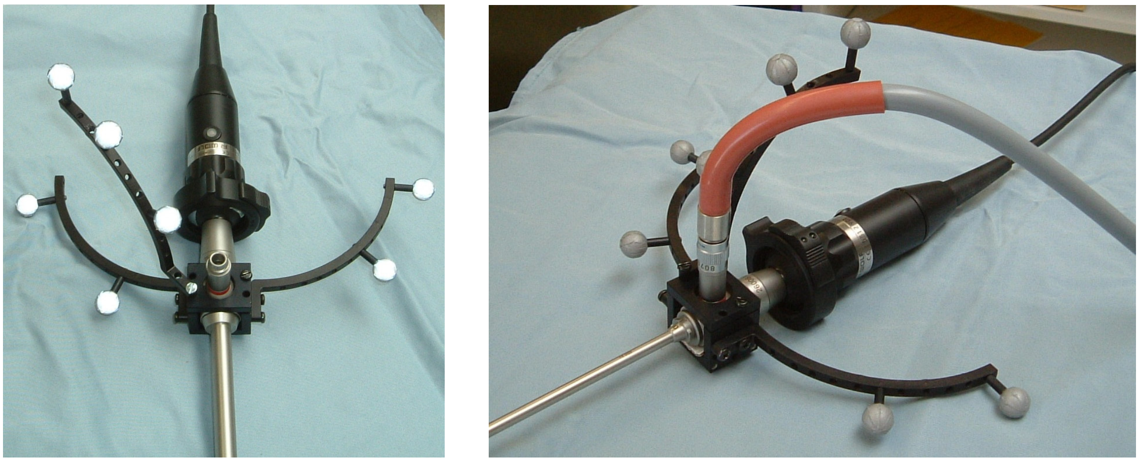


Figure 5.10: The *epee* target screwed onto the endoscope with mounted camera head (left). The *epee* target was not designed to avoid merging and occluded markers. Markers sometimes merge when the endoscope points sideways, and when the light conductor is connected (right), markers are also occluded by the light conductor.

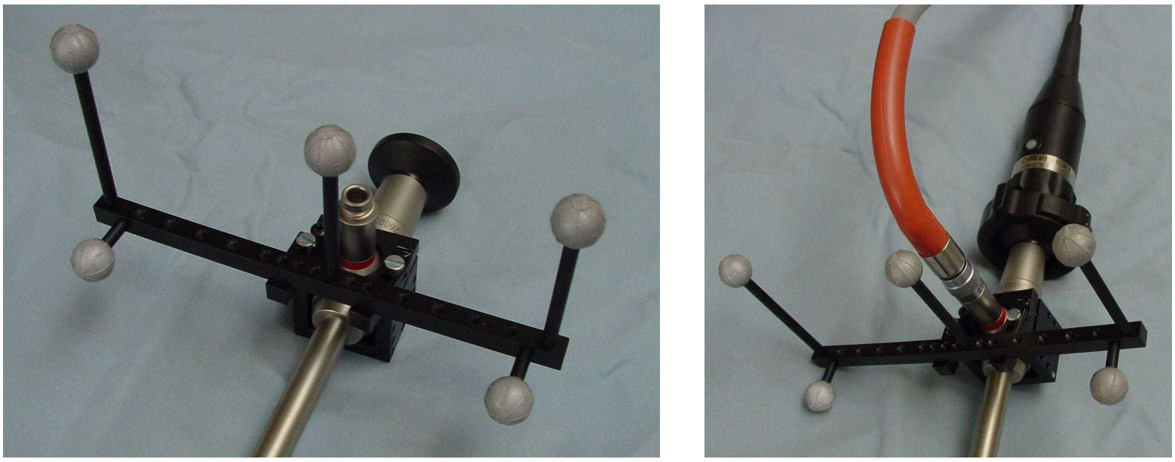


Figure 5.11: The *double decker* target. Five markers are screwed onto a connecting piece such that they lie approximately in the same plane. The left image shows the target screwed onto the endoscope without camera head and light conductor whereas on the right side both are mounted.

approximately on a plane (cf. Figure 5.11). If this plane is visible by smARTtrack1, occluding and merging markers will only occur if the endoscope points approximately $90°$ sideways.

The disadvantage of this plane-approach is that the pose accuracy perpendicular to this plane, in this case along the endoscope towards its tip, is — at least theoretically — decreased. Therefore, the idea of the *double decker 2z* target is to replace the top middle marker of the *double decker* target by two markers that are located perpendicular to the plane (see Figure 5.12). This is achieved by using the middle arc of the *epee* target, where only two markers are screwed at

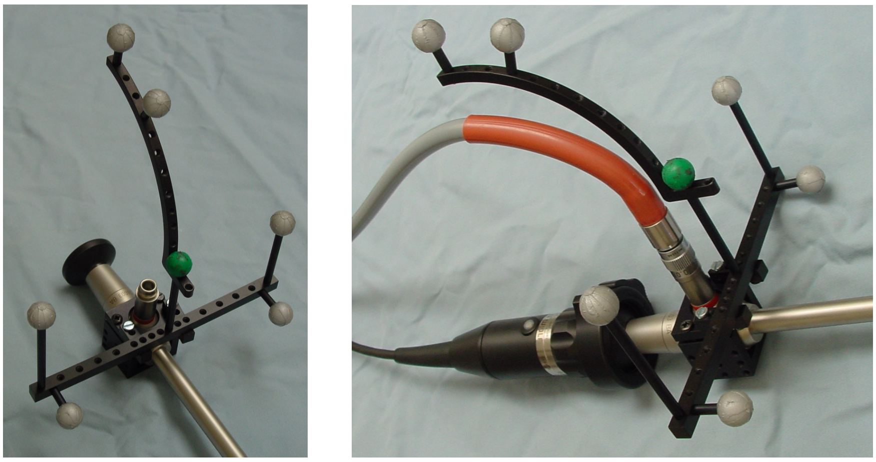


Figure 5.12: The *double decker 2z* target. Four markers lie approximately in the same plane. Additionally, the middle arc of the *epee* target was moved in front of the light conductor and two markers are placed at the end of it to provide higher pose accuracy in the direction towards the endoscope tip. The marker at the bottom of the arc is used for fixating the arc, the retro-reflective surface was removed which makes the marker invisible for smARTtrack1. The left image shows the target without camera head and light conductor whereas on the right side both are mounted.

the end of it. These two markers only merge when the endoscope points „upwards“. This case generally does not occur since smARTtrack1 „looks down“ onto the operating table. When the patient lies flat on the table it is anatomically impossible to move the endoscope upwards like this. It would only be possible when the table is moved to an extreme sloping position.

Finally, the markers and the other parts of the target have to be sterilizable. The materials that are used for target construction can be sterilized with gas. For usage during an operation the target is sterilized and put together in the operating room under sterile conditions.

5.5.2 Hand-Eye Calibration of smARTtrack1

In general, the objective of hand-eye calibration is the computation of the hand-eye transformation $\boldsymbol{T}_{\mathrm{H,E}}$ based on hand and eye poses. The eye is usually a camera mounted onto the hand of a robot. The required poses are generally obtained by capturing several images of a calibration pattern with different hand and eye poses. Eye poses are then computed by camera calibration and hand poses by applying the robot's kinematics. Here, the hand is the target and the eye is the endoscope to which the target is attached. The target/hand pose is computed by the optical tracking system and the eye pose is still computed by camera calibration, where the algorithm of Section 4.1.2, page 61, is employed.

Let N_{f} be the number of images that were captured. In the following, hand and eye poses are expressed by rotation matrices and translation vectors that are represented by 4×4 transformation matrices:

$$\boldsymbol{T}_{\mathrm{H}i} = \begin{pmatrix} \boldsymbol{R}_{\mathrm{H}i} & \boldsymbol{t}_{\mathrm{H}i} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} \qquad \boldsymbol{T}_{\mathrm{E}i} = \begin{pmatrix} \boldsymbol{R}_{\mathrm{E}i} & \boldsymbol{t}_{\mathrm{E}i} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} \quad \text{for } i = 0, \ldots, N_{\mathrm{f}} - 1 \,, \tag{5.34}$$

where $\boldsymbol{T}_{\mathrm{H}i}$ and $\boldsymbol{T}_{\mathrm{E}i}$ denote the i-th hand and eye pose, respectively. The transformation from pose i to j is represented by the 4×4 matrix $\boldsymbol{T}_{\mathrm{H}i,j}$ and $\boldsymbol{T}_{\mathrm{E}i,j}$, respectively. The problem of hand-eye calibration may then be visualized by the following diagram, which describes the relations for an arbitrary pair of hand-eye poses:

$$\begin{tikzcd}
\boldsymbol{T}_{\mathrm{H}j} \arrow[r, "\boldsymbol{T}_{\mathrm{H,E}}"] & \boldsymbol{T}_{\mathrm{E}j} \\
\boldsymbol{T}_{\mathrm{H}i} \arrow[u, "\boldsymbol{T}_{\mathrm{H}i,j}"] \arrow[r, "\boldsymbol{T}_{\mathrm{H,E}}"] & \boldsymbol{T}_{\mathrm{E}i} \arrow[u, "\boldsymbol{T}_{\mathrm{E}i,j}"']
\end{tikzcd} \quad . \tag{5.35}$$

According to this diagram the following equation holds:

$$\begin{aligned} \boldsymbol{T}_{\mathrm{E}i,j}\boldsymbol{T}_{\mathrm{H,E}} &= \boldsymbol{T}_{\mathrm{H,E}}\boldsymbol{T}_{\mathrm{H}i,j} \\ \Leftrightarrow \begin{pmatrix} \boldsymbol{R}_{\mathrm{E}i,j} & \boldsymbol{t}_{\mathrm{E}i,j} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{R}_{\mathrm{H,E}} & \boldsymbol{t}_{\mathrm{H,E}} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} &= \begin{pmatrix} \boldsymbol{R}_{\mathrm{H,E}} & \boldsymbol{t}_{\mathrm{H,E}} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{R}_{\mathrm{H}i,j} & \boldsymbol{t}_{\mathrm{H}i,j} \\ \boldsymbol{0}_3^{\mathrm{T}} & 1 \end{pmatrix} . \end{aligned} \tag{5.36}$$

One way to solve this equation is to compute $\boldsymbol{R}_{\mathrm{H,E}}$ and $\boldsymbol{t}_{\mathrm{H,E}}$ *separately* [Shi89, Tsa89, Wan92, Cho91]. Another way is to compute $\boldsymbol{R}_{\mathrm{H,E}}$ and $\boldsymbol{t}_{\mathrm{H,E}}$ *simultaneously*, either by non-linear optimization [Hor95] or by deriving a linear equation system based on representing rotations and translations as dual quaternions [Dan99, Dan01]. The latter solution is applied here. As can be seen the input for the algorithm are pairs of hand-eye poses, i. e., relative movements rather than single poses.

All methods for solving equation (5.36) have in common that at least two movement pairs with non-parallel rotation axes are necessary, i. e., $N_{\mathrm{f}} \geq 3$ and $\exists\ i, j, k, l : \boldsymbol{r}_{i,j} \nparallel \boldsymbol{r}_{k,l}$ [Tsa89, Che91] where $\boldsymbol{r}_{i,j}$ denotes the rotation axis defining the rotation from pose i to j and $\boldsymbol{r}_{k,l}$ the one from pose k to l. In this case the parameterization of rotations as rotation by an angle ϕ about a rotation axis $\boldsymbol{r} \in \mathrm{I\!R}^3$ is used. The formula for computing a rotation matrix $\boldsymbol{R}$ from ϕ and $\boldsymbol{r}$ was already given in equation (4.38), page 70. Conversely, ϕ and $\boldsymbol{r}$ can be computed from the eigenvalues of $\boldsymbol{R}$ which are 1, $\cos\phi + i\sin\phi$, and $\cos\phi - i\sin\phi$, where i is the imaginary unit. The rotation axis $\boldsymbol{r}$ is collinear to the eigenvector of $\boldsymbol{R}$ corresponding to the eigenvalue 1. The rotation angle ϕ can be obtained from either of the two other eigenvalues.

Apart from the prerequisite above, several guidelines to improve the overall accuracy were proposed by [Tsa89]. The four most important ones are:

1. Maximize the angles between different rotation axes. The error for $\boldsymbol{R}_{\mathrm{H,E}}$ is inversely proportional to the sine of the angle between different rotation axes. Thus rotation angles of $90°$ are optimal.

2. Minimize the distance between camera lens center and calibration pattern.

3. Minimize the distance between hand poses.

4. Use redundant poses (error reduction of non-systematic sources by $\sqrt{N_{\mathrm{f}}}$).

Fortunately, wide-angle lenses are employed in endoscopic surgery. This allows using small calibration patterns and moving the lens of the endoscope close to the calibration pattern. The printable asymmetric 7 × 7 pattern is employed (see Figure 5.9 and Figure 4.4, page 61), where a distance between the calibration points of 10 mm is used. About 20 images of the calibration pattern are captured by moving the endoscope by hand around the calibration pattern. The movements are performed following Tsai's guidelines. In practice it is very helpful to use a program that captures images only when the endoscope is kept still. This allows moving the endoscope around without capturing images until the next good pose is found. This is achieved by comparing successive target poses and defining „keeping still“ as moving the target less than a certain threshold, e. g., less than 1 mm. Alternatively, a whole image sequence with one or two hundred images is captured. Then, one of the data selection algorithms described in the following paragraphs should be used.

Since the input for the hand-eye algorithm is a set of relative movements from pose i to j, pairs of poses (i, j) have to be selected from the set of all possible pairs

$$\mathcal{M} = \{(i,j) \mid i,j \in \{0,\ldots,N_{\mathrm{f}}-1\}, i \neq j\} \ . \tag{5.37}$$

When the endoscope can be moved according to the guidelines, it is useful to choose the movements such that the pairs can simply be selected in temporal order. Then

$$\mathcal{M}_t = \{(i,i+1) \mid i = 0,\ldots,N_{\mathrm{f}}-2\} \tag{5.38}$$

defines the set of selected pairs. For each selected pair one equation such as (5.36) is obtained.

When it is not possible to move the endoscope according to the guidelines, i. e., when a whole image sequence is captured, using the set $\mathcal{M}_t$ leads to suboptimal results [Sch03b, Sch04a]. Then the goal is to find a subset $\mathcal{M}_s \subset \mathcal{M}$ with $|\mathcal{M}_s| \geq 3$ that fulfills the guidelines as good as possible. The size of $\mathcal{M}_s$ is usually set to a fixed value, e. g., $|\mathcal{M}_s| = 600$. Two data selection approaches are presented: exhaustive search [Sch03b] and selection based on vector quantization [Sch04a] (see also [Sch06]).

Both methods are preceded by a preprocessing step that defines a new set $\mathcal{M}' \subset \mathcal{M}$ which contains only those pairs of $\mathcal{M}$ that are well suited for hand-eye calibration. Since the rotational error is inversely proportional to the sine of the rotation angle $\phi_{i,j}$ of pair (i, j), those pairs with $\phi_{i,j}$ close to 0° or 180° are neglected. The closeness is defined by a threshold $\theta_{\text{angle}} \in [0°, 90°)$ and $\mathcal{M}'$ is then defined as

$$\begin{aligned}\mathcal{M}' = \{(i,j) \mid (i,j) \in \mathcal{M}, \quad \phi_{i,j} &\in [\theta_{\text{angle}}, 180° - \theta_{\text{angle}}] \text{ or} \\ \phi_{i,j} &\in [180° + \theta_{\text{angle}}, 360° - \theta_{\text{angle}}]\} \, .\end{aligned} \tag{5.39}$$

The interval $[180° + \theta_{\text{angle}}, 360° - \theta_{\text{angle}}]$ takes into account that a rotation about $\boldsymbol{r}_{i,j}$ by $\phi_{i,j}$ is the same as a rotation about $-\boldsymbol{r}_{i,j}$ by $360° - \phi_{i,j}$. An additional benefit of this preprocessing step is that the amount of pairs for further processing decreases and computation time with it.

Exhaustive Search

The idea of this method is to select *pairs* of movements according to the angle between their rotation axes. Let (ij, kl) denote the pair of the two movements (i, j) and (k, l). Then

$$\mathcal{S} = \{(ij, kl) \mid (i,j), (k,l) \in \mathcal{M}', i < j, k < l, ij \neq kl\} \tag{5.40}$$

defines the set of all possible pairs of movement pairs. As selection criterion the scalar product $s_{ij,kl}$ between the rotation axes of two movement pairs (i, j) and (k, l) is used:

$$s_{ij,kl} = |{\boldsymbol{r}_{i,j}}^{\mathrm{T}} \boldsymbol{r}_{k,l}| \, . \tag{5.41}$$

The value of $s_{ij,kl}$ is 1 for parallel rotation axes ($\boldsymbol{r}_{i,j} \parallel \boldsymbol{r}_{k,l}$) and 0 for perpendicular axes ($\boldsymbol{r}_{i,j} \perp \boldsymbol{r}_{k,l}$). After computing $s_{ij,kl}$ for all pairs (ij, kl) in $\mathcal{S}$, the best pairs are selected, i. e., those with the smallest values of $s_{ij,kl}$. If N_{relmov} relative movements should be selected, only the best $N_{\text{relmov}}/2$ pairs of $\mathcal{S}$ have to be selected (assuming that N_{relmov} is even) since each pair (ij, kl) consists of the *two* relative movements (i, j) and (k, l).

The described method is exhaustive since the rotation axes of all pairs of relative movements are compared. The worst case estimate of the time complexity of this approach is $O({N_\text{f}}^4)$. If no relative movements are eliminated during the preprocessing step their total number is $|\mathcal{M}'| = |\mathcal{M}| = N_\text{f}(N_\text{f} - 1)/2$. The total number of *pairs* of relative movements, i. e., the number of elements of $\mathcal{S}$, is then

$$|\mathcal{S}| = \frac{N_\text{f}(N_\text{f}-1)}{2} \cdot \frac{\left(\frac{N_\text{f}(N_\text{f}-1)}{2} - 1\right)}{2} = \frac{1}{8} \cdot ({N_\text{f}}^4 - 2{N_\text{f}}^3 - {N_\text{f}}^2 + 2N_\text{f}) = O({N_\text{f}}^4)\,. \tag{5.42}$$

Apart from the time complexity a further drawback of this approach is the selection of *pairs* of relative movements. One relative movement may be contained in several selected pairs. Since a linear system of equations for solving equation (5.36) is set up based on the selected pairs, it could happen that one relative movement is used more than once, leading to two linearly dependent equations. This would increase the number of equations unnecessarily and thus also computation time.

Selection Based on Vector Quantization

This approach selects an *optimal set* of relative movements rather than selecting a *number of optimal pairs* of relative movements. The selection of rotation axes, which are as non-parallel as possible, is achieved as follows.

The rotation axes of all $|\mathcal{M}'|$ relative movements are normalized to one. As already mentioned, the axis/angle representation is not unique, e. g., a rotation about $\boldsymbol{r}$ by ϕ is the same as a rotation about $-\boldsymbol{r}$ by $360° - \phi$. Without loss of generality those axes $\boldsymbol{r}$ with negative z-coordinate are therefore inverted, i. e., $\boldsymbol{r}' = -\boldsymbol{r}$. This assures that similar rotation axes in the sense of parallelism are represented as similar vectors in 3-D. The resulting vectors lie on the upper half ($z > 0$) of the surface of the 3-D unit sphere. Rotation axes with zero z-coordinate have to be transformed similarly (for details see [Sch06]).

Now, N_relmov cluster centers $\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{N_\text{relmov}-1}$ are computed by *vector quantization* with the LBG algorithm [Lin80], where N_relmov is the number of relative movements to be selected. The cluster centers define a partitioning of $\mathbb{R}^3$. In the case of vector quantization the cluster centers are called *codebook vectors* and each vector $\boldsymbol{x} \in \mathbb{R}^3$ is mapped to a cluster according to the following rule:

$$\boldsymbol{x} \longmapsto \boldsymbol{c}_\kappa, \text{ if } d(\boldsymbol{x}, \boldsymbol{c}_\kappa) < d(\boldsymbol{x}, \boldsymbol{c}_i)\ \forall\, i = 0, \ldots, N_\text{relmov} - 1, i \neq \kappa \tag{5.43}$$

where $d(\cdot, \cdot)$ is a distance measure. Here the Euclidean distance is used.

Normally a codebook vector does not coincide with an element of the input vector set. Therefore, $\mathcal{M}_s$ is obtained by selecting for each codebook vector $\boldsymbol{c}_i$ the rotation axis with the smallest distance to $\boldsymbol{c}_i$.

Since the LBG algorithm optimizes the codebook iteratively, a worst case estimate of the time complexity can only be specified for each iteration step. If no relative movements are eliminated during the preprocessing step, $N_\text{f}(N_\text{f} - 1)/2$ vectors have to be clustered by the LBG algorithm, where the time complexity of each iteration is $O(nN_\text{relmov})$, with n being the number of input vectors and N_relmov the number of codebook vectors. Since N_relmov is usually a constant, $O(N_\text{f}\frac{(N_\text{f}-1)}{2}N_\text{relmov}) = O(N_\text{f}{}^2)$. The number of iterations is controlled by the percentage decrease δ of the quantization error. Here, the iteration process is stopped for $\delta \leq 0.001$.

5.6 Computation of Depth and Confidence Maps

Since the objective is the reconstruction of a DC light field, depth and confidence maps have to be computed, where the already available information, i. e., the endoscope poses for each acquired image, can be used.

The confidence value for each pixel is set to 1 by default. It is only changed for substituting an image degradation using the light field. The value is then set to zero for pixels belonging to the degradation. The value of a pixel is also set to zero if this pixel contains no information. This occurs at the border of the endoscopic image since the endoscope provides a round image while the camera captures a rectangular image. The round image of the endoscope is then surrounded by a black border (e. g., see Figure 5.5, page 99). The black border pixels do not correspond to a light ray, their confidence value is therefore set to zero. The border pixels are identified as follows. A sequence of images of a white surface (sheet of paper) is captured and converted to gray images. For each pixel the minimum value of all images is computed and all pixels below a certain gray-value are defined as border pixels. Here the threshold value 50 is used. Since the images are cropped, the effect is only visible when 5 mm endoscopes are employed, because these endoscopes provide a smaller image, resulting in a visible border after cropping the image.

The computation of depth maps for the reconstruction of light fields using structure-from-motion (see Section 5.3, page 90) has already been described in Section 3.3.2, page 52. The basic concept is to use the reconstructed 3-D points to interpolate the depth value for each pixel. The reconstructed 3-D points are a result of the structure-from-motion algorithm. The technique is straight-forward: all 3-D points $\boldsymbol{w}_i$ whose 2-D projections $\boldsymbol{q}_i$ are visible in a certain image are projected into the image plane. The depth value d_i for these non-discrete 2-D points is known. Motivated by triangular mesh interpolation, for each discrete pixel $\boldsymbol{q}$ the nearest three of these projected points, $\boldsymbol{q}_1, \boldsymbol{q}_2$, and $\boldsymbol{q}_3$, contribute to the interpolated depth value $d(\boldsymbol{q})$:

$$d(\boldsymbol{q}) = \sum_{i=1}^{3} \frac{w_i}{\overline{w}}\, d_i \tag{5.44}$$

with

$$\overline{w} = \sum_{i=1}^{3} w_i \quad \text{and} \quad w_i = \frac{1}{\|\boldsymbol{q}_i - \boldsymbol{q}\| + 1}\,, \tag{5.45}$$

i. e., the weight w_i for the depth value d_i is chosen according to the distance of $\boldsymbol{q}_i$ from $\boldsymbol{q}$.

Depth maps for light fields reconstructed with AESOP or smARTtrack1 are computed analogously. However, since no 3-D surface points are available, such points have to be computed first. Additionally, the interpolation method by searching the three next neighbors for each pixel is very slow. Since light field computation should be as fast as possible so that it can be used during an operation, an alternative interpolation technique and an alternative depth map representation are described in the following. The steps for obtaining depth information for AESOP and smARTtrack1 light fields are:

1. Compute 2-D point correspondences
2. Triangulate 3-D surface points
3. Optimize the result non-linearly (optional)
4. Compute depth

The point tracking algorithm described in Section 5.3, page 90, is employed for the first step. A 3-D point is triangulated from all available 2-D correspondences for this point [Har03]. Figure 5.13 exemplarily shows the triangulation of a 3-D point from two 2-D correspondences, i. e., two views. A least median squares (LMedS) technique [Rou87] is applied to eliminate endoscope pose outliers that occur, for instance, due to merging or occluding markers when using smARTtrack1. Thereby endoscope poses with a too large back-projection error are not used for triangulation. Let $\boldsymbol{P}_0, \ldots, \boldsymbol{P}_{k-1}$ denote the projection matrices corresponding to the k images in which the 2-D projections $\boldsymbol{q}_i$, $i = 0, \ldots, k-1$, of a 3-D point $\boldsymbol{w}$ could be tracked. The steps of the applied LMedS algorithm are:

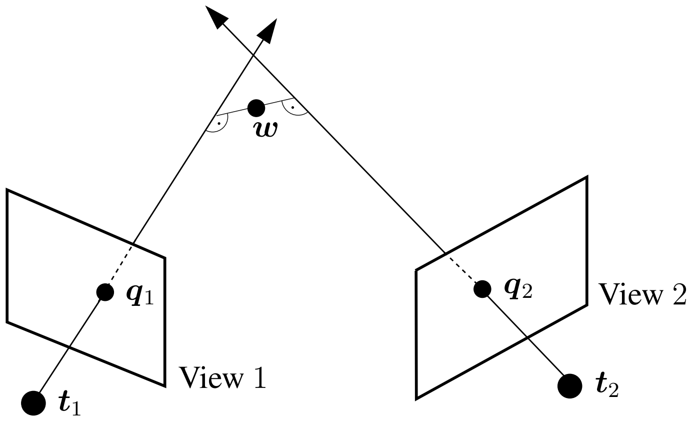


Figure 5.13: Triangulation from two views (adopted from [Tru98]): the 3-D point $\boldsymbol{w}$ is triangulated from the two 2-D correspondences $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. Thereby its distance to the rays from the camera position $\boldsymbol{t}_1$ through $\boldsymbol{q}_1$ and $\boldsymbol{t}_2$ through $\boldsymbol{q}_2$ is minimized, since the two rays will generally not intersect. The rays are computed using the intrinsic and extrinsic camera parameters corresponding to the two views. When more than two views are employed for triangulation, the mean distance of $\boldsymbol{w}$ to all rays is minimized.

1. Randomly select two projection matrices from $\{\boldsymbol{P}_0, \ldots, \boldsymbol{P}_{k-1}\}$ and triangulate the 3-D point $\boldsymbol{w}$ from these two matrices.

2. Calculate the back-projection error $\epsilon_{\mathrm{BPE},i}$ for *all* camera poses $i = 0, \ldots, k-1$: $\epsilon_{\mathrm{BPE},i} = \|\boldsymbol{q}_i - \widehat{\boldsymbol{q}}_i\|$, where $\widehat{\boldsymbol{q}}_i$ is the Euclidean point corresponding to the homogeneous projection of $\boldsymbol{w}$: $\underline{\widehat{\boldsymbol{q}}}_i = \boldsymbol{P}_i \underline{\boldsymbol{w}}$. Determine the median of the computed back-projection errors $\{\epsilon_{\mathrm{BPE},0}, \ldots, \epsilon_{\mathrm{BPE},k-1}\}$.

3. Repeat the first two steps n times and select the best solution, i. e., the minimal median value m^*. The number n is chosen such that the probability of selecting two good projection matrices (inliers) at least once is 99%. To compute n, the assumed probability of outliers p_{out} has to be defined. The 99%-condition can then be expressed as $(1-(1-p_{\mathrm{out}})^2)^n \leq 1 - 0.99$, which yields $n \geq \frac{\ln(0.01)}{\ln(1-(1-p_{\mathrm{out}})^2)}$.

4. Projection matrices are selected according to their back-projection error for the 3-D point corresponding to the triangulation where m^* was obtained. For the final triangulation only those projection matrices are used which result in a back-projection error smaller than a threshold θ_{LMedS}, where $\theta_{\mathrm{LMedS}} = 2.5 \cdot \widehat{\sigma}$ and $\widehat{\sigma} = 1.4826(1 + 5/(k-2)) \cdot \sqrt{m^*}$. Note that $\widehat{\sigma}$ is an estimate of the standard deviation of the measurements [Rou87, page 202], in this case of the point tracking.

Only the assumed probability of outliers p_{out} has to be specified. The larger p_{out}, the more samples have to be drawn until the probability of obtaining at least one sample that does not contain outliers is large enough. Finally, the 3-D point is optimized non-linearly according to the back-projection error using the Levenberg-Marquardt algorithm [Den83]. Although *endoscope pose outliers* are not employed for triangulation, the resulting back-projection error may be large due to *bad point tracking*, e. g., because of bad image quality. Therefore, only points resulting in a back-projection error below a threshold θ_{BPE} are used. The threshold θ_{BPE} allows controlling the quality of the triangulated 3-D points.

The obtained result can be optimized non-linearly: either all endoscope poses and 3-D points are optimized simultaneously by bundle adjustment [Har03], or the endoscope poses are optimized first, followed by re-triangulation of the 3-D points. In the first case, the overall back-projection error is minimized (cf. equation (5.3), page 86). In the second case, each endoscope pose is optimized independently from all other endoscope poses, according to the back-projection error of all 3-D points that are visible in the corresponding image. The disadvantage of bundle adjustment is the long computation time. Especially for usage during minimally invasive operations, the optimization of single endoscope poses followed by re-triangulation is better suited.

Instead of interpolating and storing a depth value for *each pixel* of an image, a *3-D triangular mesh* based on the triangulated 3-D points is employed. In general a triangular mesh consists of a set of triangles where each triangle is described by three vertices. Here, the vertices are triangulated 3-D points. The triangles are computed as follows (see Figure 5.14). The 2-D projections of the triangulated 3-D points are used for a 2-D Delaunay triangulation, where the Delaunay triangulation is performed with the algorithm described in [She97, She02]. The Delaunay triangulation yields a 2-D triangular mesh, i. e., triangles with 2-D vertices. The 3-D triangular mesh is obtained by substituting the 2-D vertices (projections) by the corresponding 3-D points. Thereby a 3-D triangular mesh that represents the depth information (surface geometry) can be computed for each image.

The computed triangular mesh generally contains regions where the sampling is coarse due to the low number of 2-D correspondences that could be computed in this area. Especially close to the borders of the image little or no 2-D correspondences can be computed. This is due to the inhomogeneous illumination of the scene (see Table 2.2, page 22): the amount of light is not sufficient and decreases towards the borders of the image, which makes finding 2-D correspondences more difficult. In order to refine the sampling and, above all, to compute depth values for *all* pixels, the algorithm for computing the 3-D triangular mesh is extended as follows (see Figure 5.15). At the beginning, depth values for some additional 2-D points are interpolated by using equation (5.44). The additional 2-D points are employed together

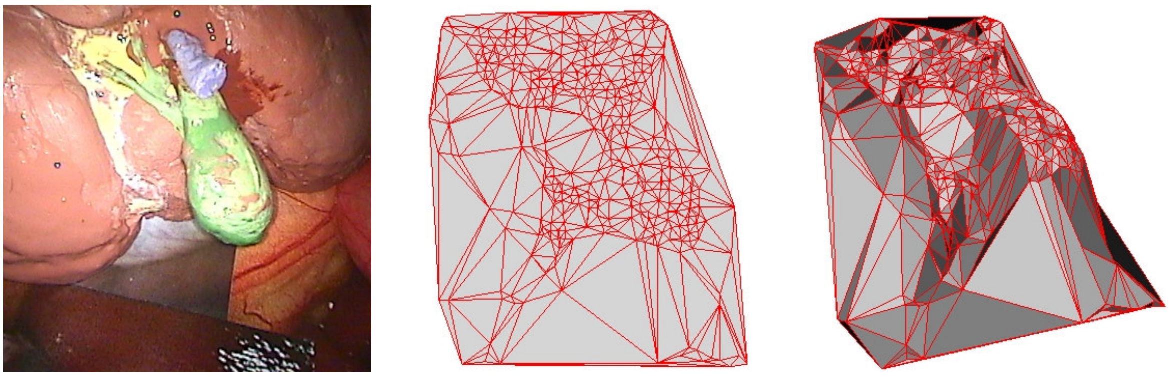


Figure 5.14: The depth map for the original image (left) is represented as a 3-D triangular mesh: the 2-D projections of the triangulated 3-D points are used for a 2-D Delaunay triangulation. The Delaunay triangulation yields a 2-D triangular mesh (middle). The 3-D triangular mesh is obtained by substituting the 2-D vertices by the corresponding 3-D points (right). The size of the original image is 512×512 pixels.

with the 2-D projections of the triangulated points for the 2-D Delaunay triangulation. The additional points are chosen to lie on a grid with fixed spacing, e. g., a 32×32 pixel grid is normally employed. For each grid point $\boldsymbol{q}$ the corresponding 3-D point $\boldsymbol{w}$ is computed from the corresponding interpolated depth value d, and the camera parameters $\boldsymbol{K}$, $\boldsymbol{R}$, and $\boldsymbol{t}$ by

$$\boldsymbol{w} = \boldsymbol{t} + d \cdot \frac{\boldsymbol{R}\boldsymbol{K}^{-1}\underline{\boldsymbol{q}}}{\|\boldsymbol{R}\boldsymbol{K}^{-1}\underline{\boldsymbol{q}}\|}\,, \tag{5.46}$$

where $\underline{\boldsymbol{q}}$ is the homogenous vector corresponding to $\boldsymbol{q}$ and the term $\boldsymbol{R}\boldsymbol{K}^{-1}\underline{\boldsymbol{q}}$ is the direction vector for the "projection ray" in world coordinates. The 3-D point is then obtained by adding d times the normed direction vector to $\boldsymbol{t}$. Without loss of generality it is assumed that the depth values used for interpolation are the Euclidean distances of the 3-D points to the camera center, i. e., $d = d_t$ (cf. Section 3.1.2, page 38).

The advantage of the 3-D triangular mesh is the reduced computational cost. Depending on the grid spacing, only a small number of 2-D depth values and corresponding 3-D points have to be computed, e. g., compared to the conventional approach, only $0.1\,\%$ of interpolations have to be performed when a 32×32 pixel grid is used: $\frac{1}{32 \cdot 32} = \frac{1}{1024} \approx 0.001$. For an image of size 512×512 pixels, only 256 grid points have to be interpolated. If it is assumed that usually about 500 3-D points are triangulated, the input for the 2-D Delaunay triangulation are 756 2-D points for which a Delaunay triangulation can be performed very fast.

Currently, only the unstructured lumigraph renderer (cf. Section 3.2.2, page 46) of the employed software [Vog05a] is capable of using the 3-D triangular meshes directly. For the use with other visualization approaches, conventional depth maps have to be computed. Nevertheless, the

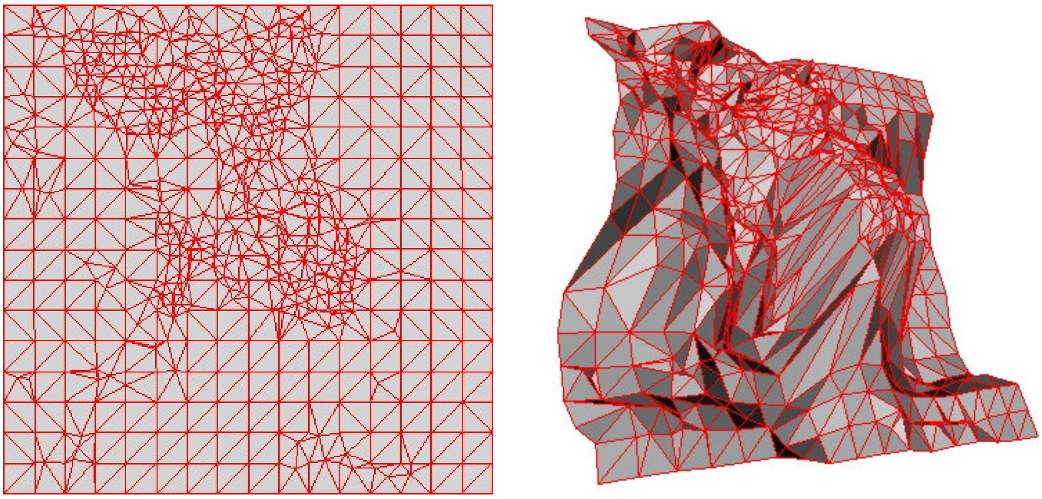


Figure 5.15: 3-D triangular mesh with grid points: before computing the 2-D Delaunay triangulation, 2-D grid points are interpolated and their corresponding 3-D point is computed. The interpolated 2-D grid points together with the 2-D projections are then used for the 2-D Delaunay triangulation (left), leading to the corresponding 3-D triangular mesh (right). The grid spacing used here is 32×32 pixels for an image of size 512×512 pixels.

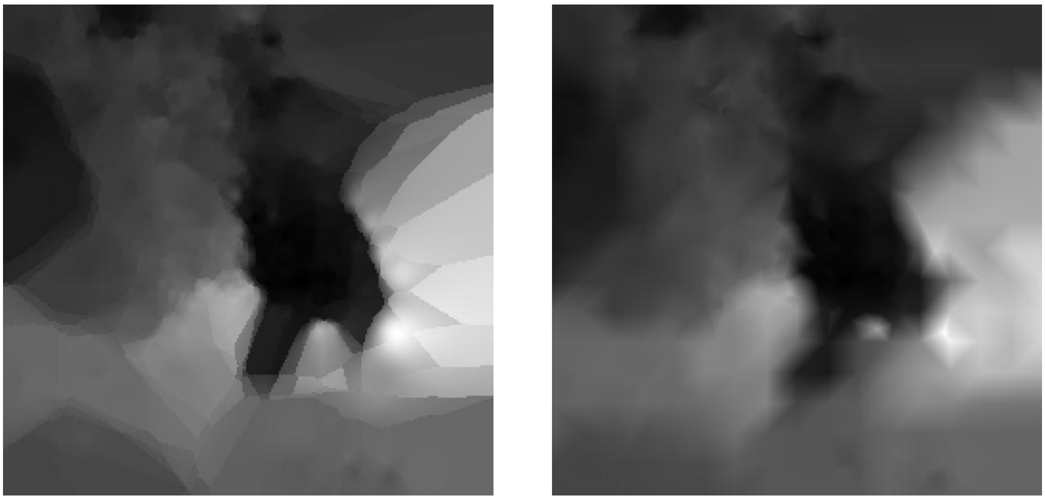


Figure 5.16: Compared to the conventional depth image (left) where the depth value *for each pixel* is interpolated, the depth image employed here is rendered by graphics hardware based on the 2-D triangular mesh (right). A 32×32 pixel grid was used.

triangular mesh can be used to decrease computation time. In this case the 2-D triangular mesh is sufficient. Based on the 2-D mesh the depth value for each 2-D point is assigned to each vertex. This allows exploiting graphics hardware for fast rendering of the corresponding depth image. The difference between this and the conventional approach is the interpolation technique: instead of using the three nearest neighbors for interpolation, the 2-D triangular mesh is used.

The smaller the grid spacing, the more similar are the rendered depth images to the conventional depth images. In Figure 5.16 an example is shown. The advantage of this approach is the reduced computation time.

Having described three approaches for light field reconstruction and the computation of depth and confidence maps, the next section deals with the visualization techniques that are employed.

5.7 Light Field Visualization

The employed software and hardware are only shortly summarized in this section as the focus of this thesis is not on methods for light field *visualization*. The visualization techniques developed and implemented in [Hei04, Vog05a] are used. The current software is called *lgf3* (lumigraph framework version 3). *lgf3* provides all visualization techniques described in Section 3.2. The result of all light field reconstruction methods presented in the previous sections is a free form light field. Here, free form light fields are rendered with the unstructured lumigraph approach (see Section 3.2.2, page 46). Additionally, two-plane light fields are used, as they allow very fast (real-time) rendering. Their drawback is the additional computation time that is required for generating the two-plane representation.

The rendered images are either displayed on a conventional computer monitor or on a 3-D monitor (cf. Section 2.5.1, page 25). The latter has the advantage that it is possible to obtain a realistic depth impression. The 3-D autostereoscopic display "C-nt" by SeeReal Technologies GmbH is employed here. The resolution of the display is 1600×1200 pixels (columns $\times$ rows) and it is connected via the digital video interface (DVI). The images are displayed with 60 Hz. The stereo images are transferred vertically interleaved, i. e., every other column corresponds to the left image, the other columns to the right image. Thus, the final horizontal resolution is only half the size of the original resolution, resulting in a final resolution of 800×1200 pixels. Note that this is still more than the resolution of the S-VHS input signal. The two stereo images are separated by using prisms that provide two images according to the observer's eyes. This means that no special glasses are required, which is an advantage for the use during an endoscopic surgery. The "C-nt" model is a (cheap) 3-D monitor where the eyes of the observer are not tracked. This feature is not important since the relative position of monitor and surgeon will remain the same during an endoscopic surgery.

The applied software can also render dynamic light fields. For these light fields a slider allows moving through the temporal dimension whereas spatial movements are still performed by moving the mouse (see [Sch04b]). The temporal slider switches between the reconstructed static light fields.

5.8 Comparison

This section summarizes the main advantages and disadvantages of the three approaches for light field reconstruction.

The structure-from-motion approach has the advantage that no additional device is necessary. Its disadvantages are the computation time and the restrictions on obtaining a usable result: no movement inside the scene, smooth camera movement during recording of the image sequence, structure in the scene for point tracking, and good illumination conditions. Even if the scene is static, the accuracy of the extrinsic camera parameters depends on the accuracy of point tracking, i. e., bad image quality due to noise or image degradations as well as movement in the scene reduces the accuracy. Another important drawback for the practical use is the sensitivity of the algorithms: a slight change of the tracking or reconstruction parameters may result in a completely different light field reconstruction. It usually takes some time to find a parameter set that leads to a good result for the sequence at hand.

Even if the mentioned prerequisites are only partly fulfilled, it is possible to reconstruct a light field with the two other approaches. The reconstruction using AESOP or smARTtrack1 has two main advantages. Firstly, the reconstruction is very fast since the endoscope's pose can be computed in real-time. Secondly, even if the conditions for point tracking are not good, a light field can be reconstructed. The scene conditions only influence the quality of the depth maps which are computed based on point tracking. Thus, only the success of error correction during rendering according to the available depth information is influenced by the scene conditions.

An advantage of AESOP is that it is possible to keep the endoscope steady when capturing an image. Thus, interlacing artefacts only occur when objects in the scene move (fast). The disadvantage of this method is that the capturing of the images takes longer. The use of AESOP has two main disadvantages: firstly, AESOP was not manufactured to provide accurate pose information. The endoscope plug is not tight enough which leads to large endoscope pose errors (cf. Table 7.12, page 164). Secondly, AESOP can only be moved when the endoscope is fixed at a certain point, usually the keyhole, which complicates hand-eye calibration. The disadvantages of the developed hand-eye calibration technique for AESOP (cf. Section 5.4.2, page 97) are that the length of the endoscope has to be measured by hand and that two steps are required to obtain the hand-eye transformation $\boldsymbol{T}_{\mathrm{H,E}}$. However, a lot of effort would be necessary in order to apply more sophisticated hand-eye calibration techniques, e. g., a sterilizable patient model would have to be constructed, where the problem of obtaining usable data for these methods remains, due to the movement restrictions of AESOP.

The advantages of smARTtrack1 compared to AESOP are the higher accuracy (cf. Table 7.12, page 164, and Table 7.17, page 168), the simpler and more accurate hand-eye calibration, and the easier handling. The latter is easier since only a target is attached to the endoscope, and the surgeon does not have to use a robot arm if he does not want to, whereas the developed targets allow using AESOP and smARTtrack1 together. The main disadvantage of smARTtrack1 is the line-of-sight requirement which is inherent to all optical tracking systems: the pose of the endoscope can only be computed when the target is visible by at least two cameras.

Finally, the three approaches are compared with respect to their suitability for augmented reality. The drawback of the structure-from-motion approach is that camera pose and scene geometry are only known up to an unknown scalar factor (cf. Section 3.3.2, page 51). This is irrelevant for 3-D visualization of the operation site, but it complicates augmenting the light field by registration and fusion with CT/MRI data, because the scale factor has to be estimated. Furthermore, only when the pose of the endoscope is available in real-time, *2-D augmented reality* can be provided, i. e., overlaying registered CT data over the endoscopic *live* image. An advantage of AESOP is the fixation of the robot arm at the operating table. Thereby the relative position of patient and AESOP remains constant when the operating table is moved. A previously computed registration with CT data will still be valid which is not the case when using smARTtrack1, unless the used cameras are also fixed at the operating table.

5.9 Summary

This chapter described three methods for reconstructing a light field of the operation site: using structure-from-motion techniques (Section 5.3), using the robot arm AESOP (Section 5.4), and using the optical tracking system smARTtrack1 (Section 5.5). One assumption is made for all three methods: the intrinsic camera parameters are constant (Section 5.2). These parameters are estimated in advance by a camera calibration technique. Three preprocessing steps precede each approach (Section 5.1): the original image is de-interlaced, undistorted, and cropped. A new representation for depth maps in terms of a 3-D triangular mesh was presented in Section 5.6. The corresponding 2-D triangular mesh is used for fast rendering of depth maps that approximate the conventional ones. Section 5.7 summarized the employed visualization techniques. The computed free form DC light fields are rendered with the unstructured lumigraph approach. Additional computation time has to be spent on the generation of two-plane light fields which allow real-time rendering. Finally, Section 5.8 compared the three approaches for light field reconstruction.

Regarding the advantages and disadvantages of the developed methods for light field reconstruction (Section 5.8), the following is proposed in order to obtain a high quality light field: preprocess the images as described in Section 5.1, compute the intrinsic camera parameters as described in Section 5.2, use smARTtrack1 for determining the endoscope's pose with one of the *double decker* targets, compute depth maps as described in Section 5.6, and use the unstructured lumigraph renderer for visualization. In order to obtain stereoscopic 3-D perception, employ a 3-D monitor or an HMD.

This chapter described two components of a typical medical augmented reality system (cf. Section 2.5, page 24): pose determination (Sections 5.4 and 5.5) and visualization (Section 5.7). The reconstructed light field is perfectly suited for 3-D visualization. The following chapter will deal with the remaining components: virtual data, registration, and fusion.

Chapter 6

Augmented Reality: Registration and Fusion of 3-D Data with Light Fields

Providing augmented reality requires a registration of virtual data with the device through which the reality is observed. In the case of endoscopic surgery, the reality, i. e., the operation site, is viewed through an endoscope. Here, CT data are employed as virtual data for augmentation. Of course, any other 3-D imaging data could also be used, for instance MRI or 3-D ultrasound.

This chapter describes a new method for markerless (*intrinsic*) rigid 3-D/3-D registration (cf. Section 2.5.4, page 30) of a rigid endoscope with CT data. This technique may be used for any endoscopic surgery where rigid endoscopes are employed, a DC light field can be reconstructed by using a pose determination system, and a rigid registration is sufficient. For those cases where a non-rigid registration is more appropriate, the obtained rigid registration may be employed as initialization. The presented method is based on reconstructing a DC light field of the operation site. The reconstructed scene geometry is then used for registration. The registration allows providing 2-D live augmented reality, where the live image of the endoscope is augmented, as well as 3-D augmented reality, where the 3-D model of the operation site, i. e., the light field, is augmented. The augmentation is achieved by fusing correctly rendered views of the CT data with the 2-D live image and the light field, respectively. The benefit of augmented reality is that important anatomical structures (e. g., vessels) are completely visible even if the structures are not or only partly visible in the endoscopic image.

Especially in medical applications markers or calibration patterns are usually employed for rigid registration, e. g., see [DB01, Sch01a, Sch03a, Vog04c, Vog04d]. Markers, also known as *fiducial markers*, can easily be identified in the CT data. Thus their 3-D position is known. A 2-D/3-D registration is then performed by capturing an image where at least three markers are visible. At least three markers are required in order to determine the six unknown parameters of the registration transformation: each marker yields two equations. Finally, the endoscope pose relative to the CT data, i. e., the registration transformation, is computed by camera calibration, based on the 3-D positions of the markers and the corresponding 2-D projections. Fiducial markers are either attached to the skin of the patient or screwed into the patient's bones. The latter provides very accurate registration but is invasive and includes a small risk of infection and damage to the underlying tissue [Sch03a]. In contrast to this, markers attached to the skin are scarcely invasive but may move several millimeters due to the elasticity of the skin. Concerning the disadvantages of both marker types for registration, a markerless technique for registration would be useful.

Exemplarily, the developed methods for registration and fusion (Sections 6.3 and 6.4) are demonstrated for a laparoscopic cholecystectomy, i. e., the minimally invasive removal of the gall bladder. Nevertheless, the techniques may be used for any other endoscopic surgery. The anatomical structures relevant for a laparoscopic cholecystectomy are described in Section 6.1. Since it is currently not common to perform a CT scan for every patient whose gall bladder is removed, a small collection of CT datasets (Section 6.2) allows choosing a suitable one for a specific patient when none is available. Thereby it is assumed that the relevant anatomy differs only slightly when the CT data is selected according to age, gender, height, and weight of the patient. Naturally it would be better to obtain 3-D data from the patient who is operated. For the future it is to be expected that this is the case for every patient, since more and more often 3-D imaging data are acquired and also non-ionizing 3-D imaging technologies like 3-D ultrasound or MRI could be used. Additionally, as mentioned before, the methods presented in this chapter are not restricted to laparoscopic cholecystectomies, and for other minimally invasive interventions, e. g., thymus resection or adrenal gland resection, a CT scan is routinely performed for each patient.

6.1 Important Anatomical Structures

Figure 6.1 illustrates the relevant anatomy for a laparoscopic cholecystectomy. The gall bladder is located directly below the liver. Cystic artery and cystic duct are ligated and cut during the operation. Therefore, and because they are moved before the dissection starts, they are not used for augmented reality. The four most interesting anatomical structures are hepatic vein (vena portae hepatis), hepatic artery (arteria hepatica propria), and bile duct (ductus choledochus and ductus hepaticus), which are all located very close to the dissection area but must not be

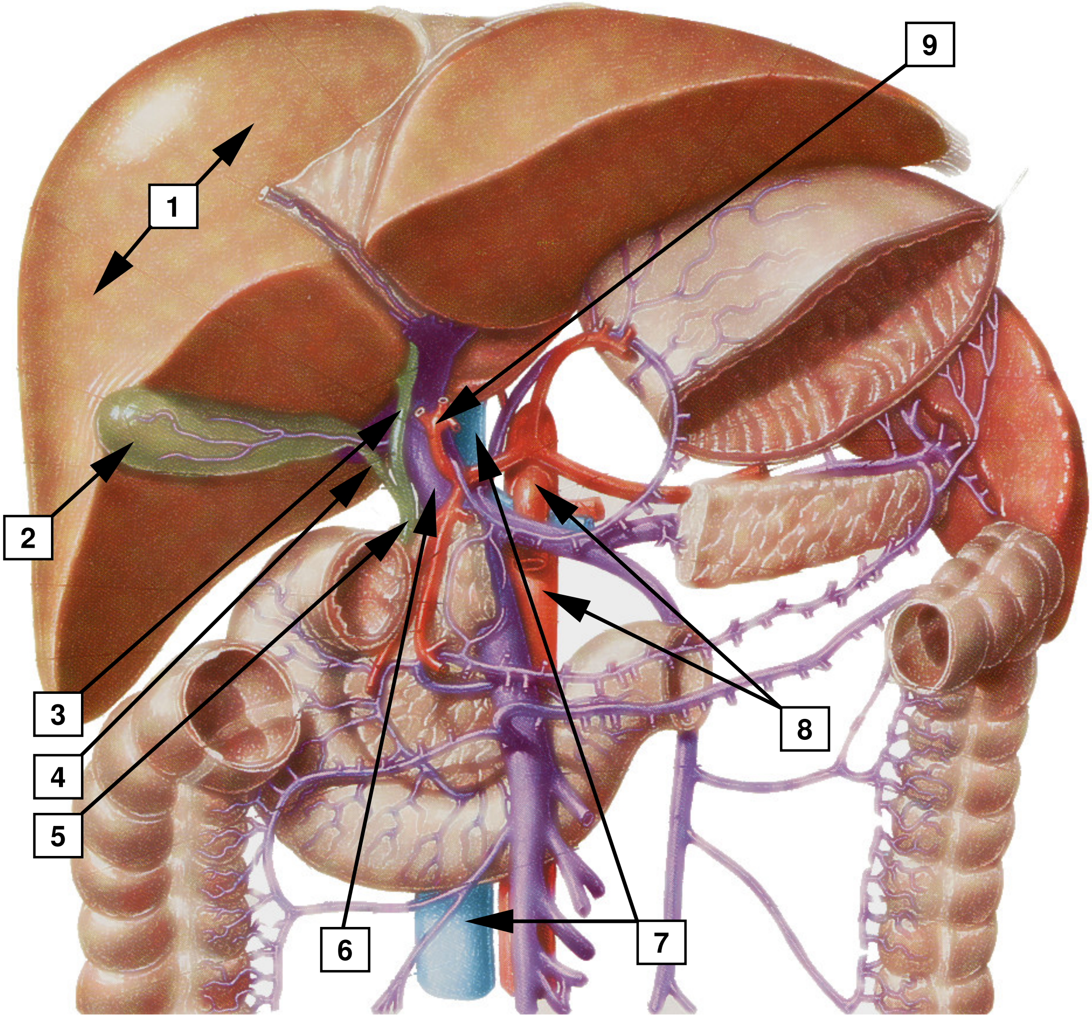


Figure 6.1: The anatomy around the gall bladder: the gall bladder (2) is located below the liver (1). Cystic duct (4) and cystic artery (not visible) are ligated and cut during the operation. Three vessels are located very close to the dissection area and must therefore not be injured during the dissection: bile duct, which can be separated into ductus choledochus (5) and ductus hepaticus (3), vena portae hepatis (6), and arteria hepatica propria (9). The two biggest vessels in the abdomen are located a little further away: vena cava inferior (7) and aorta abdominalis (8). Image adapted from [Sob04].

injured. This also applies to the two biggest vessels in the abdomen, vena cava inferior and aorta abdominalis, which are located only a few millimeters away.

Although gall bladder and liver are moved during the operation, the poses of hepatic artery and bile duct remain (relatively) constant since both structures are connected to the backplane of the abdominal cavity. The hepatic vein is more flexible. However, its interesting part, which must not be injured, runs parallel to hepatic artery and bile duct. Therefore, this part of the hepatic vein remains as fixed as hepatic artery and bile duct.

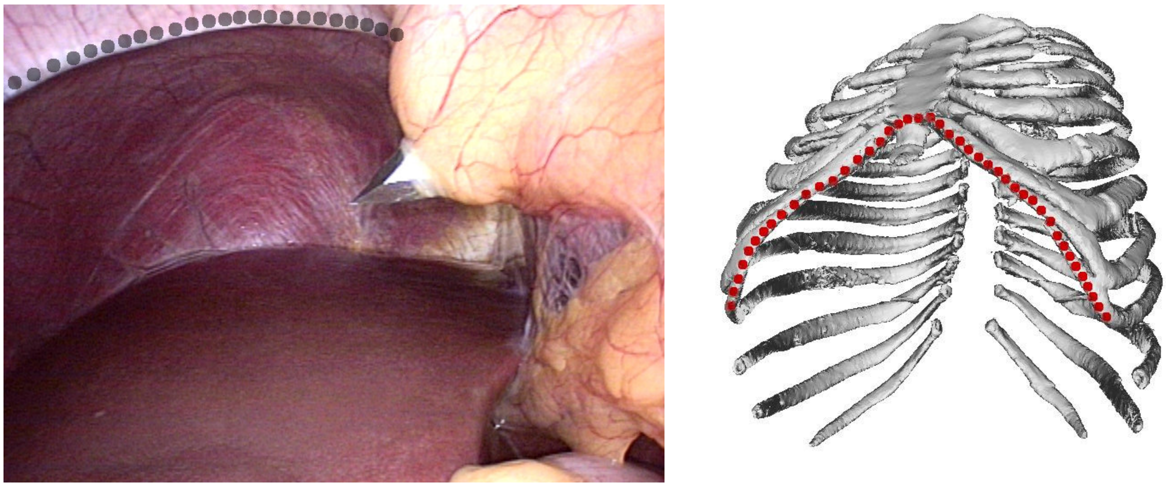


Figure 6.2: The surface of the costal arch can be seen through the endoscope during a laparoscopic cholecystectomy (left) and it can be easily segmented in a CT dataset (right). The costal arch is marked with a dotted line in both images. The CT image (right) is based on the VOXEL-MAN dataset [Höh00].

The registration technique which will be described in Section 6.3 is based on the selection of corresponding anatomical points (landmarks). The landmarks have to be selected in the CT data as well as in the light field. Which anatomical landmarks are used is not important for the technique itself. The only requirement is that an anatomical landmark does not move and has to be observable by the endoscope, i. e., it has to be located on the surface. Possible structures suitable for a laparoscopic cholecystectomy were discussed with surgeons of the Department of Surgery of the University of Erlangen-Nuremberg, where laparoscopic cholecystectomies are performed routinely. The costal arch was identified to be suitable for registration: in contrast to the organs and tissues in the abdomen it does not move. The costal arch can be easily identified and segmented in a CT dataset and its surface can be seen through the endoscope (see Figure 6.2). Since it is assumed that the interesting vessels do not move, their position relative to the costal arch is fixed.

In the laboratory, no complete patient model is available. The patient is simulated by a box with holes. The holes are covered with artificial skin that allows making small incisions and introducing the endoscope through these incisions (cf. Section 7.3, page 138). A liver/gall bladder model consisting of silicone (cf. Figure 7.2, page 138) is employed to simulate the abdominal anatomy. An artificial costal arch is not available. Liver and gall bladder are therefore the important structures for experiments in the laboratory. They are also employed for registration. The registration technique presented in Section 6.3 is universally applicable and the anatomical structures used for registration can be chosen arbitrarily. The choice is only important for augmented reality, where it is necessary that such information is used for registration and fusion that is not deformed and is not moved during the operation. In the laboratory the liver/gall bladder model is kept still which makes it possible to use it for registration.

6.2 Anatomical Database

The large CT database of MeVis [MeV05a] is accessible as part of a research cooperation. For instance, MeVis offers a service for segmentation of CT datasets for liver surgery, especially for split-liver transplantations. During recent years they have collected a large amount of segmented datasets (about 800 at the end of 2004). Naturally not all anatomical structures that are important for a laparoscopic cholecystectomy are segmented and the quality of the datasets is quite different. Therefore, only high quality datasets are selected where unsegmented anatomical structures, especially interesting vessels and the costal arch, can be segmented additionally. The segmentation is performed semi-automatically by MeVisLab [MeV05b]. MeVisLab offers the possibility of region growing and live wire contour segmentation which allows a fast segmentation provided that the contrast between the interesting structure and its surroundings is large enough. The goal is a segmentation of the following anatomical structures: costal arch, aorta, vena cava inferior, hepatic artery, portal vein, liver, gall bladder, and bile duct system. Sometimes not all structures can be segmented, e. g., the contrast for a segmentation of the bile duct system is often not large enough. The datasets are anonymized. Only gender, age, height, and weight of the patient are stored.

Another source for CT datasets is the Institute of Radiology, University of Erlangen-Nuremberg. There, suitable datasets can be acquired during routinely performed CT scans. These datasets are also segmented semi-automatically, either in co-operation with the Department of Neurosurgery of the University of Erlangen-Nuremberg or with MeVisLab.

Apart from the described CT datasets, a very exact anatomical model of the human body, the “VOXEL-MAN” [Höh00], is also part of the anatomical database. VOXEL-MAN is based on images of 770 cryotom slices and corresponding CT images. A large number of anatomical structures (about 220) were segmented by hand. Very small structures like small vessels were modeled as polygons that were fitted to the anatomy. The following structures of the VOXEL-MAN dataset are used: costal arch, liver, gall bladder, cystic duct, bile duct, aorta, vena cava inferior, common hepatic artery, proper hepatic artery, and portal vein. Bile duct, common hepatic artery, and proper hepatic artery are polygon models.

For real operations, costal arch, aorta, vena cava inferior, hepatic artery, portal vein, and main bile duct are important. Liver and gall bladder are employed under laboratory conditions.

6.3 Estimation of Registration Transformation Parameters

In the case of endoscopic surgery, two coordinate systems have to be registered: the coordinate system of the endoscope and the coordinate system of the CT data. A rigid registration is employed here, i. e., the transformation from one coordinate system to the other is expressed by a rotation matrix $\boldsymbol{R}_{\mathrm{reg}} \in \mathbb{R}^{3\times3}$ and a translation vector $\boldsymbol{t}_{\mathrm{reg}} \in \mathbb{R}^{3}$, where a 3-D point $\boldsymbol{x}_{\mathrm{ct}}$ given in CT coordinates is transformed to the corresponding 3-D point $\boldsymbol{x}_{\mathrm{endo}}$ given in endoscope coordinates by

$$\boldsymbol{x}_{\mathrm{endo}} = \boldsymbol{R}_{\mathrm{reg}}\boldsymbol{x}_{\mathrm{ct}} + \boldsymbol{t}_{\mathrm{reg}} \, . \tag{6.1}$$

A scaling factor does not have to be determined as the CT data allows computing the *real* 3-D voxel positions and *real* endoscope poses, which are obtained using a pose determination system, yield *real* 3-D information.

The computation of the registration parameters is separated into two steps: coarse and fine registration (cf. Section 2.5.4, page 30). The coarse registration is performed based on a selection of at least three corresponding 3-D points in each coordinate system. For fine registration, all available points are employed, i. e., the 3-D scene geometry of the DC light field and the triangular mesh of the segmented CT data.

The following two sections (Sections 6.3.1 and 6.3.2) describe the applied method for 3-D point selection for coarse registration. Section 6.3.3 then deals with the computation of $\boldsymbol{R}_{\mathrm{reg}}$ and $\boldsymbol{t}_{\mathrm{reg}}$.

6.3.1 Selection of 3-D Points in the Endoscope Coordinate System

Both methods for reconstructing DC light fields using a pose determination system (see Chapter 5) generate depth information based on reconstructed 3-D surface points. The visualization of the 3-D points is implemented using OpenGL. Since these points represent the scene geometry, it is possible for the surgeon to select landmarks according to the visible shape. The OpenGL selection mechanism allows for selecting 3-D points with the mouse.

In addition to the 3-D scene points, a DC light field contains texture information, namely the captured images. This information can be employed to simplify landmark identification. It is assumed that a 3-D triangular mesh is available for each image (cf. Section 5.6, page 114). One or more selected images can then be overlaid onto the 3-D triangular mesh by texture mapping. This process is supported by OpenGL where only the 2-D/3-D correspondences have to be specified. These correspondences are known since the 3-D triangular mesh was generated by performing a Delaunay triangulation of the tracked 2-D points and by assigning the corresponding 3-D point to each vertex. Now it is possible to select either 3-D points or triangles, which are both supported. When triangles are selected, the center of gravity of the three vertices is defined as the selected point. The smaller the employed grid for the 3-D triangular mesh, the more accurate is the sampling of the scene geometry by triangles. Thus, more accurate landmarks can be selected when triangles are employed for selection.

The advantage of using texture information is that it is possible to select landmarks in regions where no 2-D points could be tracked, provided that the interpolation in this area is accurate enough. This is especially useful for approximately planar surface parts, e. g., the surface of the liver, because the selection is then not restricted to the computed 3-D points. This increases the number of potentially available landmarks.

6.3.2 Selection of 3-D Points in the CT Data Coordinate System

It is assumed that a segmented CT dataset is available, e. g., one contained in the anatomical database described in Section 6.2. In order to select 3-D points in the CT data coordinate system, triangular meshes of the segmented structures are computed. These triangular meshes are then employed for selection, where either 3-D points (vertices) or triangles can be selected. Again, OpenGL is employed for rendering the triangular meshes, and when a triangle is selected, the center of gravity of the vertices is used as the selected point.

The *marching cubes* iso-surface algorithm [Lor87] is employed for computing the triangular meshes. An iso-surface defines a surface in $\mathbb{R}^3$ that contains all points $\boldsymbol{x} \in \mathbb{R}^3$ for which

$$g(\boldsymbol{x}) = c\,, \tag{6.2}$$

where $c \in \mathbb{R}$ is a constant and $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ an arbitrary real-valued function. In the case of CT data g is a discrete function where $g(\boldsymbol{x})$ is the gray-value (density) of the CT dataset at voxel $\boldsymbol{x}$. To compute an approximation of an iso-surface, the CT dataset is regarded as a grid where each voxel corresponds to a grid point. The algorithm processes logical *cubes* created from eight voxels: four are selected from slice k and four from slice $k + 1$ (see Figure 6.3). The basic idea is to visit all cubes that are intersected by the iso-surface and to define triangles approximating the iso-surface for this cube. Thereby the algorithm *marches* from cube to cube. At first, binary values are assigned to all grid points: the value is 1 if the gray-value of the corresponding voxel

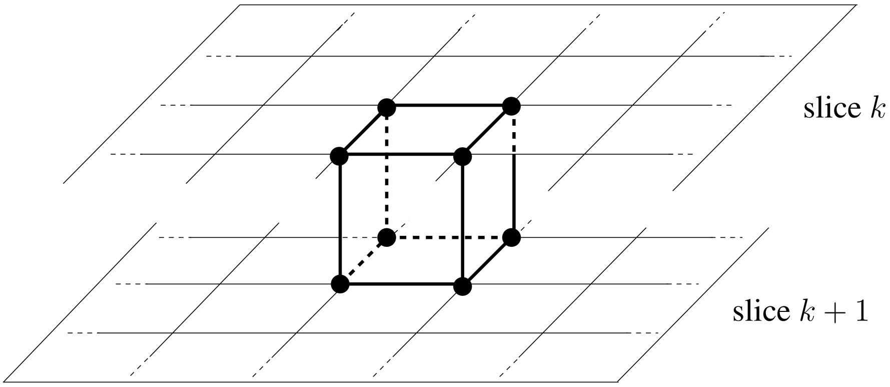


Figure 6.3: Marching cubes algorithm: a logical cube is created from eight voxels of a CT dataset, four each from two adjacent slices k and $k+1$. A binary value is assigned to each grid point: 1 if the gray-value of the corresponding voxel exceeds or equals the iso-surface constant c, 0 otherwise. The iso-surface intersects those cubes where at least two grid points have different values.

exceeds or equals c, zero otherwise. The iso-surface intersects those cubes where at least two grid points have different values. For eight grid points per cube and two states for each grid point there are $2^8 = 256$ ways an iso-surface can intersect the cube. This number can be reduced to 14 by identifying symmetric intersections. All 14 cases are shown in [Lor87]. Since the iso-surface intersects one or more edges of the cube somewhere between two grid points, the exact intersection point for each edge is computed by linear interpolation. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two voxels with $g(\boldsymbol{x}_1) < c \leq g(\boldsymbol{x}_2)$. The intersection point $\boldsymbol{x}_\mathrm{s}$ of the iso-surface with the edge between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is then determined by:

$$\boldsymbol{x}_\mathrm{s} = \boldsymbol{x}_2 + \frac{g(\boldsymbol{x}_2) - c}{g(\boldsymbol{x}_2) - g(\boldsymbol{x}_1)}(\boldsymbol{x}_2 - \boldsymbol{x}_1)\,. \tag{6.3}$$

At the most, twelve intersection points have to be computed for each cube. Based on these points, up to four triangles are defined. The resulting triangular mesh approximates the iso-surface $g(\boldsymbol{x}) = c$.

Since the voxels of the segmented anatomical structures generally do not contain the same gray-value, the algorithm cannot be applied directly in order to obtain a triangular mesh of the segmented structures. At first, the segmentation mask is used to generate a binary CT dataset, where voxels belonging to segmented structures are set to one and the others to zero. The iso-surface to be approximated is then $g(\boldsymbol{x}) = 1$. For this special case the interpolation of the intersection points is unnecessary: if the "iso-surface" intersects a cube, the intersection point is always the grid point with the value 1. Thus, the resulting triangular mesh is very accurate. The Amira software package [Ami05] employed here uses a standard implementation of the marching cubes algorithm. Since the triangular meshes are currently computed offline and not during an operation, a reduction of computation time by omitting the interpolation step is not necessary. Besides, only a slight reduction is to be expected.

The following options of Amira are utilized here: subsampling the dataset before applying the marching cubes algorithm, reducing the number of triangles of the generated mesh, and smoothing the generated mesh. Unfortunately, details on the implementation of the algorithms are not described in the Amira manual. A $2 \times 2 \times 2$ subsampling is employed here, i. e., each new voxel with double side length is computed from eight original voxels. Then the marching cubes algorithm is applied as described above. Afterwards, the number of triangles is reduced by 50 %. The reduction is achieved by collapsing edges, where an error criterion is minimized. Details on the error criterion are not given. Finally, the triangular mesh is smoothed by iteratively shifting its vertices. Each vertex $\boldsymbol{v}$ is shifted towards the mean value $\boldsymbol{v}_{\mathrm{m}}$ of its neighbors. Two parameters control the smoothing process: the number of iterations N_{iter} to be performed and a coefficient $\rho \in [0, 1]$ which specifies the amount of movement for each iteration, where the new vertex $\boldsymbol{v}'$ is obtained by:

$$\boldsymbol{v}' = \boldsymbol{v} + \rho(\boldsymbol{v}_{\mathrm{m}} - \boldsymbol{v}) \, . \tag{6.4}$$

6.3.3 Computation of Registration Transformation Parameters

The rigid transformation from CT coordinates (virtual data) to endoscope coordinates (reality) is expressed by a rotation matrix $\boldsymbol{R}_{\mathrm{reg}} \in \mathrm{I\!R}^{3\times3}$ and a translation vector $\boldsymbol{t}_{\mathrm{reg}} \in \mathrm{I\!R}^3$, where each 3-D point $\boldsymbol{x}_{\mathrm{ct}}$ in CT coordinates is transformed to the registered point $\boldsymbol{x}_{\mathrm{ct,reg}}$ in endoscope coordinates by

$$\boldsymbol{x}_{\mathrm{ct,reg}} = \boldsymbol{R}_{\mathrm{reg}} \boldsymbol{x}_{\mathrm{ct}} + \boldsymbol{t}_{\mathrm{reg}} \, . \tag{6.5}$$

The registration parameters $\boldsymbol{R}_{\mathrm{reg}}$ and $\boldsymbol{t}_{\mathrm{reg}}$ are computed in two steps: the coarse registration with the parameters $\boldsymbol{R}_{\mathrm{c}}$ and $\boldsymbol{t}_{\mathrm{c}}$ transforms $\boldsymbol{x}_{\mathrm{ct}}$ to $\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg}}$ by:

$$\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg}} = \boldsymbol{R}_{\mathrm{c}} \boldsymbol{x}_{\mathrm{ct}} + \boldsymbol{t}_{\mathrm{c}} \, . \tag{6.6}$$

The final point $\boldsymbol{x}_{\mathrm{ct,reg}}$ is then obtained by performing a fine registration, i. e., by transforming $\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg}}$ with the parameters $\boldsymbol{R}_{\mathrm{f}}$ and $\boldsymbol{t}_{\mathrm{f}}$:

$$\begin{aligned} \boldsymbol{x}_{\mathrm{ct,reg}} &= \boldsymbol{R}_{\mathrm{f}} \widetilde{\boldsymbol{x}}_{\mathrm{ct,reg}} + \boldsymbol{t}_{\mathrm{f}} \\ &= \boldsymbol{R}_{\mathrm{f}} (\boldsymbol{R}_{\mathrm{c}} \boldsymbol{x}_{\mathrm{ct}} + \boldsymbol{t}_{\mathrm{c}}) + \boldsymbol{t}_{\mathrm{f}} \\ &= \underbrace{\boldsymbol{R}_{\mathrm{f}} \boldsymbol{R}_{\mathrm{c}}}_{\boldsymbol{R}_{\mathrm{reg}}} \boldsymbol{x}_{\mathrm{ct}} + \underbrace{\boldsymbol{R}_{\mathrm{f}} \boldsymbol{t}_{\mathrm{c}} + \boldsymbol{t}_{\mathrm{f}}}_{\boldsymbol{t}_{\mathrm{reg}}} \, . \end{aligned} \tag{6.7}$$

The computation of $\boldsymbol{R}_{\mathrm{c}}, \boldsymbol{t}_{\mathrm{c}}$ and $\boldsymbol{R}_{\mathrm{f}}, \boldsymbol{t}_{\mathrm{f}}$ will be described in the following.

The coarse registration is performed based on the selected corresponding 3-D points. Let $\mathcal{C}$ denote the set of all points that can be selected in the triangular mesh generated from the segmented CT dataset, i. e., vertices or points computed from selected triangles. Let $\mathcal{E}$ denote the set of all points available in the DC light field, and $\mathcal{C}_{\mathrm{s}} \subset \mathcal{C}$ and $\mathcal{E}_{\mathrm{s}} \subset \mathcal{E}$ denote the sets of selected corresponding 3-D points (cf. Sections 6.3.1 and 6.3.2), where at least three correspondences have to be available, i. e., $|\mathcal{C}_{\mathrm{s}}| = |\mathcal{E}_{\mathrm{s}}| \geq 3$. Let the 3-D point $\boldsymbol{x}_{\mathrm{ct},i} \in \mathcal{C}_{\mathrm{s}}$ correspond to the 3-D point $\boldsymbol{x}_{\mathrm{endo},i} \in \mathcal{E}_{\mathrm{s}}$, for $i = 0, \ldots, N_{\mathrm{corr}} - 1$, where $N_{\mathrm{corr}} = |\mathcal{C}_{\mathrm{s}}| = |\mathcal{E}_{\mathrm{s}}|$. The optimal solution for $\boldsymbol{R}_{\mathrm{c}}$ and $\boldsymbol{t}_{\mathrm{c}}$ is then defined by using the sum of squared distances of the corresponding point pairs as error measure:

$$(\boldsymbol{R}_{\mathrm{c}}, \boldsymbol{t}_{\mathrm{c}}) = \underset{(\boldsymbol{R},\boldsymbol{t})}{\operatorname{argmin}} \sum_{i=0}^{N_{\mathrm{corr}}-1} \| \boldsymbol{x}_{\mathrm{endo},i} - \boldsymbol{R} \boldsymbol{x}_{\mathrm{ct},i} - \boldsymbol{t} \|^2 \, . \tag{6.8}$$

In [Lor95] four methods for the computation of the optimal solution according to equation (6.8) are compared: involving the SVD of a correlation matrix, involving orthonormal matrices, involving unit quaternions, and involving dual quaternions. The SVD algorithm, developed by Arun et al [Aru87], is employed here since the conclusion of [Lor95] is that "… *the SVD algorithm provides the best overall accuracy and stability* …" (in the presence of noise).

The SVD solution is obtained as follows (see [Lor95]). The rotation component is computed first. By defining the centers of gravity

$$\overline{\boldsymbol{x}}_{\mathrm{ct}} = \frac{1}{N_{\mathrm{corr}}} \sum_{i=0}^{N_{\mathrm{corr}}-1} \boldsymbol{x}_{\mathrm{ct},i} \quad \text{and} \quad \overline{\boldsymbol{x}}_{\mathrm{endo}} = \frac{1}{N_{\mathrm{corr}}} \sum_{i=0}^{N_{\mathrm{corr}}-1} \boldsymbol{x}_{\mathrm{endo},i} \quad , \tag{6.9}$$

a 3×3 correlation matrix $\boldsymbol{A}$ of the two point sets after a transformation to the origin is given as

$$\boldsymbol{A} = \sum_{i=0}^{N_{\mathrm{corr}}-1} (\boldsymbol{x}_{\mathrm{ct},i} - \overline{\boldsymbol{x}}_{\mathrm{ct}})(\boldsymbol{x}_{\mathrm{endo},i} - \overline{\boldsymbol{x}}_{\mathrm{endo}})^{\mathrm{T}} \, . \tag{6.10}$$

The singular value decomposition of the correlation matrix, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$, yields the optimal rotation matrix $\boldsymbol{R}_{\mathrm{c}} = \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}$, i. e.,

$$\boldsymbol{R}_{\mathrm{c}} = \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} = \underset{\boldsymbol{R}}{\operatorname{argmin}} \sum_{i=0}^{N_{\mathrm{corr}}-1} \left\|(\boldsymbol{x}_{\mathrm{endo},i} - \overline{\boldsymbol{x}}_{\mathrm{endo}}) - \boldsymbol{R}(\boldsymbol{x}_{\mathrm{ct},i} - \overline{\boldsymbol{x}}_{\mathrm{ct}})\right\|^2 . \tag{6.11}$$

The optimal translation vector is then obtained by comparing equation (6.11) with equation (6.8):

$$\begin{aligned} & \sum_{i=0}^{N_{\mathrm{corr}}-1} \left\|(\boldsymbol{x}_{\mathrm{endo},i} - \overline{\boldsymbol{x}}_{\mathrm{endo}}) - \boldsymbol{R}(\boldsymbol{x}_{\mathrm{ct},i} - \overline{\boldsymbol{x}}_{\mathrm{ct}})\right\|^2 \rightarrow \min \\ \Leftrightarrow \quad & \sum_{i=0}^{N_{\mathrm{corr}}-1} \left\|\boldsymbol{x}_{\mathrm{endo},i} - \boldsymbol{R}_{\mathrm{c}}\boldsymbol{x}_{\mathrm{ct},i} \underbrace{-\overline{\boldsymbol{x}}_{\mathrm{endo}} + \boldsymbol{R}_{\mathrm{c}}\overline{\boldsymbol{x}}_{\mathrm{ct}}}_{=-\boldsymbol{t}_{\mathrm{c}}}\right\|^2 \rightarrow \min . \end{aligned} \tag{6.12}$$

Therefore, in order to minimize equation (6.8), $\boldsymbol{t}_{\mathrm{c}}$ has to be defined as

$$\boldsymbol{t}_{\mathrm{c}} = \overline{\boldsymbol{x}}_{\mathrm{endo}} - \boldsymbol{R}_{\mathrm{c}}\overline{\boldsymbol{x}}_{\mathrm{ct}} . \tag{6.13}$$

For planar datasets or in the presence of noise, the determinant of $\boldsymbol{R}_{\mathrm{c}}$ can be -1 instead of 1, i. e., $\boldsymbol{R}_{\mathrm{c}}$ performs a reflection rather than a rotation [Lor95]. This is corrected by inverting the last (third) column of $\boldsymbol{V}$, corresponding to the singular value of $\boldsymbol{A}$ that is zero.

For fine registration, a recently published extension of the *iterative-closest-point (ICP)* algorithm [Bes92], the Picky ICP algorithm [Zin03], is employed here. Before the algorithm is applied, the points/vertices of the triangular mesh of the segmented CT dataset are transformed according to equation (6.6). Let $\widetilde{\mathcal{C}} = \{\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i}\}_{i=0}^{|\widetilde{\mathcal{C}}|-1}$ be the set of the transformed points and $\mathcal{E} = \{\boldsymbol{x}_{\mathrm{endo},i}\}_{i=0}^{|\mathcal{E}|-1}$ be the set containing the 3-D points of the DC light field. Similar to coarse registration, the problem is to compute a rotation matrix $\boldsymbol{R}_{\mathrm{f}}$ and a translation vector $\boldsymbol{t}_{\mathrm{f}}$ which produce the best alignment of $\widetilde{\mathcal{C}}$ and $\mathcal{E}$. The difference is that generally $|\widetilde{\mathcal{C}}| \neq |\mathcal{E}|$ and that the solution is found iteratively. The original ICP algorithm can be summarized as follows [Tru99]:

1. Set $k = 1$.

2. Compute the subset $\mathcal{Y} = \{\mathrm{cp}(\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},0}), \ldots, \mathrm{cp}(\widetilde{\boldsymbol{x}}_{\mathrm{ct},|\widetilde{\mathcal{C}}|-1})\}$, where

$$\mathrm{cp}(\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i}) = \underset{\boldsymbol{x}_{\mathrm{endo}} \in \mathcal{E}}{\operatorname{argmin}} \|\boldsymbol{x}_{\mathrm{endo}} - \widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i}\| \tag{6.14}$$

is the closest point to $\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i}$ in $\mathcal{E}$.

3. Compute a least-squares estimate $\boldsymbol{R}_{\mathrm{f},k}$ and $\boldsymbol{t}_{\mathrm{f},k}$ that aligns $\widetilde{\mathcal{C}}$ and $\mathcal{Y}$. Since $|\widetilde{\mathcal{C}}| = |\mathcal{Y}|$, the estimate is obtained by the SVD algorithm that was used for coarse registration.

4. Apply $\boldsymbol{R}_{\mathrm{f},k}$ and $\boldsymbol{t}_{\mathrm{f},k}$ to $\widetilde{\mathcal{C}}$, i. e., update each point $\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i} \in \widetilde{\mathcal{C}}$ by

$$\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i} = \boldsymbol{R}_{\mathrm{f},k}\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i} + \boldsymbol{t}_{,k} \, . \tag{6.15}$$

5. Go to step 2 and set $k = k + 1$, or proceed if one of the following conditions is satisfied:

   - the mean square registration error (MSRE)

   $$\epsilon_{\mathrm{MSRE}} = \frac{1}{|\widetilde{\mathcal{C}}|} \sum_{i=0}^{|\widetilde{\mathcal{C}}|-1} \left\| \widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i} - \mathrm{cp}(\widetilde{\boldsymbol{x}}_{\mathrm{ct,reg},i}) \right\|^2 . \tag{6.16}$$

   is sufficiently small;

   - the MSRE difference between two successive iterations is sufficiently small;
   - the maximum allowed number of iterations has been reached.

6. The final parameters are defined as $\boldsymbol{R}_{\mathrm{f}} = \boldsymbol{R}'_{\mathrm{f},k-1}$ and $\boldsymbol{t}_{\mathrm{f}} = \boldsymbol{t}_{\mathrm{f},k-1}$.

The Picky ICP algorithm makes several extensions to increase robustness in the presence of noise and to reduce computation time. The two most relevant extensions are:

- Erroneous point pairs can be rejected according to the distance of the two points. The threshold for rejection is thereby computed by a LMedS technique [Rou87].

- It is prevented that a point $\boldsymbol{x}_{\mathrm{endo}} \in \mathcal{E}$ is utilized in more than one pair, which can be the case when $\boldsymbol{x}_{\mathrm{endo}}$ is the closest point for several points in $\widetilde{\mathcal{C}}$. Only the pair with the smallest distance is then employed. This is especially useful for partially overlapping datasets.

The rejection of point pairs increases the robustness in the presence of noise, but it slows the convergence of the algorithm and the proof of convergence presented in [Bes92] no longer holds [Zin03]. Due to the rejection of point pairs the MSRE can also increase between two successive iterations. Therefore, the change of the motion parameters $\boldsymbol{R}_{\mathrm{f},k}$ and $\boldsymbol{t}_{\mathrm{f},k}$ is monitored. If this change is too small, the algorithm stops. Finally, a maximum number of iterations should always be specified in order to prevent infinite loops. In [Zin03] it was concluded that the Picky ICP algorithm is as robust as the ICP extension presented in [Tru99], while reducing the computation time.

Although the Picky ICP algorithm already rejects point pairs with too large a distance, it is necessary to introduce another parameter that defines the maximal allowed distance of point pairs. The reason is that it is usually assumed that the two datasets describe the same real data, where the overlap is unknown but rather large. In the case of CT data and light fields this assumption is not true. The triangular mesh of the segmented CT data describes complete 3-D surfaces of anatomical structures whereas the 3-D points of the light field are always located on the surface that is *visible* through the endoscope. Since the CT data is registered to the light field, there will usually always exist a large number of points of the CT data triangular mesh for which it makes no sense to find a correspondence. However, since these points are far away from the visible surface, they can easily be rejected by a manually set threshold. The actual threshold value to be used depends on the accuracy of the coarse registration: the better the coarse registration, the more accurate are the two datasets already aligned and the smaller the threshold can be chosen.

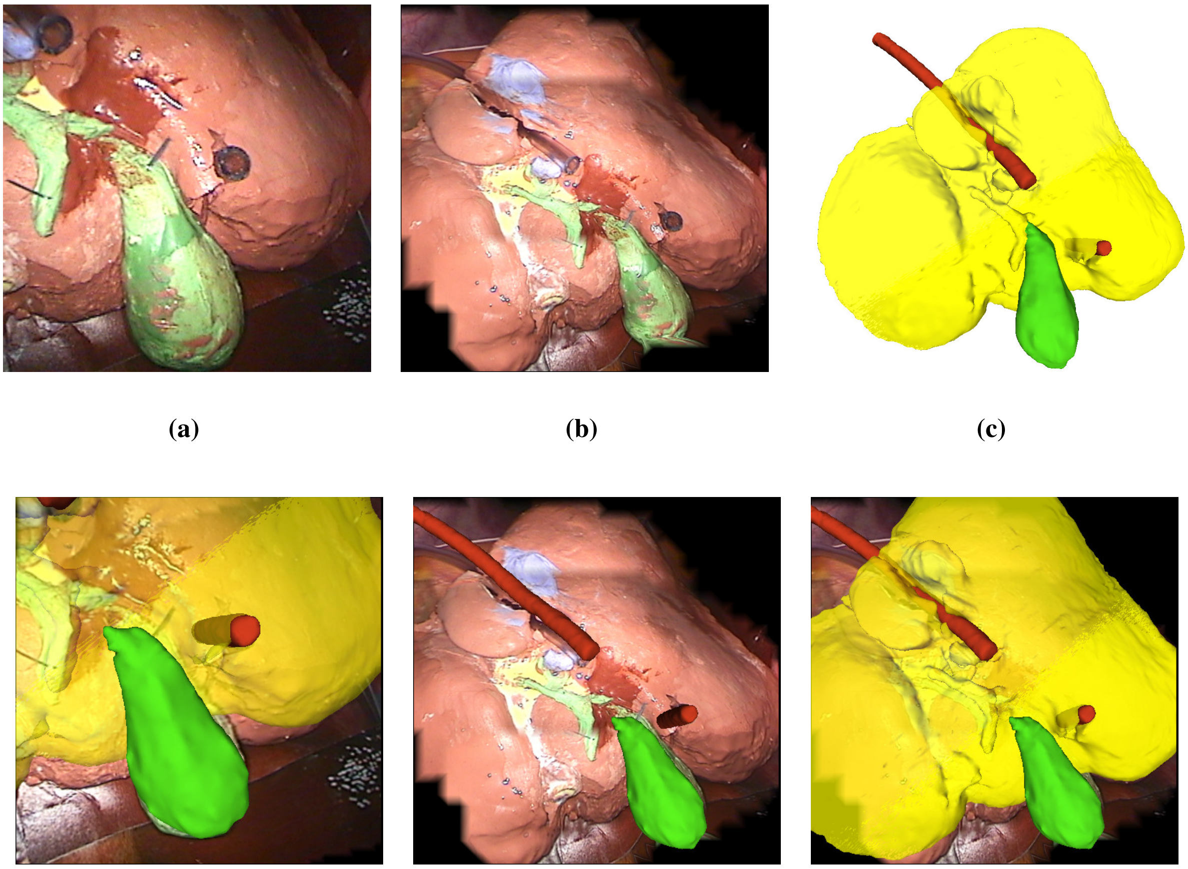
Given the transformation parameters of the coarse and fine registration, $\boldsymbol{R}_{\mathrm{c}}$, $\boldsymbol{t}_{\mathrm{c}}$, $\boldsymbol{R}_{\mathrm{f}}$, and $\boldsymbol{t}_{\mathrm{f}}$, the final transformation parameters of the registration, $\boldsymbol{R}_{\mathrm{reg}}$ and $\boldsymbol{t}_{\mathrm{reg}}$, are computed according to equation (6.7).

Finally, the registration result can be refined by moving the CT data manually in 3-D space so that it fits better to the reconstructed scene. In this case texture mapping is employed for rendering the scene, where depth maps in terms of 3-D triangular meshes and the corresponding images are employed.

6.4 Fusion by Multi-Modality Visualization

Two registered but different types of 3-D models are now available: a DC light field and a triangular mesh of the segmented CT dataset. It is assumed that the vertices of the triangular mesh of the segmented CT dataset are transformed according to the registration parameters. The augmented reality is visualized by rendering both models according to the presently given pose of the observer/camera. The rendered CT data triangular mesh is then overlaid onto the rendered light field image. Since the reality (the light field image) is occluded by the CT data, the CT data should be displayed transparently. With OpenGL this can be implemented easily. The degree of transparency, including no transparency at all, can be set by the physician according to his preferences.

Apart from the described 3-D augmented reality visualization, 2-D live augmented reality



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 6.4: Augmented reality example: (a) original image of a silicone model of liver and gall bladder that contains two tubes that simulate vessels, (b) image rendered from the corresponding DC light field, (c) triangular mesh of the segmented CT dataset of the silicone model, where the liver is shown in yellow, the gall bladder in green, and the tubes in red (d) 2-D augmented reality: the original image is overlaid by a rendered image of the triangular mesh, (e,f) 3-D augmented reality: the DC light field image is overlaid with the gall bladder and the tubes (e) and additionally with the liver (f). The benefit of augmented reality can be seen when the original image (a) is compared to the augmented images (d)-(f): the tube/vessel can be seen completely and not only the part that is visible in the original endoscopic image.

can also be provided when a pose determination system is employed. Once the registration is performed, the current pose of the endoscope and the corresponding view of the segmented CT dataset are known. Therefore, the triangular mesh can be rendered and overlaid onto the 2-D live image. This is particularly useful when the registration is performed at the beginning of an operation. The 2-D live image can then be augmented during the remaining operation. Figure 6.4 shows an example of 2-D and 3-D augmented reality.

For displaying the 3-D augmented reality on a 3-D monitor, a stereo image pair, i. e., two images, have to be rendered for each 3-D model. When only a projection is regarded on a conventional monitor, one image of each model is sufficient.

6.5 Summary

This chapter described an approach for providing 3-D and 2-D augmented reality based on the reconstruction of a DC light field and a rigid 3-D/3-D registration with segmented CT data. After generating a triangular mesh from the segmented CT data with the marching cubes algorithm, the registration is performed in two steps: for coarse registration corresponding points are selected manually, for fine registration the Picky ICP algorithm is utilized.

Exemplarily important anatomical structures for a laparoscopic cholecystectomy were considered. A database of segmented CT datasets allows providing augmented reality even when no CT dataset of the patient is available: a suitable one is then chosen according to gender, age, height, and weight of the patient, where it is assumed that the anatomy differs only slightly. For research purposes this method is sufficient to show the potential and benefit of augmented reality in endoscopic surgery. The presented techniques are not restricted to laparoscopic cholecystectomies. For other minimally invasive operations, e. g., thymus resection or adrenal gland resection, a CT scan is routinely performed for each patient and thus available for augmented reality.

Chapter 7

Experiments and Evaluation

This chapter describes experiments and evaluations of the developed methods for computer assisted endoscopic surgery that were presented in Chapters 4 to 6. Figure 7.1 depicts the experimental setup in the laboratory using the robot arm AESOP and the optical tracking system smARTtrack1. A box is utilized in the laboratory as an artificial patient. The lid of the box contains holes which are covered with artificial skin. This allows making small incisions and introducing the endoscope and surgical instruments. The inner walls of the box are lined with printed color images that were acquired during minimally invasive operations, namely cholecystectomies. As it is possible to open the lid of the box, arbitrary objects can be put inside the "patient". A silicone model of the liver/gall bladder of a corpse is employed to simulate

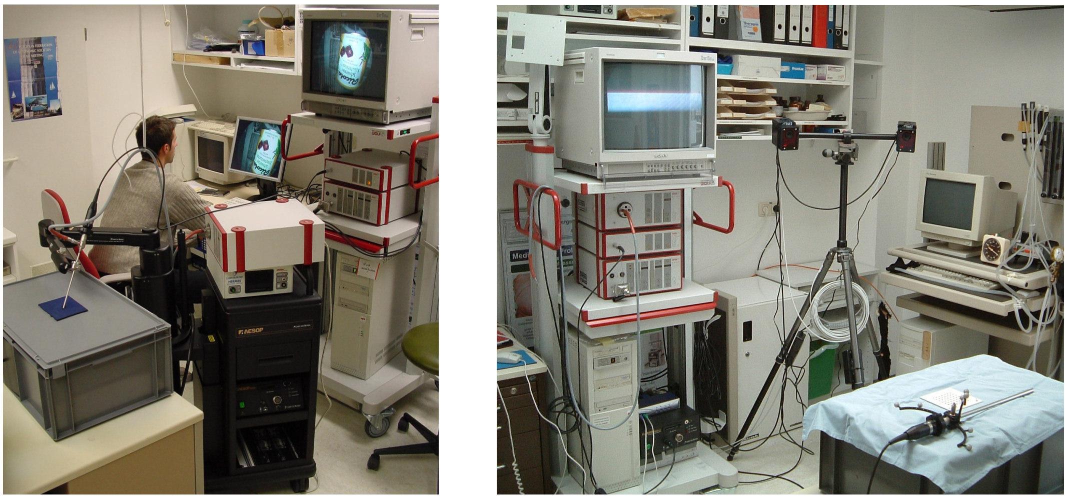


Figure 7.1: Experimental setup in the laboratory using the robot arm AESOP (left) and the optical tracking system smARTtrack1 (right). Additionally, the video-endoscopic system is visible in both images as well as the box that is utilized as an artificial patient (cf. Figure 7.2).

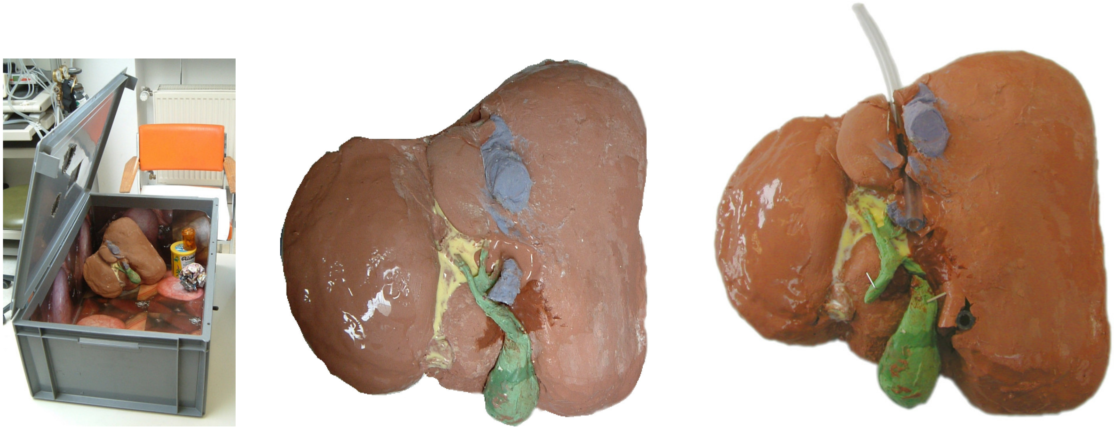


Figure 7.2: A box is used as an artificial patient (left). The lid of the box contains holes which allow introducing the endoscope and surgical instruments. The inner walls of the box are lined with printed color images of minimally invasive operations (cholecystectomies). The middle and right image show the liver/gall bladder model that is used inside the box to simulate the abdominal anatomy. The model is made of silicone. The image on the right shows the model extended by two tubes which are used to simulate vessels for augmented reality experiments. The big brown structure is the liver, the small green structure is the gall bladder that ends in the cystic duct. The upper blue structure is a part of the vena cava, the lower blue structure is a part of the hepatic artery. The colors were chosen by the person who made the model and only the color of the liver approximately corresponds to the color of a real liver.

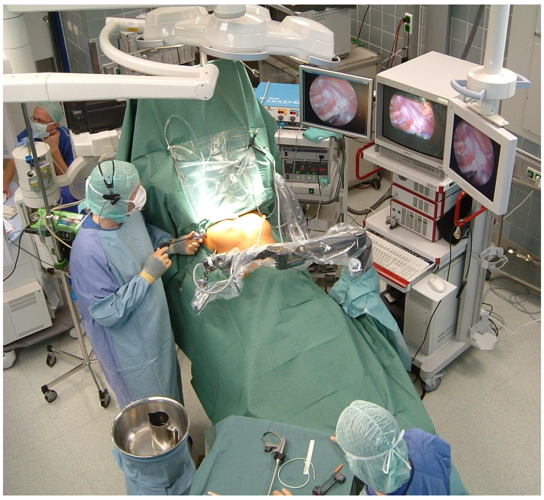


Figure 7.3: Setup in the operating room for light field reconstruction using AESOP during a minimally invasive operation (thoracoscopy). The extended video-endoscopic system is located to the right of the patient, the surgeon stands to the left. In this case an additional third monitor was available.

the abdominal anatomy inside the "patient". Figure 7.2 displays the artificial patient and the liver/gall bladder model. Figures 7.3 and 7.4 show the setup in the operating room using AESOP and smARTtrack1. A modern operating room for minimally invasive surgery contains more

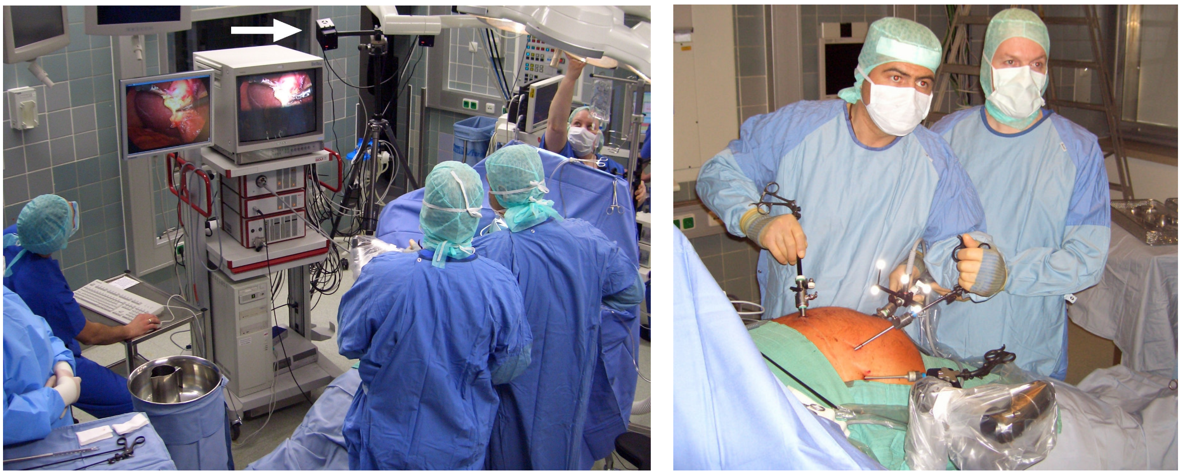


Figure 7.4: Setup in the operating room for light field reconstruction using smARTtrack1 during a minimally invasive operation (cholecystectomy). The left image shows the arrangement of the extended video-endoscopic system and smARTtrack1 on its tripod (marked by an arrow). Only the left of the two cameras of smARTtrack1 is visible. Two surgeons stand at the operating table and the author of this thesis is sitting at the keyboard. The image on the right shows the two surgeons in action, where the endoscope with the attached target is moved by the surgeon on the right side, below the arm of the surgeon on the left side.

than one monitor for displaying the image of the endoscope (cf. Figure 2.4, page 20). These already available monitors can also be used for displaying the processed and rendered images. The first monitor then displays the original image, the second one the enhanced and augmented live-image, and the third one the augmented light field visualization. Up to now, the 3-D monitor has only been used in the laboratory, but the second monitor of the extended video-endoscopic system can easily be exchanged for the 3-D monitor.

Figure 7.5 illustrates the process of hand-eye calibration of AESOP and smARTtrack1 in the operating room. In both cases the calibration pattern is placed on an unsterile table next to the sterile patient. Since only two images of the calibration pattern are necessary for hand-eye calibration of AESOP (cf. Section 5.4.2, page 100), a symmetric pattern is sufficient. For hand-eye calibration of smARTtrack1, usually 20 images of an asymmetric calibration pattern are acquired according to Tsai's guidelines (cf. Section 5.5.2, page 108). Furthermore, a program is utilized that captures images only when the endoscope is kept still.

After describing the setup in the laboratory and in the operating room, in the following the necessary methodology for the evaluation of the experiments is introduced in Section 7.1. Section 7.2 describes results of real-time endoscopic image enhancement. Static and dynamic light field reconstruction results can be found in Section 7.3. Experiments on the removal of image degradations by light fields are described in Section 7.4. Section 7.5 presents examples of 2-D

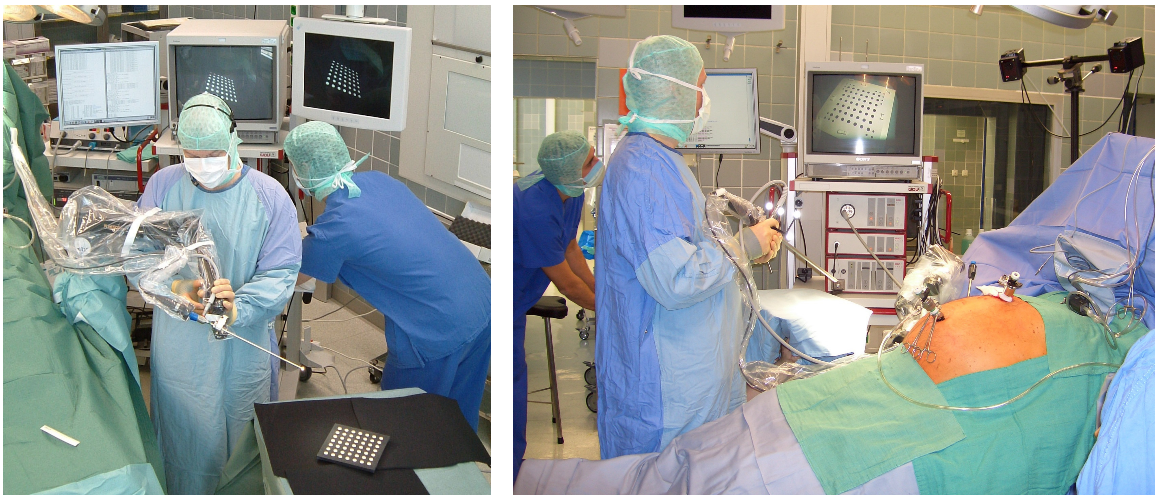


Figure 7.5: Hand-eye calibration of AESOP (left) and smARTtrack1 (right). The calibration pattern is placed on an unsterile table next to the sterile patient. The surgeon moves the endoscope for the acquisition of calibration images.

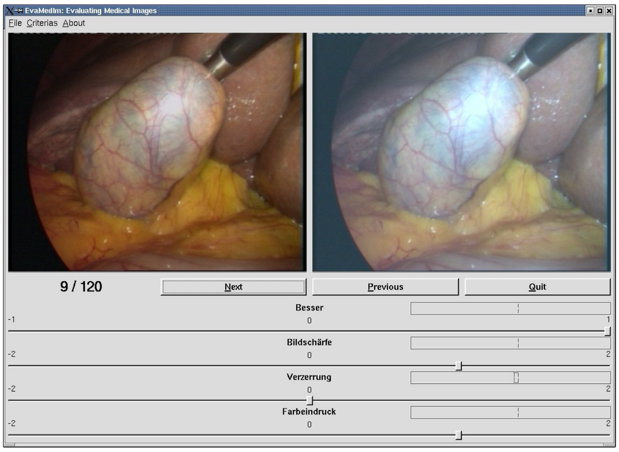
and 3-D augmented reality. Finally, the obtained results are discussed in Section 7.6.

7.1 Methodology for Evaluation

This section introduces the methodology for evaluating the developed methods. The methodology is described here in order to allow for a compact presentation of the results at a later stage. The methods for the evaluation of real-time image enhancement are described in Section 7.1.1. A description of error measures for pose determination based on ground truth data can be found in Section 7.1.2. Section 7.1.3 describes a technique for the objective evaluation of image and light field quality based on ground truth data.

7.1.1 Real-Time Image Enhancement

Single images are used for the subjective evaluation of the real-time image enhancement methods *distortion correction*, *color normalization*, and *temporal filtering*. Image pairs consisting of original endoscopic image and processed image are presented on a computer monitor. The evaluation is carried out in a *double blind setup*, i. e., neither the surgeon nor the person who is carrying out the evaluation (the tutor) know which one is the original image. *EvaMedIm*, a program for "Evaluating Medical Images", was developed for displaying the image pairs randomly in order to ensure double blindness (see Figure 7.6). The surgeon evaluates the displayed image



Besser
-1 0 1

Bildschärfe
-2 0 2

Verzerrung
-2 0 2

Farbeindruck
-2 0 2

Figure 7.6: Screenshot of *EvaMedIm*, the developed program for "Evaluating Medical Images". The original and processed image are displayed simultaneously, where their arrangement, i. e., which one is displayed on which side, is chosen randomly. Beneath the images the evaluation criteria are displayed. A value for each criterion can be set by using the corresponding slider. The program is also capable of displaying movies.

pair by setting a value $v \in \mathbb{Z}$ for each evaluation criterion. In an intermediate step, positive values mean that the image on the right is preferred, negative values that the image on the left is preferred with respect to the evaluated criterion. When no difference between the displayed pair is observable, v is set to zero. Before the evaluation result is stored, the obtained values are transformed such that positive values mean that the *processed* image is preferred and negative values that the *original* image is preferred. For the evaluation of real-time endoscopic image enhancement the range of the evaluation criteria was either $\{-1, 0, 1\}$ or $\{-2, -1, 0, 1, 2\}$. The latter range allows weighting the answer, e. g, $v = 1$ means the processed image is *better* and $v = 2$ means the processed image is *much better*.

It cannot be assumed that the evaluation values are normally distributed. This was verified by the Lilliefor test [Lil67]: the null hypothesis of normal distribution was never rejected at significance level $\alpha = 0.01$. Thus the appropriate test for demonstrating a significant difference between original and processed images is the Wilcoxon signed rank test[1] [Gib03, Wil45]. The Wilcoxon test is a two-sided rank test of the null hypothesis that the data originate from a distribution whose *median* is zero, i. e., "no observable difference". A significant difference between original and processed images is obtained if the null hypothesis is rejected. Given the limited range of the evaluation value, a *mean value* larger than zero indicates that the processed images are significantly better, otherwise the original images are significantly better. The larger the dif-

---

[1] Also known as Mann-Whitney U test

ference the smaller the p-value of the test, i. e., the probability of observing the data by chance if the null hypothesis “median is zero” were true.

7.1.2 Pose Error

An error measure for pose data is necessary in order to judge the accuracy of hand-eye calibration and of the methods for pose determination of an endoscope. It is assumed that ground truth data are available. Here, the poses obtained from camera calibration are used as ground truth data, where the camera calibration algorithm of Section 4.1.2 is applied. Theoretically, it is possible to compute absolute pose errors, but as the origin is defined arbitrarily for camera calibration as well as for the employed pose determination systems, the two coordinate systems would have to be registered first. A possible solution would be to define the poses of the first (or any other) camera to be identical and to transform the computed poses according to this definition. All poses should then be equal and absolute errors could be computed. The disadvantage of this method is that all depends on one camera pose and the error of this pose influences the errors of all other camera poses. Therefore, another method for obtaining pose errors is proposed. The concept is based on the fact that only relative camera poses are relevant for 3-D reconstruction of a scene. Thus, a large number of pose *pairs* is chosen randomly. The pose error for each pair is computed and the mean value of all pairs is defined as the pose error for the whole image sequence.

Similar to hand-eye calibration, the transformation in terms of a rotation matrix and a translation vector between the first and the second pose is used. Without errors, this transformation should be equal for ground truth poses and computed poses. The camera pose of the i-th image defines the transformation from a 3-D world point $\boldsymbol{w}$ to camera coordinates ${}^{\mathrm{c}}\boldsymbol{w}_i$ (cf. equation (3.6), page 36). Using homogeneous coordinates this transformation can be written as

$$
{}^{\mathrm{c}}\underline{\boldsymbol{w}}_i = \underbrace{\begin{pmatrix} \boldsymbol{R}_i{}^{\mathrm{T}} & -\boldsymbol{R}_i{}^{\mathrm{T}}\boldsymbol{t}_i \\ \boldsymbol{0}_3{}^{\mathrm{T}} & 1 \end{pmatrix}}_{=:\boldsymbol{T}_i} \underline{\boldsymbol{w}} \,. \tag{7.1}
$$

Let $\boldsymbol{T}_i$ denote the ground truth pose of the i-th image and $\widehat{\boldsymbol{T}}_i$ the corresponding computed pose. The transformation $\boldsymbol{T}_{i,j}$ from pose i to j is then obtained by:

$$
\boldsymbol{T}_{i,j} = \boldsymbol{T}_j \boldsymbol{T}_i{}^{-1}\,, \text{ where } \boldsymbol{T}_{i,j} := \begin{pmatrix} \boldsymbol{R}_{i,j}^{\mathrm{T}} & -\boldsymbol{R}_{i,j}^{\mathrm{T}}\boldsymbol{t}_{i,j} \\ \boldsymbol{0}_3{}^{\mathrm{T}} & 1 \end{pmatrix} \,. \tag{7.2}
$$

$\widehat{\boldsymbol{T}}_{i,j}$ is computed analogously. $\boldsymbol{T}_{i,j}$ and $\widehat{\boldsymbol{T}}_{i,j}$ are the *relative* transformations from pose i to j and

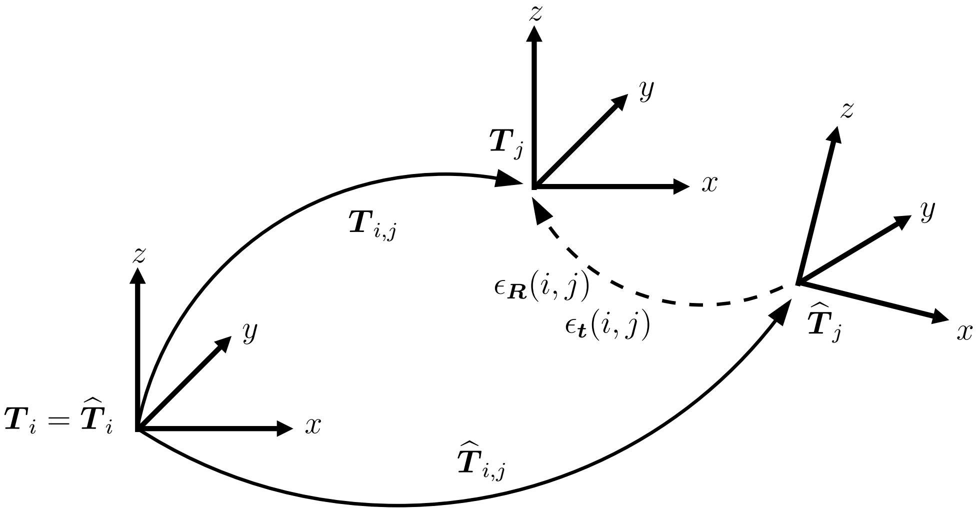

z
y
x
$\boldsymbol{T}_j$
z
y
x
$\boldsymbol{T}_{i,j}$
z
$\epsilon_{\boldsymbol{R}}(i,j)$
y
$\epsilon_{\boldsymbol{t}}(i,j)$
$\widehat{\boldsymbol{T}}_j$
x
$\boldsymbol{T}_i = \widehat{\boldsymbol{T}}_i$
x
$\widehat{\boldsymbol{T}}_{i,j}$

Figure 7.7: The idea for measuring the error of a pose pair (i, j) is sketched. The camera poses are represented by 4×4 transformation matrices $\boldsymbol{T}_i$, $\boldsymbol{T}_j$ for the ground truth camera poses and $\widehat{\boldsymbol{T}}_i$, $\widehat{\boldsymbol{T}}_j$ for the computed camera poses. Without loss of generality $\boldsymbol{T}_i = \widehat{\boldsymbol{T}}_i$ because only the *relative* transformations from pose i to j are important for error computation. Without errors, the ground truth transformation $\boldsymbol{T}_{i,j}$ from pose i to j should be equal to the computed transformation $\widehat{\boldsymbol{T}}_{i,j}$. The difference (dashed line) is the error which can be specified in terms of a rotation error $\epsilon_{\boldsymbol{R}}(i,j)$ and a translation error $\epsilon_{\boldsymbol{t}}(i,j)$.

should be equal. Two types of errors can now be computed:

- translation error and
- rotation error.

The idea for the computation of the errors is sketched in Figure 7.7. The translation error $\epsilon_{\boldsymbol{t}}(i,j)$ of pose pair (i,j) is defined as the norm of the translation difference:

$$\epsilon_{\boldsymbol{t}}(i,j) = ||\boldsymbol{t}_{i,j} - \widehat{\boldsymbol{t}}_{i,j}|| \,. \tag{7.3}$$

It is assumed that $\epsilon_{\boldsymbol{t}}(i,j)$ depends on the distance between the two camera positions: the larger the distance, the larger the error. Thus, a relative translation error $\epsilon_{\boldsymbol{t},\mathrm{rel}}(i,j)$ for the pose pair (i,j) is additionally defined: it is obtained by dividing $\epsilon_{\boldsymbol{t}}(i,j)$ by the norm of the ground truth translation vector $\boldsymbol{t}_{i,j}$:

$$\epsilon_{\boldsymbol{t},\mathrm{rel}}(i,j) = \frac{\epsilon_{\boldsymbol{t}}(i,j)}{||\boldsymbol{t}_{i,j}||} = \frac{||\boldsymbol{t}_{i,j} - \widehat{\boldsymbol{t}}_{i,j}||}{||\boldsymbol{t}_{i,j}||} \,. \tag{7.4}$$

The first idea for computing the rotation error might probably be to compute it analogously to the translation error by using the Frobenius norm of the rotation matrices. But the interpretation of the Frobenius norm of a difference of rotation matrices is difficult. Furthermore, the Frobenius norm of the ground truth rotation is constant[2], which means that it makes no sense to compute a relative rotation error in this way. Therefore, rotation errors are computed by using the axis/angle and the Cardan angle representations. The computation of the rotation axis $\boldsymbol{r}$ and the rotation angle ϕ from the eigenvalues of a rotation matrix was described in Section 5.5.2, page 107. The Cardan angle representation[3] defines a rotation matrix $\boldsymbol{R}$ by multiplying three simple rotation matrices, each one about one of the coordinate axes:

$$\boldsymbol{R} = \boldsymbol{R}_z(\gamma)\boldsymbol{R}_y(\beta)\boldsymbol{R}_x(\alpha)\,, \tag{7.5}$$

where the three rotation matrices define rotations about the x, y, and z-axis by angles α, β, and γ (cf. equations (5.26) to (5.28), page 96). The computation of Cardan angles from a rotation matrix is tricky because one has to take care of a lot of special cases. A detailed description is given in [Sla05].

An error measure for rotation is obtained by firstly computing the “difference” between $\boldsymbol{R}_{i,j}$ and $\widehat{\boldsymbol{R}}_{i,j}$ in terms of a rotation matrix $\boldsymbol{R}_{\text{diff},i,j}$:

$$\boldsymbol{R}_{\text{diff},i,j} = \widehat{\boldsymbol{R}}_{i,j}^{\mathrm{T}}\boldsymbol{R}_{i,j}\,. \tag{7.6}$$

Without errors $\boldsymbol{R}_{\text{diff},i,j}$ would equal the identity matrix. Secondly, the axis/angle and Cardan angle representation is computed from $\boldsymbol{R}_{\text{diff},i,j}$ and $\boldsymbol{R}_{i,j}$. Let $\boldsymbol{r}_{\text{diff},i,j}$ and $\phi_{\text{diff},i,j}$ be the axis/angle representation of $\boldsymbol{R}_{\text{diff},i,j}$, and $\alpha_{\text{diff},i,j}$, $\beta_{\text{diff},i,j}$, and $\gamma_{\text{diff},i,j}$ the corresponding Cardan angle representation. Let $\boldsymbol{r}_{i,j}$ and $\phi_{i,j}$ be the axis/angle representation of the ground truth transformation from pose i to j and $\alpha_{i,j}$, $\beta_{i,j}$, and $\gamma_{i,j}$ the corresponding Cardan angle representation. The axis/angle rotation error $\epsilon_{\boldsymbol{R}}(i,j)$ and the *relative* axis/angle rotation error $\epsilon_{\boldsymbol{R},\text{rel}}(i,j)$ of a pose pair (i,j) are defined as:

$$\begin{aligned} \epsilon_{\boldsymbol{R}}(i,j) &= |\phi_{\text{diff},i,j}| \text{ and} && (7.7)\\ \epsilon_{\boldsymbol{R},\text{rel}}(i,j) &= \frac{\epsilon_{\boldsymbol{R}}(i,j)}{|\phi_{i,j}|} = \frac{|\phi_{\text{diff},i,j}|}{|\phi_{i,j}|}\,. && (7.8)\end{aligned}$$

The Cardan angle rotation errors $\epsilon_{\boldsymbol{R},\text{C},\alpha}(i,j)$, $\epsilon_{\boldsymbol{R},\text{C},\beta}(i,j)$, and $\epsilon_{\boldsymbol{R},\text{C},\gamma}(i,j)$, and the *relative* Cardan

---

[2] The Frobenius norm of any 3×3 rotation matrix is $\sqrt{3}$ (the norm of each column vector is 1).

[3] Sometimes the Cardan and Euler angle representations are confounded with each other, e. g., in [Sla05]. The definition of a rotation matrix based on Euler angles is $\boldsymbol{R} = \boldsymbol{R}_z(\gamma)\boldsymbol{R}_y(\beta)\boldsymbol{R}_z(\alpha)$ [McK91].

angle rotation error $\epsilon_{\boldsymbol{R},\mathrm{rel,C}}(i,j)$ of a pose pair (i,j) are defined as:

$$\begin{aligned}
\epsilon_{\boldsymbol{R},\mathrm{C},\alpha}(i,j) &= \left|\alpha_{\mathrm{diff},i,j}\right|, && (7.9)\\
\epsilon_{\boldsymbol{R},\mathrm{C},\beta}(i,j) &= \left|\beta_{\mathrm{diff},i,j}\right|, && (7.10)\\
\epsilon_{\boldsymbol{R},\mathrm{C},\gamma}(i,j) &= \left|\gamma_{\mathrm{diff},i,j}\right|, \text{ and} && (7.11)\\
\epsilon_{\boldsymbol{R},\mathrm{rel,C}}(i,j) &= \frac{\epsilon_{\boldsymbol{R},\mathrm{C},\alpha}(i,j)+\epsilon_{\boldsymbol{R},\mathrm{C},\beta}(i,j)+\epsilon_{\boldsymbol{R},\mathrm{C},\gamma}(i,j)}{|\alpha_{i,j}|+|\beta_{i,j}|+|\gamma_{i,j}|}\\
&= \frac{|\alpha_{\mathrm{diff},i,j}|+|\beta_{\mathrm{diff},i,j}|+|\gamma_{\mathrm{diff},i,j}|}{|\alpha_{i,j}|+|\beta_{i,j}|+|\gamma_{i,j}|}. && (7.12)
\end{aligned}$$

The relative errors are computed by dividing the sum of rotation angles of $\boldsymbol{R}_{\mathrm{diff},i,j}$ by the sum of angles of the ground truth rotation matrix $\boldsymbol{R}_{i,j}$.

In order to compute the pose errors for an image sequence, an arbitrary number of pose pairs, usually 100 or 1000, is chosen randomly. It makes sense to choose only pairs (i,j) with $i \neq j$ or even with a minimal frame distance $\varDelta_f$, i. e., $|i-j| \geq \varDelta_f$. The pose error for each pair is computed and the mean values $\bar{\epsilon}_{\boldsymbol{t}}, \bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}, \bar{\epsilon}_{\boldsymbol{R}}, \bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}, \bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\alpha}, \bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\beta}, \bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\gamma}$, and $\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel,C}}$ are defined as the pose errors for the image sequence.

7.1.3 Quality of Light Fields and Image Quality

Quality criteria used to evaluate video coding algorithms can also be used to objectively evaluate image and light field quality. The quality of a compressed/processed gray-value image $\widehat{f}$ with respect to the original image f can be measured by the following three criteria [Wan02]:

- **Mean absolute difference (MAD)**

  $$Q_{\mathrm{MAD}} = \frac{1}{N_{\mathrm{r}}N_{\mathrm{c}}}\sum_{y=0}^{N_{\mathrm{r}}-1}\sum_{x=0}^{N_{\mathrm{c}}-1}\left|f(x,y)-\widehat{f}(x,y)\right|, \tag{7.13}$$

  where N_{r} and N_{c} are the number of rows and columns of f.

- **Signal to noise ratio (SNR)**

  $$Q_{\mathrm{SNR}} = 10\log_{10}\frac{\sum_{y=0}^{N_{\mathrm{r}}-1}\sum_{x=0}^{N_{\mathrm{c}}-1} f^2(x,y)}{\sum_{y=0}^{N_{\mathrm{r}}-1}\sum_{x=0}^{N_{\mathrm{c}}-1}\left(f(x,y)-\widehat{f}(x,y)\right)^2}, \tag{7.14}$$

  where the numerator totalizes the *squared signal* (ground truth) and the denominator totalizes the *squared noise*.

- **Peak signal to noise ratio (PSNR)**

$$Q_{\mathrm{PSNR}} = 10\log_{10}\frac{\sum_{y=0}^{N_\mathrm{r}-1}\sum_{x=0}^{N_\mathrm{c}-1}255^2}{\sum_{y=0}^{N_\mathrm{r}-1}\sum_{x=0}^{N_\mathrm{c}-1}\left(f(x,y)-\widehat{f}(x,y)\right)^2} \tag{7.15}$$

$$= 10\log_{10}\frac{255^2}{\frac{1}{N_\mathrm{r}N_\mathrm{c}}\sum_{y=0}^{N_\mathrm{r}-1}\sum_{x=0}^{N_\mathrm{c}-1}\left(f(x,y)-\widehat{f}(x,y)\right)^2}\,. \tag{7.16}$$

The difference to Q_{SNR} is that the numerator, i. e., the sum of the squared signal, is substituted by the *squared peak value* of the signal, i. e., 255^2 in the case of gray-value images. The denominator of equation (7.16) is also known as *mean squared error (MSE).*

The extension to color images is straight forward: $|f(x,y)-\widehat{f}(x,y)|$ is substituted by

$$\frac{1}{3}\left(|I_\mathrm{r}(x,y)-\widehat{I}_\mathrm{r}(x,y)|+|I_\mathrm{g}(x,y)-\widehat{I}_\mathrm{g}(x,y)|+|I_\mathrm{b}(x,y)-\widehat{I}_\mathrm{b}(x,y)|\right)\,,$$

where I_r, I_g, and I_b are the intensity values of the red, green, and blue channel of the color image $\boldsymbol{f}(x,y)=(I_\mathrm{r}(x,y),I_\mathrm{g}(x,y),I_\mathrm{b}(x,y))^\mathrm{T}$. Analogously, $f^2(x,y)$ is substituted by

$$\frac{1}{3}\left({I_\mathrm{r}}^2(x,y)+{I_\mathrm{g}}^2(x,y)+{I_\mathrm{b}}^2(x,y)\right)$$

and $(f(x,y)-\widehat{f}(x,y))^2$ by

$$\frac{1}{3}\left((I_\mathrm{r}(x,y)-\widehat{I}_\mathrm{r}(x,y))^2+(I_\mathrm{g}(x,y)-\widehat{I}_\mathrm{g}(x,y))^2+(I_\mathrm{b}(x,y)-\widehat{I}_\mathrm{b}(x,y))^2\right)\,.$$

Image sequences can now be evaluated by computing the mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ over the whole sequence.

The idea for evaluating light fields has been introduced in [Nie05]. The authors propose to compare the original images with images that are rendered at the original camera poses using the original intrinsic camera parameters. Without errors the original and the rendered image would be identical and the quality of the rendered image can be measured by the criteria described above, where the original image is used as ground truth. In order to evaluate the interpolation and extrapolation properties, the original image is not used for rendering. This idea is extended as follows: an arbitrarily chosen amount of neighboring images is not used for rendering. Thus, the color values of the rendered image have to be interpolated from the remaining images. Apart from the accuracy of the camera parameters, the result depends on the accuracy of the available depth information in the light field.

The proposed method for omitting neighboring images is based on the fact that all evaluations of light field quality described later in this chapter were performed with the unstructured lumigraph rendering approach (cf. Section 3.2.2, page 46). For unstructured lumigraph rendering, weights for each camera are computed according to penalties, and only those cameras with the k smallest penalties are used for rendering the current image. The cameras that must not be used during rendering (the "neighbors") are defined as the l cameras with the smallest penalties. These l cameras are deleted from the list of available cameras *before* the k best suited cameras for rendering are selected from this list. A light field is then evaluated by rendering an image for each original camera pose while only the remaining camera images are used for rendering. The mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ of all images define the quality of the light field. For all light field evaluations presented later, $l = 3$ was used.

7.2 Endoscopic Image Enhancement

The objective of *real-time* endoscopic image enhancement is to grab, process, and display 25 PAL color images per second, i. e., not to exceed 40 msec computation time per image. This objective was achieved (see Table 7.1). Apart from the camera calibration, which has to be performed only once at the beginning of an operation, the slowest algorithm is distortion correction with 37 msec. Strictly speaking, the computation time for grabbing and displaying an image has to be taken into account and added to the computation time of the algorithm, yielding 42.7 msec. This results in 23.4 frames per second (fps) instead of 25 fps. In general, such a small difference cannot be noticed by a human observer. Although each *single* image enhancement method can be provided in real-time, most combinations of the available methods lead to a more or less reduced frame rate. For instance, the fastest combination of all three methods, namely color normalization with step size $k = 10$, temporal filtering with filter size $s = 3$, and distortion correction requires 76.4 msec which corresponds to 13.1 fps. This is quite close to real-time and it can be expected that the next PC generation allows real-time processing for the combination of all three image enhancement methods. Even if horizon equalization is additionally used, the total computation time is only 100.4 msec corresponding to 10.0 fps.

As color normalization, temporal filtering, and distortion correction process the image pixelwise, defining a region of interest (ROI) may further reduce computation time, where a ROI containing only half the number of pixels of the original image will be processed about two times faster.

Method	**Computation Time [msec]**			
Color Normalization (Step k)	34 ($k = 1$)	25 ($k = 10$)		
Temporal Filtering (Size s)	8.7 ($s = 3$)	14 ($s = 5$)	98 ($s = 7$)	180 ($s = 9$)
Distortion Correction	37			
Horizon Equalization	24			
Grabbing	3.5			
Displaying (Zoom Factor z)	2.2 ($z = 1.0$)	2.3 ($z = 1.3$)	2.3 ($z = 1.6$)	2.3 ($z = 1.9$)
Camera Calibration	877			

Table 7.1: Computation times of image enhancement methods on a 3.2 GHz Pentium 4 PC. Mean values computed from processing 100 color images of size 768×576 pixels (PAL) are stated. Since the computation of the rotation matrix $\boldsymbol{R}_{\mathrm{C}}$ for color normalization alone requires 10 msec (= 29 % of 34 msec) the computation time can be reduced if $\boldsymbol{R}_{\mathrm{C}}$ is only computed for every k-th image and only every k-th pixel is used for its computation. For $k = 10$, color normalization requires only 25 msec instead of 34 msec for $k = 1$. Temporal filtering is only possible in real-time for filter sizes $s \leq 5$. Zooming with zoom factors $1 \leq z \leq 1.9$ does not increase computation time which is due to the use of graphics hardware. The computation time for camera calibration is less than one second, where ten images were used. Note that camera calibration has to be done only once at the beginning of an operation.

Camera	$\sigma_{\mathcal{P}}$(red)	$\sigma_{\mathcal{P}}$(green)	$\sigma_{\mathcal{P}}$(blue)
Sony, Firewire	1.9	1.6	1.8
Sony, S-VHS	2.4	2.1	2.8
Wolf Endocam, S-VHS, 5 mm Endoscope	6.6	5.9	9.1
Wolf Endocam, S-VHS, 10 mm Endoscope	7.5	6.0	9.3

Table 7.2: Comparison of the sensor noise of the employed endoscopic camera (Wolf Endocam 5512) and a standard consumer camera (Sony 3-CCD). The sensor noise is specified in terms of the mean standard deviation $\sigma_{\mathcal{P}}(\cdot)$ for each color channel (red, green, blue) for a set of pixels $\mathcal{P}$, where 10 images and an area of size 500×500 pixels was used, i. e., $\sigma_{\mathcal{P}}(\cdot)$ was computed as mean value from the sensor noise $\sigma(\cdot)$ of $|\mathcal{P}| = 250,000$ pixels (cf. equations (4.1) and (4.3), page 57). The sensor noise of the Sony camera depends on the method for data transfer: digital data transfer via the Firewire bus results in a lower sensor noise compared to using the analog S-VHS signal.

Before describing the experiments on the real-time enhancement methods in more detail in the following four sections (Sections 7.2.1 to 7.2.4), the quality of the acquired endoscopic images in terms of the sensor noise is considered. Table 7.2 states the sensor noise of the employed endoscopic camera (Wolf Endocam 5512) and a normal digital video camera (Sony 3-CCD). The sensor noise using the Wolf Endocam 5512 is about three times larger compared to the Sony

Sequence	ϵ_{BPE}	F_{x}	F_{y}	C_{x}	C_{y}	κ_1	κ_2	p_1	p_2
ALFcc 2	0.13	491.5	491.4	339.6	288.1	-0.31	0.15	0.00091	-0.0011
ALFcc 3	0.12	491.5	491.4	345.4	272.6	-0.31	0.15	0.00096	-0.0011
ALFcc 4	0.16	482.4	482.6	348.9	287.4	-0.29	0.14	$-1.4 \cdot 10^{-5}$	-0.0012
ALFcc 5	0.21	498.0	497.8	353.1	287.4	-0.32	0.15	0.0018	0.00067
ART 11	0.18	495.5	495.4	397.1	266.4	-0.33	0.20	0.00099	-0.0028
$\overline{x}$	0.16	491.8	491.7	356.8	280.4	-0.31	0.16	0.00093	-0.0011
σ	0.037	5.93	5.79	23.1	10.2	0.015	0.024	0.00064	0.0012

Table 7.3: Camera calibration results using a 5 mm endoscope. The intrinsic camera parameters $F_{\mathrm{x}}, F_{\mathrm{y}}, C_{\mathrm{x}}, C_{\mathrm{y}}, \kappa_1, \kappa_2, p_1$, and p_2 as well as the back-projection error ϵ_{BPE} over all images are specified for each sequence. Additionally, the mean value $\overline{x}$ and the standard deviation σ of each column are stated.

camera, while there is only a slight difference whether a 5 mm or 10 mm endoscope is used. The Sony camera allows transferring the images digitally by the Firewire bus which further reduces sensor noise. Unfortunately, endoscopic cameras in general do not provide digital data transfer and the Wolf Endocam 5512 is no exception. Thus, the S-VHS signal and a frame grabber card have to be utilized.

7.2.1 Camera Calibration and Distortion Correction

Distortion correction and light field reconstruction rely on camera calibration. Different types of calibration patterns have already been described (see Figure 4.4, page 61). For light field reconstruction with AESOP, the manufactured symmetric 7×7 pattern of white circles on black background was used with a distance of 20 mm between the calibration points. The printable asymmetric 7×7 pattern with a distance of 10 mm between the calibration points was employed for light field reconstruction using smARTtrack1. The reason for not using the same calibration pattern was that the more sophisticated printable pattern and the corresponding calibration algorithm were developed after the experiments with AESOP had been accomplished. Figure 7.8 shows the printable pattern fixed by white magnets to a flat steel plate and examples of intermediate results of the calibration algorithm (cf. Section 4.1.2, page 62).

Results of camera calibration with a 5 mm and a 10 mm endoscope can be found in Tables 7.3 and 7.4. Each result was obtained from a sequence of ten images. The threshold for the circu-

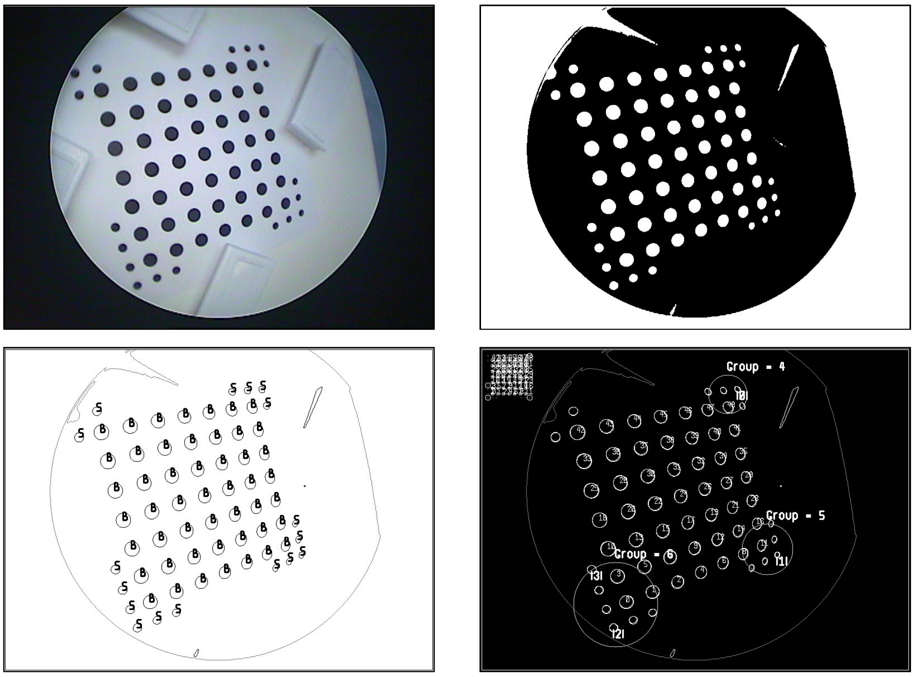


Figure 7.8: Three intermediate results of the camera calibration algorithm are depicted: the original image of the calibration pattern (top left) was converted to a gray-value image, inverted, and then binarized using a threshold value of 124, where gray-value 0 = black and 255 = white (top right). The bottom left image displays the fitted contours together with the computed assignment of small (S) and big (B) circles. Three groups of small circles were found (bottom right image): with four (Group = 4), five (Group = 5), and six small circles (Group = 6).

Sequence	ϵ_{BPE}	F_{x}	F_{y}	C_{x}	C_{y}	κ_1	κ_2	p_1	p_2
`ART 30`	0.13	557.3	553.3	373.4	247.3	-0.17	0.22	0.0062	-0.0024
`ART 32`	0.16	556.6	554.6	374.1	245.1	-0.17	0.22	0.0039	-0.00081
`ART 36`	0.17	552.6	551.5	375.2	249.5	-0.16	0.20	0.0035	-0.0016
`ART 38`	0.18	549.4	548.2	374.9	255.9	-0.17	0.20	0.0045	-0.0025
`ART 50`	0.20	552.2	551.2	374.7	247.4	-0.16	0.19	0.0046	-0.0020
$\overline{x}$	0.17	553.6	551.8	374.5	249.0	-0.17	0.21	0.0045	-0.0019
σ	0.026	3.29	2.42	0.716	4.14	0.0055	0.013	0.0010	0.00069

Table 7.4: Camera calibration results using a 10 mm endoscope. The intrinsic camera parameters $F_{\text{x}}, F_{\text{y}}, C_{\text{x}}, C_{\text{y}}, \kappa_1, \kappa_2, p_1$, and p_2 as well as the back-projection error ϵ_{BPE} over all images are specified for each sequence. Additionally, the mean value $\overline{x}$ and the standard deviation σ of each column are stated.

larity criterion C was set to 15 for the symmetric pattern (`ALFcc` sequences[4]) and to 18 for the asymmetric pattern (`ART` sequences[5]). The number of contour points of valid contours had to be in the interval $[20, \infty)$ pixels and the contour area had to be in the interval $[50, 1000]$ square pixels. In addition to the calibrated intrinsic camera parameters, the obtained back-projection error ϵ_{BPE} over all images is specified (cf. equation (3.36), page 51). In order to get an impression of the variability of camera calibration, the mean value and standard deviation of the five sequences was computed for each intrinsic parameter and the back-projection error. For both endoscope types the mean back-projection error was smaller than 0.2 pixels. The back-projection error measures the fit (in pixels) of the estimated camera model to the data, including the error due to noise. For each sequence the camera head was mounted arbitrarily onto the endoscope optics. This explains the rather large standard deviation for the principal point using the 5 mm endoscope ($\sigma(C_{\mathrm{x}}) = 23.1$ pixels and $\sigma(C_{\mathrm{y}}) = 10.2$ pixels). Interestingly, the effect of this manual mounting was smaller for the 10 mm endoscope ($\sigma(C_{\mathrm{x}}) = 0.7$ pixels and $\sigma(C_{\mathrm{y}}) = 4.1$ pixels). As expected, the standard deviation of the effective focal lengths and the distortion parameters was small, but considering the fact that the effective focal lengths should be *equal*, the differences were rather large (standard deviation approximately $1\,\%$ of the mean value for the 5 mm endoscope and $0.5\,\%$ for the 10 mm endoscope). The distortion parameters of the 5 mm endoscope and the 10 mm endoscope differed mainly in the first radial parameter which was about two times larger for the 5 mm endoscope ($\overline{x}(\kappa_1) = -0.31$ compared to $\overline{x}(\kappa_1) = -0.17$). The conclusion stated in many publications on camera calibration (e. g., see [Hei04, Tru98, Zha96, Tsa87]), that radial distortions represent the main part of divergence to the pinhole camera model and tangential distortions (p_1 and p_2) are negligible, was reproducable.

Concluding this section, Figure 7.9 displays three examples of distortion correction. Noticeable distortions, which can nevertheless be corrected, occurred only when 5 mm endoscopes were used (middle and bottom image).

7.2.2 Color Normalization

Altogether five examples of color normalization are presented. Figure 7.10 shows the first four images (`good`, `blood I`, `blood II`, and `blood III`). The effect of setting the center of the transformed color cluster $\boldsymbol{\mu}'$ manually is visualized in Figure 7.11, where a extremely bad image (`blood IV`) was chosen in order to show the strength of the approach. The computed values of the rotation angle ϕ and the center of the transformed color cluster $\boldsymbol{\mu}'$ for all five images

---

[4] ALFcc is an acronym for AESOP Light Field camera calibration.

[5] ART is an abbreviation of smARTtrack1.

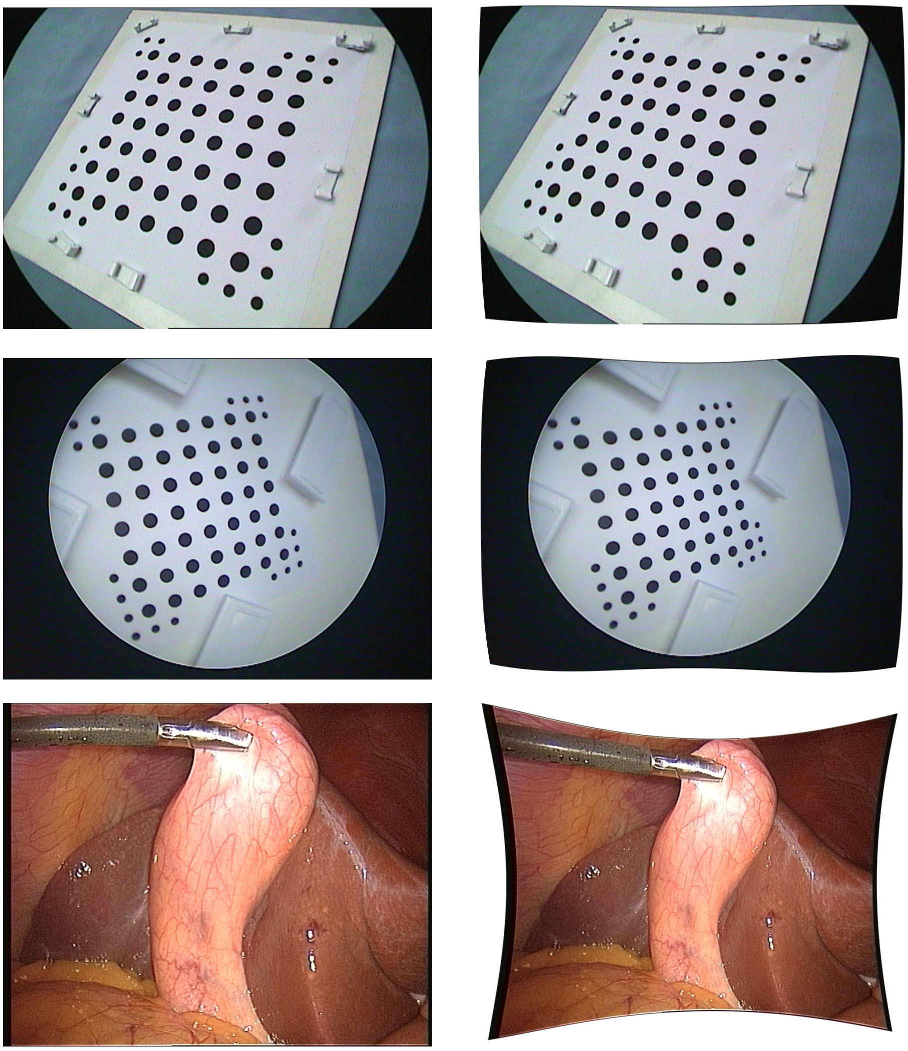


Figure 7.9: Examples of distortion correction. The undistorted images are displayed to the right of the original images. For acquiring the top image a 10 mm endoscope was used. The distortion is barely visible. For the middle and bottom images, a 5 mm endoscope was used. Distortions are clearly visible and can be corrected.

are listed in Table 7.5. As expected, the rotation angle of the good image was smaller than the rotation angles of the bad images.

7.2.3 Temporal Filtering

The implementation of the temporal color median filter described in Section 4.3, page 71, is based on the fast spatial color median filter of the Intel Image Processing Library (IPL). Computation times for the application of color median filters with sizes 3×3 and 5×5 to a PAL color

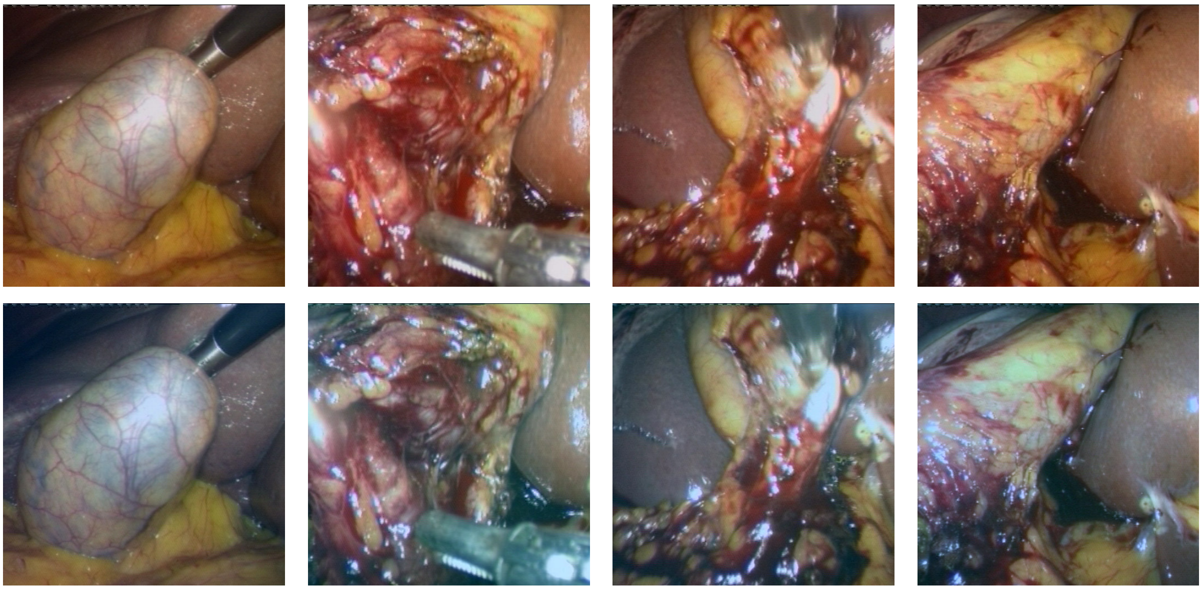


Figure 7.10: Color normalization: the original image is shown on top of the processed image. Four images are shown: a good one (`good`), captured at the beginning of the operation (left), and three bad ones (`blood I`, `blood II`, and `blood III`, from left to right), captured during the operation. The center of the transformed color cluster $\boldsymbol{\mu}'$ was computed by equation (4.41), page 70.

Image	`good`	`blood I`	`blood II`	`blood III`	`blood IV`
Rotation Angle ϕ	0.44	2.0	1.6	8.6	7.4
Cluster Center $\boldsymbol{\mu}'$	94.3	111	98.9	102	112

Table 7.5: Color normalization: the computed values for the rotation angle ϕ [degree] and the center of the transformed color cluster $\boldsymbol{\mu}'$ [gray-value] for the five images `good`, `blood I`, `blood II`, `blood III`, and `blood IV` are stated.

image are stated in Table 7.6. Exemplarily the computation times of two widely used spatial filters, namely 3×3 Sobel and 3×3 Gauß, are also specified. In Section 4.3, two possibilities for implementation of a spatial color median filter were described:

Filter	Sobel 3×3	Gauß 3×3	Median 3×3	Median 5×5
Computation Time [msec]	1.8	2.8	2.7	13

Table 7.6: Computation times of some IPL filters applied to a PAL color image (size 768×576 pixels). The mean value of 100 filtering operations is stated.

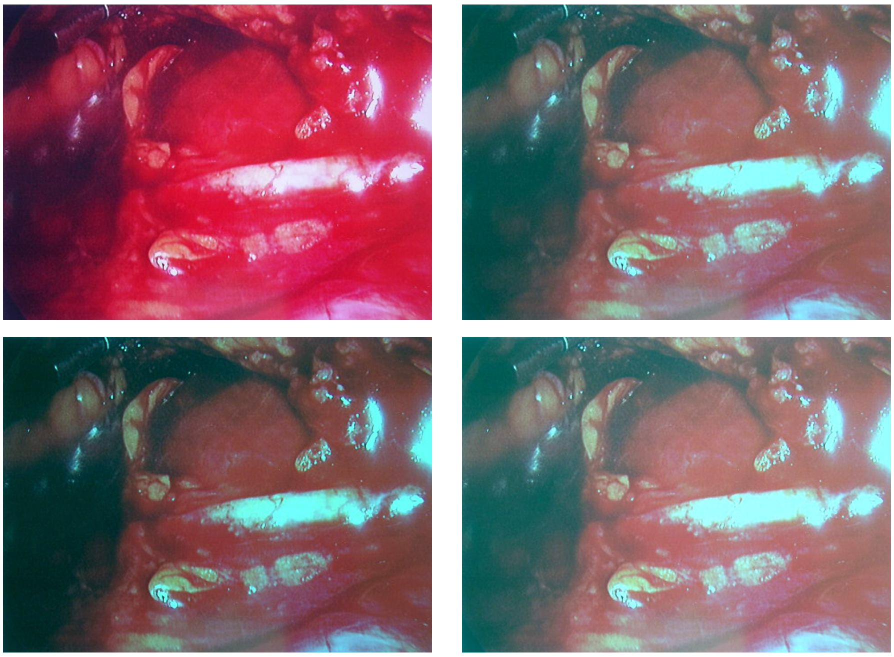


Figure 7.11: Color normalization example with different values for the center of the transformed color cluster $\boldsymbol{\mu}'$. The original image (`blood IV`) is displayed top left, the corresponding processed image with $\boldsymbol{\mu}'$ computed by equation (4.41), page 70, is shown top right, where $\boldsymbol{\mu}' = (112, 112, 112)^{\mathrm{T}}$ (cf. Table 7.5). For the bottom left and right images, $\boldsymbol{\mu}'$ was chosen $(80, 80, 80)^{\mathrm{T}}$ and $(130, 130, 130)^{\mathrm{T}}$.

- *Single channel median filter*: each color channel is filtered separately.
- *Vector median filter*: the pixels contained in the filter mask are sorted according to the norm of the color vector (cf. equation (4.42), page 72).

As the two filters yield different result images and the single channel median filter was employed for the implementation of the *temporal* color median filter (cf. Section 4.3, page 72), the question arises as to how large the difference between the two spatial versions of the filter is. In order to answer this question, 50 randomly generated images and 50 endoscopic images were processed with each spatial median filter. The difference between the resulting images was computed in terms of the mean value $\overline{Q}_{\mathrm{MAD}}$ (see equation (7.13), page 145). In this case the question regarding which of the two images was used as ground truth is irrelevant since only the difference was of interest. Table 7.7 shows the result[6]. As expected $\overline{Q}_{\mathrm{MAD}}$ was large for randomly generated images (17 to 24 gray-values) but very small for endoscopic images (0.13 to 0.80 gray-values). The small difference of the filtered endoscopic images and the drastically reduced computation

---

[6] The 50 endoscopic images were chosen from endoscopic image sequences that were recorded with a DV (digital video) recorder in DV PAL format, i. e., size 720×576 pixels.

Median Size (rows × columns)	3×1	5×1	7×1	3×3	5×5	7×7
$\overline{Q}_{\mathrm{MAD}}$ `random`	17	22	24	24	21	19
$\overline{Q}_{\mathrm{MAD}}$ `endoscopic`	0.13	0.25	0.35	0.46	0.70	0.80
Comp. Time Vector Median [msec]	900	1400	1800	2600	9700	29000
Comp. Time Channel Median [msec]	2.0	3.0	12	3.0	13	69

Table 7.7: Comparison of two kinds of spatial color median filters: *single channel median* and *vector median*. 50 random (`random`) and 50 endoscopic (`endoscopic`) images were processed and $\overline{Q}_{\mathrm{MAD}}$ was computed. The image size was 720×576 pixels. $\overline{Q}_{\mathrm{MAD}}$ is large (17 to 24 gray-values) for randomly generated images but very small (0.13 to 0.80 gray-values) for endoscopic images. In addition to $\overline{Q}_{\mathrm{MAD}}$, the computation time for each kind of filter is stated.

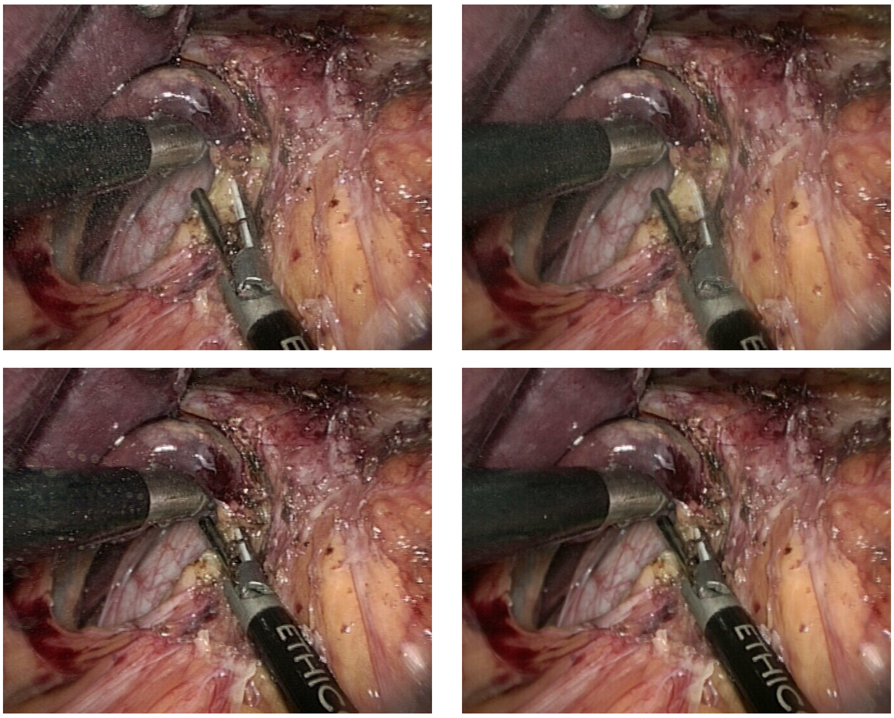


Figure 7.12: Two examples of temporal filtering with filter size 5. The filtered image is displayed to the right of the original image. Especially the disturbing small flying particles that are clearly visible in front of the black surgical instrument were removed.

time justify the use of the spatial *single channel* median filter for the implementation of the *temporal* color median filter. Examples of temporal filtering with filter size 5 are displayed in Figure 7.12.

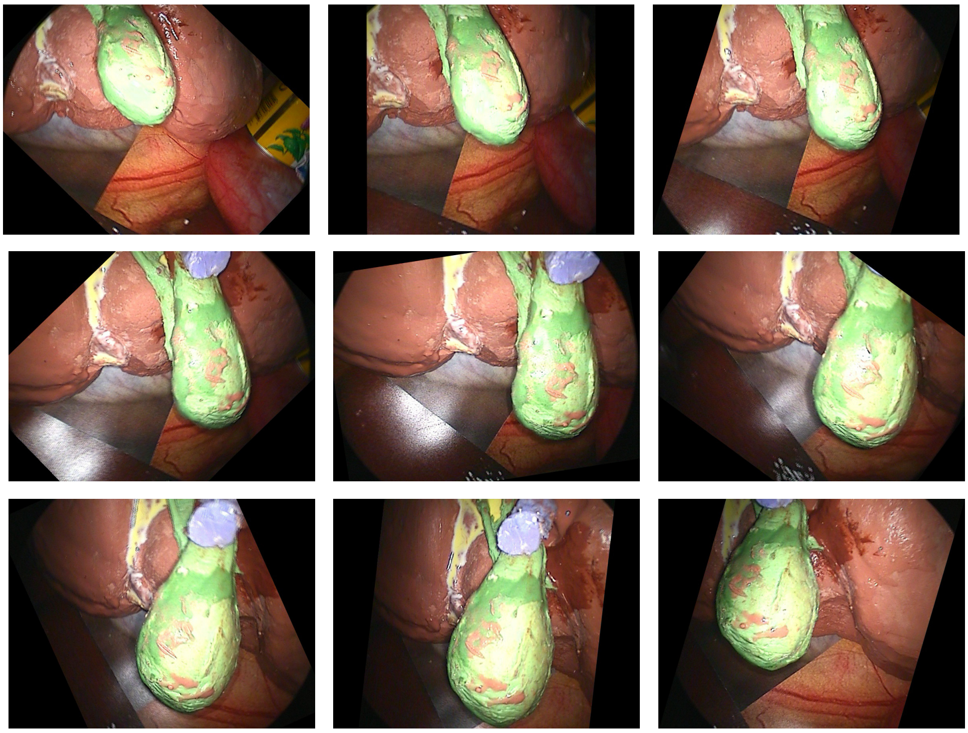


Figure 7.13: Horizon equalization: the endoscope was rotated approximately 240° with fixed camera head (top left to bottom right image). The experiment was performed in the laboratory using the liver/gall bladder model. Although the endoscope was rotated about 240° the horizon is kept constant, e. g., the gall bladder always “points downwards”.

7.2.4 Image Geometry Transformations

Examples of horizon equalization are shown in Figure 7.13. The benefit of digital zoom becomes clear when regarding Figure 7.14: a close view of the operation site *and* the original image can be displayed *simultaneously* on two monitors next to each other.

7.2.5 Evaluation

A subjective evaluation was performed for the following image enhancement methods: distortion correction, color normalization, and temporal filtering with filter sizes 3 and 5. The applied technique using the evaluation program *EvaMedIm* was described in Section 7.1.1, page 140. For each enhancement method 30 images were selected randomly. This yielded 90 image pairs. As temporal filtering was evaluated for filter sizes 3 and 5, another 30 image pairs were added, totaling 120 image pairs. In order to obtain the temporally filtered images the whole sequence was filtered and the corresponding images were taken from the filtered sequence. The number of

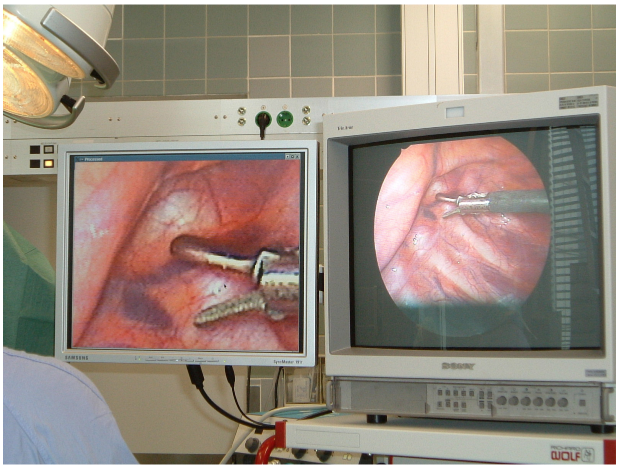


Figure 7.14: Digital zooming allows providing a very close view of the operation site (left monitor) while simultaneously displaying the original image (right monitor).

evaluating surgeons was 14, seven of them having more than five years experience and the other seven having no experience concerning minimally invasive operations. The evaluation criteria were:

- *Better/Worse*: Which of the two images do you prefer?
- *Sharpness*: Which of the two images looks sharper?
- *Distortion*: Which of the two images is less distorted?
- *Color Impression*: Which of the two images do you prefer regarding its color and the possibility of distinguishing different types of tissue?

The corresponding range for the evaluation value v was $\{-1, 0, 1\}$ for “Better/Worse” and $\{-2, -1, 0, 1, 2\}$ for the three other criteria. Recall that positive values signify that the processed image was preferred, negative values that the original image was preferred, and if no difference between the displayed pair was observable, v was set to zero. Since each of the 14 surgeons evaluated 30 image pairs per enhancement method, a total of 420 evaluations were obtained for each method. The mean values $\overline{v}_{\mathrm{All}}$ of these 420 evaluations are summarized in Table 7.8. Additionally, the mean values $\overline{v}_{\mathrm{Exp}}$ for the group of experienced physicians as well as $\overline{v}_{\mathrm{Unexp}}$ for the group of unexperienced physicians are stated separately. The corresponding p-values of the Wilcoxon signed rank test of the null hypothesis that the data originate from a distribution whose *median* is zero, i. e., “no observable difference”, can be found in Table 7.9. The p-values indicate the significance of the result: the smaller the p-value the more significant the result,

Method	Group	Better/Worse	Sharpness	Distortion	Color
Color Normalization	All	0.28	0.17	0.081	0.019
	Exp.	0.16	0.057	0.010	-0.24
	Unexp.	0.40	0.28	0.15	0.28
Distortion Correction	All	0.43	-0.064	0.52	-0.019
	Exp.	0.46	-0.057	0.58	-0.024
	Unexp.	0.41	-0.071	0.46	-0.014
Temporal Filtering (Size 3)	All	-0.069	-0.083	-0.043	-0.033
	Exp.	-0.038	-0.052	-0.033	-0.033
	Unexp.	-0.10	-0.11	-0.052	-0.033
Temporal Filtering (Size 5)	All	-0.21	-0.25	-0.055	-0.043
	Exp.	-0.24	-0.24	-0.10	-0.033
	Unexp.	-0.17	-0.26	-0.010	-0.052

Table 7.8: Subjective evaluation of image enhancement methods: 14 surgeons evaluated 30 image pairs for each method. The mean values of the resulting 420 evaluations are stated according to the evaluation criteria “Better/Worse”, “Sharpness”, “Distortion”, “Color” (impression). Additionally, the mean values for the group of experienced and the group of unexperienced physicians are stated. The range of the evaluation value was $\{-1, 0, 1\}$ for “Better/Worse” and $\{-2, -1, 0, 1, 2\}$ for the three other criteria. Positive values mean that the *processed* image was preferred and negative values that the *original* image was preferred. When no difference between the displayed pair was observable, the evaluation value for this pair was zero.

i. e., the lower the probability of obtaining the result if the null hypothesis were true. The improvement of the image quality by color normalization ($\overline{v}_{\text{All}} = 0.28$ for “Better/Worse” and $\overline{v}_{\text{All}} = 0.17$ for “Sharpness”) and distortion correction ($\overline{v}_{\text{All}} = 0.43$ for “Better/Worse” and $\overline{v}_{\text{All}} = 0.52$ for “Distortion”) is clearly visible. These results are highly significant with $p \ll 10^{-4}$ (see Table 7.9). For temporal filtering with filter size 3, $\overline{v}_{\text{All}}$ was not significantly different from zero ($p > 0.01$). Using filter size 5 the original images were preferred ($\overline{v}_{\text{All}} = -0.21$ for “Better/Worse” and $\overline{v}_{\text{All}} = -0.25$ for “Sharpness” with $p \ll 10^{-7}$).

Color normalization was the only enhancement method where the results of experienced and unexperienced surgeons differed noticeably. For “Better/Worse” and “Sharpness” the mean values were both positive, but quite different ($\overline{v}_{\text{Exp}} = 0.16$ in contrast to $\overline{v}_{\text{Unexp}} = 0.40$, and $\overline{v}_{\text{Exp}} = 0.057$ in contrast to $\overline{v}_{\text{Unexp}} = 0.28$), but the largest difference was obtained for “Color” ($\overline{v}_{\text{Exp}} = -0.24$ in contrast to $\overline{v}_{\text{Unexp}} = 0.28$). The interpretation of this result is that the color normalized images look uncommon to the experienced surgeons as they are accustomed to the “wrong” color of the endoscopic images. Interestingly, the experienced surgeons preferred the

Method	Group	Better/Worse	Sharpness	Distortion	Color
Color Normalization	All	$4.7\cdot 10^{-10}$	$6.7\cdot 10^{-5}$	$6.4\cdot 10^{-3}$	$6.6\cdot 10^{-1}$
	Exp.	$9.7\cdot 10^{-3}$	$2.6\cdot 10^{-1}$	$7.4\cdot 10^{-1}$	$5.6\cdot 10^{-4}$
	Unexp.	$1.1\cdot 10^{-9}$	$3.5\cdot 10^{-5}$	$3.7\cdot 10^{-3}$	$3.9\cdot 10^{-4}$
Distortion Correction	All	$1.2\cdot 10^{-26}$	$1.3\cdot 10^{-2}$	$5.8\cdot 10^{-28}$	$1.2\cdot 10^{-1}$
	Exp.	$4.9\cdot 10^{-16}$	$2.0\cdot 10^{-1}$	$4.5\cdot 10^{-19}$	$1.8\cdot 10^{-1}$
	Unexp.	$2.4\cdot 10^{-12}$	$7.1\cdot 10^{-3}$	$2.7\cdot 10^{-11}$	$6.3\cdot 10^{-1}$
Temporal Filtering (Size 3)	All	$3.0\cdot 10^{-2}$	$1.3\cdot 10^{-2}$	$2.4\cdot 10^{-2}$	$3.1\cdot 10^{-2}$
	Exp.	$3.7\cdot 10^{-1}$	$2.5\cdot 10^{-1}$	$1.3\cdot 10^{-1}$	$1.1\cdot 10^{-1}$
	Unexp.	$3.5\cdot 10^{-2}$	$2.1\cdot 10^{-2}$	$9.3\cdot 10^{-2}$	$1.4\cdot 10^{-1}$
Temporal Filtering (Size 5)	All	$1.4\cdot 10^{-8}$	$3.9\cdot 10^{-10}$	$1.1\cdot 10^{-2}$	$2.7\cdot 10^{-3}$
	Exp.	$1.9\cdot 10^{-6}$	$6.4\cdot 10^{-6}$	$3.9\cdot 10^{-4}$	$3.9\cdot 10^{-2}$
	Unexp.	$1.0\cdot 10^{-3}$	$1.2\cdot 10^{-5}$	$7.7\cdot 10^{-1}$	$3.4\cdot 10^{-2}$

Table 7.9: Significance in terms of p-values for the Wilcoxon ranksum test of the null hypothesis that the median of the evaluation values was zero, i. e., "no observable difference" (see Table 7.8).

processed images *in general* ($\overline{v}_{\mathrm{Exp}} = 0.16$ for "Better/Worse") although they did not like the new color ($\overline{v}_{\mathrm{Exp}} = -0.24$ for "Color").

Another question is why the original images were preferred for temporal filtering, especially in the case of filter size 5. The answer is that temporal filtering reduces the sharpness of the image ($\overline{v}_{\mathrm{All}} = -0.25$ for "Sharpness" and filter size 5). This is due to the fact that the prerequisite for "perfect" temporal filtering, a static scene and camera, is only partly fulfilled during minimally invasive operations. When comparing single images, the reduction of sharpness seems to be more relevant to the surgeons than the benefit of reduced temporal noise. This result was not satisfying, especially with regard to the good results obtained by temporal filtering that have been shown in Figure 7.12, page 155. Therefore, two temporally filtered image *sequences* were presented to the 14 surgeons. Both sequences were filtered with filter sizes 3 and 5. For this additional evaluation the surgeons could only decide which of the two times three sequences they prefer (original, temporal filter size 3, or temporal filter size 5). The result was: 4 votes for "I do not see a difference", zero votes for the original sequences, 8 votes for temporal filtering with filter size 3, and 16 votes for filter size 5. As only two sequences were evaluated by the 14 surgeons a statistical test did not make sense. Nevertheless, the result was obvious: the original sequences were *never* preferred but the majority of surgeons ($86\,\%$) voted for temporal filtered sequences, especially with filter size 5 ($57\,\%$). With regard to image sequences it can therefore be concluded that, in contrast to single images, the benefit of reduced temporal noise is larger than the loss

Filter	no	t3	t5	t7
$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	25.49	25.14	25.07	25.04
$\overline{Q}_{\mathrm{SNR}}$ [dB]	9.255	9.445	9.495	9.517
$\overline{Q}_{\mathrm{PSNR}}$ [dB]	18.56	18.76	18.82	18.84

Table 7.10: Objective evaluation of temporal filtering. The mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ are stated for the original sequence (no) and for the temporally filtered sequences with filter sizes 3 (t3), 5 (t5), and 7 (t7). The length of the sequence was 200 frames. AESOP was used to fix the endoscope in the laboratory. The ground truth sequence was captured first, then smoke was introduced into the artificial patient. PAL color images were captured (size 768×576 pixels). The ground truth image was computed by averaging 50 images of the ground truth sequence.

of quality due to reduced sharpness. In any case, the reduced sharpness of single images in an image sequence is only visible if several consecutive images are blurred and not only one or two. The results of this subjective evaluation were published in [Vog03a, Krü03a, Krü03b, Krü04], but without appropriate statistical proof of significance.

Temporal filtering was also evaluated objectively. The idea was to first obtain ground truth data. Temporal noise was then introduced and the filtered result compared to the ground truth data. The experiment was carried out in the laboratory using AESOP to fix the endoscope while capturing a sequence of the liver/gall bladder silicone model. A ground truth image was obtained by averaging 50 images of the ground truth sequence. Then, smoke was introduced into the artificial patient by lighting several matches inside the box, which has a small shutter through which this was done. This caused image distortions similar to those occurring when cutting tissue with high frequency diathermy during minimally invasive operations. After closing the shutter, an image sequence was captured. This sequence was temporally filtered with filter sizes 3, 5, and 7. The final sequence contained 200 frames. Q_{MAD}, Q_{SNR}, and Q_{PSNR} were computed for each image of the original (noisy) and the three filtered sequences. The mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ are shown in Table 7.10. An improvement of image quality was identified: the larger the temporal filter size the smaller gets $\overline{Q}_{\mathrm{MAD}}$ and the larger get $\overline{Q}_{\mathrm{SNR}}$ and $\overline{Q}_{\mathrm{PSNR}}$. Since mainly small flying particles were removed by temporal filtering, the *quantitative* improvement was not very large, e. g., $\overline{Q}_{\mathrm{MAD}}$ was only decreased by $1.8\,\%$ (from $\overline{Q}_{\mathrm{MAD}} = 25.49$ gray-values to $\overline{Q}_{\mathrm{MAD}} = 25.04$ gray-values).

Sequence	Frames	Preprocessing	Image Size	Comp. Time [sec]
`Gall-20020708`	141	di-int	400×400	743
`Gall-Tape4`	61	di-int	512×512	296
`Hyp-20010425`	121	di-int	400×400	94
`Stomach`	138	di-sub	256×256	1063

Table 7.11: Light field reconstruction using structure-from-motion: the number of frames, the applied preprocessing algorithm, the image size, and the computation time are stated for each sequence. The first three sequences were de-interlaced by interpolation ("di-int") whereas the `Stomach` sequence was de-interlaced by subsampling ("di-sub"). All sequences were captured during cholecystectomies, except for `Hyp-20010425` which was captured during a thoracoscopic operation.

7.3 Light Field Reconstruction and Visualization

This section describes the results of light field reconstruction using structure-from-motion techniques, the robot arm AESOP, and the optical tracking system smARTtrack1. Results of static light field reconstructions can be found in Sections 7.3.1 to 7.3.3, examples of dynamic light field reconstructions are shown in Section 7.3.4. For the presentation of the results only a few expressive light field reconstructions were selected as altogether 100 light fields were reconstructed using AESOP and smARTtrack1. All figures displaying the results of static light field reconstructions are located at the end of the last section on static light field reconstruction (Section 7.3.3), starting from page 177. This layout was chosen in order to describe the results without several interruptions of the text by pages containing only figures.

7.3.1 Static Light Fields Using Structure-From-Motion

Four sequences were employed for light field reconstruction by structure-from-motion techniques. Table 7.11 summarizes the properties of the reconstructions of these sequences. The endoscope diameters were 5 mm for `Hyp-20010425` and 10 mm for the other three sequences. The computation time was large although at most 141 frames were processed. Additionally, the computation time was not correlated to the number of frames, e. g., light field reconstruction of the `Hyp-20010425` sequence took only 94 sec whereas the reconstruction of the `Stomach` sequence required almost 18 min for only 17 more images with smaller image size. The images were cropped such that no black border was left which resulted in 512×512 pixels images as well as 400×400 pixels images. Otherwise confidence maps would have had to be used. The `Stomach` sequence was de-interlaced by subsampling which led to 256×256 pixels image size, the other three sequences were de-interlaced by interpolation (cf. Section 5.1.1, page 84). All four sequences were recorded during surgery without additionally acquiring images of a calibration pattern. Thus, no camera calibration and distortion correction could be performed. Nevertheless, light field reconstruction was possible but it turned out that using self-calibration, i. e., the determination of the intrinsic camera parameters for each image, led to a failure of the algorithm. In this case failure means that either only a small percentage of camera poses could be estimated or that the computed depth information was not usable. Therefore, the principal point was always set to the middle point of the images, e. g., $(C_{\mathrm{x}}, C_{\mathrm{y}})^{\mathrm{T}} = (256, 256)^{\mathrm{T}}$ for image size 512×512 pixels, and the focal lengths were fixed to a certain value, which was set heuristically. Light field reconstructions where the intrinsic camera parameters were calibrated and the original sequence was de-interlaced, undistorted, and cropped (cf. Section 5.1, page 84) will be presented later when comparing the structure-from-motion reconstructions to those obtained by using AESOP and smARTtrack1.

During the experiments with structure-from-motion light field reconstruction it turned out that the result depends strongly on the chosen parameters. Finding a good parameter set was difficult and it took some time to find one. Unfortunately, small changes of parameters may lead to a failure of the algorithm. The following list describes those parameters that had to be adapted depending on the sequence at hand:

- The number of points that should be tracked was set to 1000 for the two gall sequences (`Gall-20020708` and `Gall-Tape4`) and to 500 for the sequences `Hyp-20010425` and `Stomach`.

- The minimal distance for feature points was set to 10 pixels for `Gall-20020708` and `Gall-Tape4`, to 7 pixels for `Hyp-20010425`, and to 5 pixels for `Stomach`.

- The maximal allowed length of the initial sequence (cf. Section 3.3.2, page 50) was set to 20 for all sequences except `Hyp-20010425`, where a value of 120 was used. Actually, the finally selected initial sequence contained 100 frames. As the extension of the initial sequence takes much longer than the factorization of an equivalent sequence, the computation time was much lower for `Hyp-20010425` compared to the other three sequences. But using larger values for the maximally allowed length for the other sequences did not yield usable reconstruction results.

- The effective focal lengths F_{x} and F_{y} were set to 600 (`Gall-20020708`), to 800 (`Gall-Tape4` and `Hyp-20010425`), and to 900 (`Stomach`).

For all four sequences a window size of 11×11 pixels was used for feature detection as well as for feature tracking. It has to be noted that better parameters for feature tracking, which allow tracking the 2-D feature points longer, did not necessarily lead to an improvement of the reconstruction. It even occurred that "better" tracking parameters resulted in a failure of the reconstruction algorithm.

The results of light field reconstruction of the `Gall-20020708` and `Hyp-20010425` sequences are illustrated in Figures 7.18 and 7.19, pages 177 and 178. These two figures as well as many others use a "standard" presentation of light field reconstruction results:

- The top row shows three images of the original sequence.
- The middle row displays the computed camera path (left image) and two views of the computed 3-D points (middle and right image). Pyramids are used to represent camera poses, the tip being the camera center and the base being parallel to the image plane.
- The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the reconstructed light field (middle and right image) using the unstructured lumigraph rendering approach (cf. Section 3.2.2, page 46).

The computed 3-D points and the 3-D triangular mesh show the shape of the surface of the operating field. Although no calibration pattern was used, the quality of the rendered images is good. Nevertheless, artefacts are visible especially for the `Hyp-20010425` sequence (see Figure 7.19, page 178).

7.3.2 Static Light Fields Using AESOP

Table 7.12 states the errors of endoscope pose determination using the robot arm AESOP (cf. Section 7.1.2, page 142). The two image sequences of a calibration pattern contained 55 and 100 images, respectively. For both sequences $\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$ was about the same and very large ($56\,\%$ and $57\,\%$, respectively). The relative rotation error $\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$ of the sequence `ALF 40` was smaller than that of sequence `ALF 14`, but still large. The two types of relative rotation errors, computed from axis/angle and Cardan angle representation, led to approximately the same value, i. e., $\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}} \approx \bar{\epsilon}_{\boldsymbol{R},\mathrm{rel,C}}$. The threshold for the circularity criterion C was set to 16 for both sequences. The number of contour points of valid contours had to be in the interval $[30, \infty)$ (`ALF 14`) and $[50, \infty]$ (`ALF 40`). The contour area had to be in the interval $[20, 1000]$ square pixels (`ALF 14`) and $[50, 1000]$ square pixels (`ALF 40`). The thresholds used for binarization were 110 gray-values (`ALF 14`) and 84 gray-values (`ALF 40`), where the symmetric 7×7 calibration pattern

Sequence	**Translation Error**		**Rotation Error**		**Rotation Error (Cardan)**			
	$\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\overline{\epsilon}_{\boldsymbol{t}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	$\overline{\epsilon}_{\boldsymbol{R}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel,C}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\alpha}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\beta}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\gamma}$
`ALF 14`	57 %	3.9 mm	36 %	3.0 °	36 %	1.8 °	1.5 °	1.2 °
`ALF 40`	56 %	2.9 mm	26 %	1.9 °	24 %	0.70 °	0.46 °	1.5 °

Table 7.12: Endoscope pose errors of the sequences `ALF 14` (55 images) and `ALF 40` (100 images). The pose errors for each sequence were computed from 1000 randomly selected pose pairs with minimal frame distance $\Delta_f = 5$.

Sequence	**Frames**	θ_{BPE}	**Prep. [sec]**	**Track. [sec]**	**Depth [sec]**	$\sum$ **[sec]**
`ALF 12`	100	15	5.0	17	18	40
`ALF 52`	128	15	8.0	22	12	42
`ALF 53`	100	15	4.0	18	12	34
`ALF 65`	141	10	8.0	20	29	57
`ALF 67`	144	15	12	22	18	52

Table 7.13: Light field reconstruction using AESOP: the number of frames, the threshold θ_{BPE} for the back-projection error, and the computation times for preprocessing, point tracking, and computation of depth are stated. For depth computation LMedS and non-linear optimization was used. The total computation time is also stated ($\sum$ [sec]).

was used. The parameters determined by hand-eye calibration were $l_{\mathrm{e}} = 200\,\mathrm{mm}$, $\alpha_{\mathrm{plug}} = 356.6°$ (`ALF 14`), $\alpha_{\mathrm{plug}} = 118.3°$ (`ALF 40`), $\alpha_{\mathrm{c2o}} = 273.1°$ (`ALF 14`), and $\alpha_{\mathrm{c2o}} = 271.7°$ (`ALF 40`). For all experiments with AESOP a 5 mm endoscope was utilized, where the angle of the side view optics α_{opt} was 30°.

Table 7.13 summarizes the properties of five selected light field reconstructions using AESOP. The parameters determined by hand-eye calibration were $l_{\mathrm{e}} = 245\,\mathrm{mm}$, $\alpha_{\mathrm{plug}} = 72.7°$, $\alpha_{\mathrm{c2o}} = 298.3°$ (`ALF 12`), $l_{\mathrm{e}} = 245\,\mathrm{mm}$, $\alpha_{\mathrm{plug}} = 154.8°$, $\alpha_{\mathrm{c2o}} = 324.6°$ (`ALF 52` and `ALF 53`), and $l_{\mathrm{e}} = 235\,\mathrm{mm}$, $\alpha_{\mathrm{plug}} = 6.4°$, $\alpha_{\mathrm{c2o}} = 319.3°$ (`ALF 65` and `ALF 67`). The tracking parameters were the same for all sequences: the number of points that should be tracked was set to 500, the minimal distance of feature points was set to 5 pixels, and a window size of 11×11 pixels was used for feature detection as well as for feature tracking. Apart from the threshold for the back-projection error θ_{BPE} (see Table 7.13), the parameters of depth computation were the same for all sequences: a 2-D feature point had to be tracked longer than 10 frames in order to be taken into account, the probability of outliers p_{out} was set to 0.3 for the application of LMedS, depth values outside the interval $[0, 10000]$ mm were discarded, and a 32×32 pixel grid was used for the interpolation of additional depth points. The maximal number of iterations for non-linear optimization of the extrinsic camera parameters was set to 25.

The reconstruction of light fields using AESOP was very fast in comparison to using structure-from-motion techniques (see Table 7.11, page 161, and Table 7.13). The computation times were not equal for image sequences with the same number of frames as the depth computation is based on 2-D point tracking and the computation time of point tracking depends on the image sequence at hand. The number of points that can be tracked, the number of points that are lost from frame to frame, and the number of *trails* influence the computation time, where a trail contains all 2-D point correspondences that were obtained by tracking a certain feature point. A two-plane light field was computed for the sequences `ALF 53` and `ALF 65`. This took 66 sec and 92 sec, respectively. Examples of light field reconstruction using AESOP in the laboratory and in the operating room are displayed in Figures 7.20 and 7.21, pages 179 and 180 respectively. The computed depth information of the operating room light field was very noisy and almost unusable. In contrast to this the computed depth information of the laboratory sequence clearly shows the scene's surface.

For a comparison of structure-from-motion light field reconstruction to light field reconstruction using AESOP, see Figure 7.22, page 181. A sequence of a city map paper ball (`ALF 67`) was chosen which provides texture information that could be used for point tracking. The point tracking parameters were the same as for light field reconstruction using AESOP. The maximal allowed length of the initial sequence (cf. Section 3.3.2, page 50) was set to 20, the effective focal lengths determined by camera calibration were used, $F_\mathrm{x} = 510.0$ and $F_\mathrm{y} = 508.2$, and the principal point was set to the middle point of the image, i. e., $(C_\mathrm{x}, C_\mathrm{y})^\mathrm{T} = (256, 256)^\mathrm{T}$. Using the real principal point resulted in a worse result. For this sequence the rectangular camera path of AESOP was reconstructed by the structure-from-motion approach but the 3-D points are very flat. Compared to this, the camera path computed using AESOP's kinematics is noisier but the computed 3-D points look more like the shape of a ball.

Depth computation can be improved by an LMedS technique to eliminate endoscope pose outliers and by non-linear optimization (cf. Section 5.6, page 112). The quality of two light fields (`ALF 65` and `ALF 67`) was evaluated without LMedS, with LMedS, and with LMedS and non-linear optimization of the extrinsic camera parameters (see Table 7.14). Additionally, the quality of a light field reconstructed using structure-from-motion is stated. Without LMedS no usable depth information was computed ($\overline{Q}_\mathrm{MAD} = 127$ and $\overline{Q}_\mathrm{MAD} = 129$, respectively). The application of the LMedS technique yielded usable results and with additional non-linear optimization a

Sequence	$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	$\overline{Q}_{\mathrm{SNR}}$ [dB]	$\overline{Q}_{\mathrm{PSNR}}$ [dB]
`ALF 65`	127	0.00	4.60
`ALF 65` LMedS	25.3	11.8	16.4
`ALF 65` LMedS Opt.	16.5	15.1	19.7
`ALF 65` SFM	-	-	-
`ALF 67`	129	0.00	4.49
`ALF 67` LMedS	33.0	10.2	14.7
`ALF 67` LMedS Opt.	28.9	11.2	15.7
`ALF 67` SFM	16.6	15.4	19.9

Table 7.14: Evaluation of light field reconstruction for different types of depth computation: without LMedS and optimization, with LMedS, and with LMedS and non-linear optimization of the extrinsic camera parameters. The mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ are stated. Additionally, the light field quality using structure-from-motion (SFM) was compared to the light field quality using AESOP. For the sequence `ALF 65` the structure-from-motion approach failed since images were only captured when AESOP moved the endoscope from left to right. Thus, no continuous image stream was available and the algorithm stopped after reconstructing the first of five movements from left to right.

major improvement was achieved ($\overline{Q}_{\mathrm{MAD}} = 16.5$ in contrast to $\overline{Q}_{\mathrm{MAD}} = 25.3$ and $\overline{Q}_{\mathrm{MAD}} = 28.9$ in contrast to $\overline{Q}_{\mathrm{MAD}} = 33.0$). Light field reconstruction using structure-from-motion failed for the sequence `ALF 65` because this sequence was not captured continously but only when the endoscope was moved from left to right. This causes a problem when using structure-from-motion since usually all tracked 2-D points are lost at the end of a movement from left to right. At this point the algorithm cannot proceed to the next image and therefore stops. The sequence `ALF 67` was captured continously and the structure-from-motion approach succeeded. In this case the quality of the structure-from-motion light field was better ($\overline{Q}_{\mathrm{MAD}} = 16.6$ compared to $\overline{Q}_{\mathrm{MAD}} = 28.9$) although the computed depth information is flat and does not look like a paper ball (cf. Figure 7.22, page 181). With respect to the quality of light fields it seems that more accurate camera poses with less accurate depth information are superior to less accurate camera poses with more accurate depth information.

7.3.3 Static Light Fields Using smARTtrack1

Three targets for endoscope tracking with smARTtrack1 were designed (cf. Section 5.5.1, page 103): the *"Epee"*, the *"DD 2z"*, and the *"DD"* target. The error of hand-eye calibration using these targets is compared in Table 7.15. In order to judge these and the following errors, the

Target	**Translation Error**		**Rotation Error**		**Rotation Error (Cardan)**			
	$\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\overline{\epsilon}_{\boldsymbol{t}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	$\overline{\epsilon}_{\boldsymbol{R}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel,C}}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\alpha}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\beta}$	$\overline{\epsilon}_{\boldsymbol{R},\mathrm{C},\gamma}$
Epee	5.0 %	2.1 mm	1.1 %	0.43 °	1.1 %	0.28 °	0.24 °	0.11 °
DD 2z	4.2 %	1.3 mm	0.70 %	0.27 °	0.74 %	0.17 °	0.12 °	0.12 °
DD	3.4 %	1.2 mm	0.96 %	0.31 °	0.97 %	0.17 °	0.15 °	0.14 °

Table 7.15: Comparison of hand-eye calibration errors with smARTtrack1 using the *Epee*, the *DD 2z*, and the *DD* target. The errors are determined by comparing the pose data obtained by transforming the hand poses by the computed hand-eye transformation to the eye data obtained by camera calibration (ground truth). The pose errors were computed from 100 randomly selected pose pairs with minimal frame distance $\Delta_f = 1$.

accuracy of smARTtrack1 has to be taken into account (cf. Section 5.5, page 102): 0.19 mm position error in x- and y-direction, 0.36 mm position error in z-direction, and $0.14°$ rotation error. Note that the errors specified by the manufacturer are usually lower than the ones obtained in real applications [Wag02]. Thus, the accuracy of the application at hand should be evaluated in order to determine the actual errors. The relative translation errors of hand-eye calibration were similar (3 % to 5 %) as well as the two types of relative rotation errors (about 1 %). The effect of data selection (cf. Section 5.5.2, page 110) and non-linear optimization for hand-eye calibration is depicted in Table 7.16. Only data selection based on vector quantization was examined since it turned out that this approach is superior to the exhaustive search method [Sch04a]. The objective function given in [Hor95] was utilized for non-linear optimization with the Levenberg-Marquardt algorithm [Den83], where the sum of the mean squared errors according to the central hand-eye equation is minimized (cf. equation (5.36), page 107). Based on the seven examined sequences no optimal method could be determined since the lowest errors were either obtained without data selection (`ART 52`), with data selection but without non-linear optimization (`ART 92`), or with data selection *and* non-linear optimization (`ART 115`). As the computation of the hand-eye transformation is very fast when using about 20 images and has to be done only once at the beginning of an operation, all three solutions were computed and the one with the lowest error was chosen for further processing.

After having evaluated the accuracy of hand-eye calibration, the endoscope pose accuracy using smARTtrack1 was examined. Table 7.17 compares the pose errors for the three targets. The smallest endoscope pose errors were obtained using the *DD* target ($\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}} = 3.8\,\%$ and $\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}} = 2.7\,\%$).

In order to evaluate the accuracy of the 3-D reconstruction, namely the computed depth infor-

Sequence	Temporal		VQ		VQ Opt.	
	$\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$
`ART 21`	5.0 %	1.1 %	4.7 %	1.1 %	4.4 %	0.86 %
`ART 32`	3.9 %	1.0 %	4.0 %	1.0 %	4.0 %	0.96 %
`ART 38`	4.2 %	0.70 %	3.9 %	0.66 %	3.9 %	0.62 %
`ART 52`	3.4 %	0.96 %	3.5 %	1.1 %	4.0 %	0.96 %
`ART 74`	5.0 %	1.2 %	4.9 %	1.1 %	5.5 %	0.87 %
`ART 92`	4.7 %	1.2 %	4.4 %	1.2 %	6.7 %	1.5 %
`ART 115`	4.1 %	4.2 %	2.7 %	2.6 %	2.5 %	2.2 %

Table 7.16: Hand-eye calibration of smARTtrack1. The relative translation and rotation errors ($\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}, \bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$) without data selection (Temporal), with data selection based on vector quantization (VQ), and with additional non-linear optimization of the obtained result (VQ Opt.) are stated. The pose errors were computed from 100 randomly selected pose pairs with minimal frame distance $\Delta_f = 1$.

Target	Translation Error		Rotation Error		Rotation Error (Cardan)			
	$\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{t}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	$\bar{\epsilon}_{\boldsymbol{R}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel,C}}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\alpha}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\beta}$	$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\gamma}$
Epee	7.3 %	1.7 mm	3.1 %	0.54 °	3.2 %	0.20 °	0.21 °	0.40 °
DD 2z	7.6 %	1.9 mm	4.1 %	0.93 °	4.4 %	0.32 °	0.37 °	0.67 °
DD	3.8 %	1.5 mm	2.7 %	0.63 °	2.5 %	0.25 °	0.41 °	0.28 °

Table 7.17: Endoscope pose accuracy using smARTtrack1. For each available target the pose errors were computed from 1000 randomly selected pose pairs with minimal frame distance $\Delta_f = 5$.

mation, a sequence of a textured sphere in front of a black background was recorded. Thus, only points on the surface of the sphere were tracked. As the radius of the sphere was known, the error of the computed 3-D points was determined as follows. A sphere was fitted to the 3-D points by estimating the center and radius of the sphere. The equation for a 3-D sphere with radius r and center $(c_x, c_y, c_z)^{\mathrm{T}}$ is $(x - c_x)^2 + (y - c_y)^2 + (z - c_z)^2 = r^2$. Expanding this equation yields

$$\frac{1}{r^2 - c_x^2 - c_y^2 - c_z^2} \Big((x^2 + y^2 + z^2) - 2c_x x - 2c_y y - 2c_z z \Big) = 1 \,. \tag{7.17}$$

Defining $r' := 1/(r^2 - c_x^2 - c_y^2 - c_z^2)$ this can be written as the following scalar product:

$$\big(x^2 + y^2 + z^2, x, y, z \big) \left(r', -2r'c_x, -2r'c_y, -2r'c_z \right)^{\mathrm{T}} = 1 \,. \tag{7.18}$$

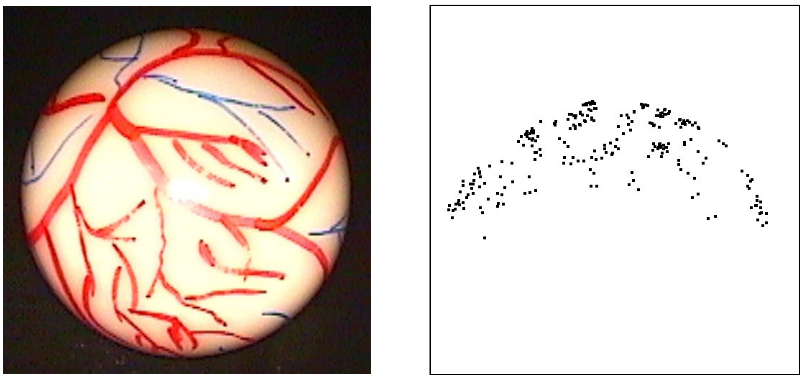


Figure 7.15: A sequence of a textured sphere in front of a black background is used for evaluation of the 3-D reconstruction accuracy using smARttrack1 (left image). The image on the right displays an example of the computed 3-D points that are used for estimating the radius and center of the sphere.

Target	Radius error	Shape error
Epee	7.4 % = 1.7 mm	1.3 mm
DD 2z	1.4 % = 0.31 mm	0.22 mm
DD	0.92 % = 0.21 mm	0.36 mm

Table 7.18: Accuracy of 3-D reconstruction using smARTtrack1. A sphere with radius 22.5 mm was reconstructed. After estimating the center and radius of the sphere based on the computed 3-D points (cf. Figure 7.15), the absolute and relative radius error and the mean distance of the 3-D points from the estimated sphere surface (shape error) were determined.

Then, for each computed 3-D point one equation such as (7.18) is obtained resulting in a linear system of equations of the form $\boldsymbol{A}\boldsymbol{x} = \mathbf{1}_n$, where $\boldsymbol{A} \in \mathrm{I\!R}^{n \times 4}$ for n 3-D points, each element of $\mathbf{1}_n \in \mathrm{I\!R}^n$ is 1, and $\boldsymbol{x} := (r', -2r'c_x, -2r'c_y, -2r'c_z)^{\mathrm{T}}$. The solution vector $\boldsymbol{x}$ is obtained by multiplying $\mathbf{1}_n$ by the pseudo-inverse $\boldsymbol{A}^+$ of $\boldsymbol{A}$ from the left (see Appendix B). Let x_i be the i-th element of the solution vector $\boldsymbol{x}$. Then,

$$c_x = -\frac{x_2}{2r'} = -\frac{x_2}{2x_1}, \quad c_y = -\frac{x_3}{2x_1}, \quad c_z = -\frac{x_4}{2x_1}, \text{ and} \tag{7.19}$$

$$r = \sqrt{r' + c_x^2 + c_y^2 + c_z^2} = \sqrt{x_1 + c_x^2 + c_y^2 + c_z^2}\,. \tag{7.20}$$

The absolute and relative radius error was then computed, where the radius of the used sphere was 22.5 mm. Figure 7.15 displays the utilized sphere and an example of computed 3-D points. Table 7.18 states the errors for the three targets. The lowest error was obtained with the *DD* target: 0.92 % (0.21 mm). Additionally, a shape error ϵ_{shape} was computed in terms of the mean

Sequence	**Frames**	θ_{BPE}	**Prep. [sec]**	**Track. [sec]**	**Depth [sec]**	$\sum$ **[sec]**
`ART 22`	155	2	12	34	25	71
`ART 75`	165	2	15	42	32	89
`ART 93`	342	3	28	93	48	169
`ART 118`	511	2	59	156	91	306

Table 7.19: Light field reconstruction using smARTtrack1: the number of frames, the threshold θ_{BPE} for the back-projection error, and the computation times for preprocessing, point tracking, and computation of depth are stated. For depth computation LMedS and non-linear optimization was used. The total computation time is also stated ($\sum$ [sec]). The sequence `ART 118` was recorded by using a *conventional video camera* in order to test the transferability of the developed methods for endoscope pose determination and light field reconstruction (see also Figure 7.16, page 171).

distance of the 3-D points to the sphere's surface:

$$\epsilon_{\mathrm{shape}} = \sum_{i=1}^{n} \left| \| \widehat{\boldsymbol{w}}_i - (c_x, c_y, c_z)^{\mathrm{T}} \| - r \right| , \tag{7.21}$$

where $\widehat{\boldsymbol{w}}_i$ is the i-th of the n computed 3-D points. In this case, the *DD 2z* target yielded the best results ($\epsilon_{\mathrm{shape}} = 0.22\,\mathrm{mm}$). The error obtained with the *DD* target was slightly larger ($\epsilon_{\mathrm{shape}} = 0.36\,\mathrm{mm}$).

Two examples of light field reconstruction using smARTtrack1 in the laboratory are shown in Figures 7.23 and 7.27, pages 182 and 184. An example of light field reconstruction in the operating room is illustrated in Figure 7.28, page 185. The shape of the reconstructed scene is clearly visible when regarding the computed 3-D points, e. g., see Figure 7.23 (gall bladder) and Figure 7.28 (liver).

Table 7.19 summarizes the properties of four selected light field reconstructions using smARTtrack1. A $10\,\mathrm{mm}$ endoscope was utilized for all endoscopic sequences. The tracking parameters were the same for all sequences: the number of points that should be tracked was set to 500, the minimal distance of feature points was set to 10 pixels, and a window size of 15×15 pixels was used for feature detection as well as for feature tracking. Apart from the threshold for the back-projection error θ_{BPE}, the parameters of depth computation were the same for all sequences: a 2-D feature point had to be tracked longer than 10 frames in order to be taken into account, the probability of outliers p_{out} was set to 0.3 for the application of LMedS, depth values outside the interval $[0, 10000]$ mm were discarded, and a 32×32 pixel grid was used for the interpolation of additional depth points. The maximal number of iterations for non-linear optimization of the

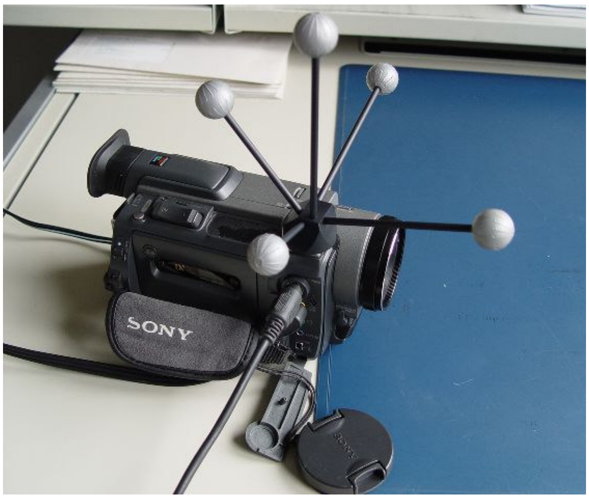


Figure 7.16: A target was attached to a conventional video camera (Sony 3-CCD) in order to test the transferability of the developed methods for light field reconstruction. The result of the reconstructed light field is shown in Figure 7.32, page 187.

extrinsic camera parameters was set to 25.

For a sequence with about 155 frames the reconstruction of a light field took less than one and a half minutes (see Table 7.19, sequence `ART 22`). As the computation time depends mainly on the number of images, longer sequences accordingly required more computation time, e. g., 5 min for the sequence `ART 118` with 511 frames.

In order to test the transferability of the developed methods for light field reconstruction using *endoscopes*, a target was attached to a *conventional video camera* (see Figure 7.16). Using this setup 20 images for hand-eye calibration were acquired (cf. Table 7.16, page 168, sequence `ART 115`) and an image sequence of a toy crawler was captured (sequence `ART 118`). Light field reconstruction could be performed without changing any parameters, i. e., exactly the same parameters as for light field reconstruction with an endoscope in the laboratory were employed ($\theta_{\mathrm{BPE}} = 2$). The result of the light field reconstruction using a conventional video camera is shown in Figure 7.32, page 187. The shape of the toy crawler is clearly visible. Another four light fields were reconstructed with this setup. In each case the light field was reconstructed without changing any parameters.

A comparison of the light field quality for different types of depth computation and for the reconstruction using structure-from-motion can be found in Table 7.20. Without LMedS no usable depth information was computed for the sequence `ALF 75` ($\overline{Q}_{\mathrm{MAD}} = 128$). The non-linear optimization results in a considerable improvement of the light field quality for all three sequences. The quality of the operating room light field (`ART 93`) is worse compared to the

Sequence	$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	$\overline{Q}_{\mathrm{SNR}}$ [dB]	$\overline{Q}_{\mathrm{PSNR}}$ [dB]
`ART 22`	9.27	19.9	24.7
`ART 22` LMedS	8.87	20.4	25.1
`ART 22` LMedS Opt.	5.43	24.9	29.7
`ART 22` SFM	5.26	25.3	30.0
`ART 75`	128	0.00	4.83
`ART 75` LMedS	10.6	18.8	23.7
`ART 75` LMedS Opt.	8.09	21.3	26.1
`ART 93`	13.0	17.2	21.6
`ART 93` LMedS	13.0	17.2	21.7
`ART 93` LMedS Opt.	10.2	19.2	23.7
`ART 93` SFM	10.4	19.0	23.5

Table 7.20: Evaluation of light field reconstruction for different types of depth computation: without LMedS and optimization, with LMedS, and with LMedS and non-linear optimization of the extrinsic camera parameters. The mean values $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ are stated. The results are stated for laboratory (`ART 22` and `ART 75`) and one operating room sequence (`ART 93`). Additionally, the light field quality using structure-from-motion (SFM) was compared for two sequences (`ART 22` and `ART 93`) to the light field quality using smARTtrack1.

laboratory light fields (`ART 22` and `ART 75`). The larger threshold for the back-projection error also indicates this result (cf. Table 7.19). This may be due to the small movements that occur, e. g., by heart beat and breathing of the patient.

The quality of the structure-from-motion light fields is comparable to the light fields reconstructed using smARTtrack1. The reconstruction results of the two structure-from-motion light fields in terms of the computed camera path and the computed 3-D points are displayed in Figure 7.24, page 183, and Figure 7.29, page 186. These results look very similar to those obtained by using smARTtrack1. The point tracking parameters were the same as for light field reconstruction using smARTtrack1. The maximally allowed length of the initial sequence was set to 20 (cf. Section 3.3.2, page 50), the effective focal lengths determined by camera calibration were used ($F_{\mathrm{x}} = 547.2$ and $F_{\mathrm{y}} = 545.3$ for sequence `ART 22` and $F_{\mathrm{x}} = 550.5$ and $F_{\mathrm{y}} = 550.0$ for sequence `ART 93`), and the principal point was set to the middle point of the image, i. e., $(C_{\mathrm{x}}, C_{\mathrm{y}})^{\mathrm{T}} = (256, 256)^{\mathrm{T}}$. Using the real principal point resulted in a worse result. The computation times of light field reconstruction using structure-from-motion are stated in Table 7.21. For these two sequences, light field reconstruction using structure-from-motion took much longer than the reconstruction using smARTtrack1, e. g., it took almost ten minutes for the sequence

Sequence	Prep. [sec]	Track. [sec]	Rec. [sec]	Depth [sec]	$\sum$ [sec]
`ART 22` SFM	12	34	506	8.0	560
`ART 93` SFM	28	93	906	18	1045

Table 7.21: Computation times for structure-from-motion light field reconstruction of the sequences `ART 22` and `ART 93` with 155 and 342 frames, respectively.

Sequence	Normal [sec]	LMedS [sec]	LMedS and Optimization [sec]
`ART 22`	8.0	11	25
`ART 75`	9.0	13	32
`ART 93`	16	20	48
`ART 118`	16	25	91

Table 7.22: Computation times for different approaches for depth computation. Compared to the "normal" approach the "LMedS" approach is only a few seconds slower. Naturally, the additionally required computation time for non-linear optimization depends on the number of frames of the sequence (see Table 7.19, page 170).

`ART 22` compared to slightly more than one minute when using smARTtrack1 (cf. Table 7.19, page 170).

The effects of using the LMedS technique and non-linear optimization of the extrinsic camera parameters for depth computation of the two sequences `ART 22` and `ART 93` are illustrated in Figure 7.25 and Figure 7.26, page 183, and Figure 7.30 and Figure 7.31, page 186. Figures 7.25 and 7.30 visualize the improvement by LMedS and non-linear optimization for a selected 3-D triangular mesh. For both sequences the benefit of the application of LMedS is barely visible whereas the benefit of non-linear optimization is clearly visible: the resulting 3-D mesh contains more and more accurate 3-D points, e. g., if the shapes of the gall bladder of the left and the right image in Figure 7.25 are compared. The non-linearly optimized extrinsic camera parameters often look smoother (see Figures 7.26 and 7.31). The computation times for the three approaches for depth computation are summarized in Table 7.22. The application of LMedS took only a few seconds longer, even for a large image sequence like `ART 118`. Non-linear optimization depends mainly on the number of frames, e. g., for 511 frames (`ART 118`) it took 75 sec longer compared to the "normal" method, whereas it took only 17 sec longer for 155 frames (`ART 22`).

Finally, the influence of different hand-eye calibration results on the quality of the computed

Sequence	$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	$\overline{Q}_{\mathrm{SNR}}$ [dB]	$\overline{Q}_{\mathrm{PSNR}}$ [dB]
`ART 93` temp.	10.2	19.2	23.7
`ART 93` VQ	10.1	19.2	23.7
`ART 93` VQ Opt.	10.3	19.2	23.6
`ART 118` temp.	12.8	16.1	21.1
`ART 118` VQ	11.0	17.1	22.1
`ART 118` VQ Opt.	10.8	17.2	22.2

Table 7.23: Influence of data selection and non-linear optimization for hand-eye calibration on the light field quality. $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ were computed for each light field. Three light fields were computed for the two sequences `ART 93` and `ART 118`: without data selection for hand-eye calibration (temp.), with data selection based on vector quantization (VQ), and with additional non-linear optimization of the hand-eye calibration result (VQ Opt.).

light field was examined. Two sequences were selected such that the first one (`ART 93`) resulted in larger errors for hand-eye calibration with vector quantization and non-linear optimization and the second one (`ART 118`) resulted in smaller errors (cf. Table 7.16, page 168, where sequence `ART 92` is the hand-eye calibration sequence for sequence `ART 93` and sequence `ART 115` the one for sequence `ART 118`). Table 7.23 states the results. For each of the two sequences three light fields were reconstructed and their quality measured in terms of $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$. The difference of the light field quality for the sequence `ART 93` was small (at most 0.2 gray-values). For the sequence `ART 118` the difference was larger: $\overline{Q}_{\mathrm{MAD}}$ was reduced by 2.0 gray-values by data selection based on vector quantization and non-linear optimization. The reason for this difference is presumably due to the relative rotation error $\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$ (see Table 7.15, page 167): the difference of the hand-eye calibration methods "temp." and "VQ Opt." was 2.0 percentage points for sequence `ART 115` but only 0.3 percentage points for the sequence `ART 92`. The translation error difference between the two hand-eye calibration methods was comparable.

In addition to the objective evaluation of light field quality as shown above, a subjective evaluation of the light fields was performed. Four light fields were selected, two reconstructed from laboratory sequences (`ART 22` and `ART 75`) and two from sequences recorded in the operating room (`ART 93` and `ART 99`). Five image pairs were evaluated for each light field. Each image pair consisted of an original image and of an image rendered from the light field using the camera parameters corresponding to the original image, but omitting the three cameras/images best suited for rendering. Thus, similar to objective evaluation, the rendered image had to be interpolated from neighboring images and the result depended on the accuracy of the computed camera parameters and depth information.

Altogether, 20 image pairs were presented to 10 surgeons. The original image was always identified as the better one since the rendered image was obtained by interpolation. The surgeons had to quantify the difference: no difference (grade 1), small difference which is only visible after regarding the images for a few seconds (grade 2), large difference but the quality of the worse image is sufficient to work with (grade 3), large difference and the quality of the worse image is *not* sufficient to work with (grade 4). In order to simplify the evaluation for the tutor, and as it was expected that the original image will always be identified as the better one, a "blind" instead of a "double-blind" setup was used. This means the tutor knew which one of the displayed images was the original one but the evaluating surgeon did not. In addition to the image pairs, the original sequence as well as a sequence rendered from the light fields were presented to the surgeons, i. e., four times two sequences. The surgeons had to judge the quality of the light fields *in general* according to these movies. This means *one* grade was set by each surgeon. Again, four grades were available: very good (grade 1), good (grade 2), bad (grade 3), and very bad (grade 4).

The mean value of the 200 image pair evaluations was 2.9. This means the difference is visible but the quality of the rendered images is sufficient to work with. Subdividing the evaluation into laboratory light fields and operating room light fields, the mean value was 2.5 for the laboratory light fields and 3.3 for the operating room light fields. This reinforces the already stated difference of light field quality in the laboratory and in the operating room (cf. Table 7.20, page 172). The histograms of the grades are displayed in Figure 7.17. They illustrate this fact additionally. Interestingly, the mean value of the movie evaluation was 1.9, i. e., in this case the quality of light fields in general was "good".

As mentioned at the beginning of Section 7.3, page 161, the following figures (pages 177 to 187) display results of static light field reconstruction. The following section on dynamic light field reconstruction (Section 7.3.4) starts on page 188.

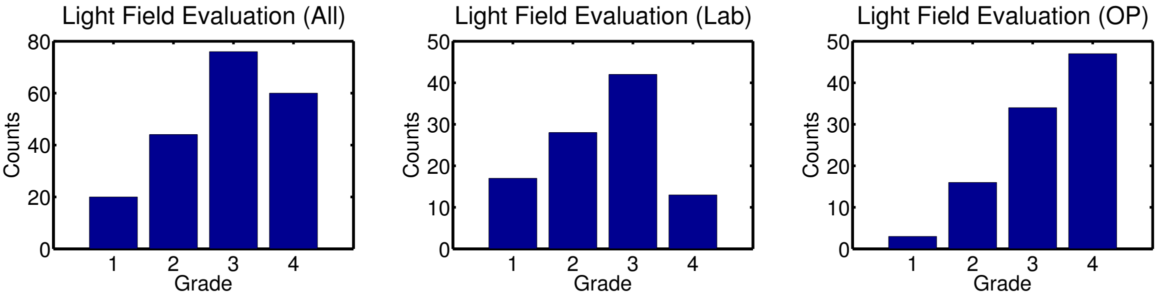


Figure 7.17: Subjective light field evaluation by 10 surgeons based on 20 image pairs. Image pairs consisting of an original and a rendered image were evaluated. The difference was quantified by four grades: no difference (grade 1), small difference which is only visible after regarding the images for a few seconds (grade 2), large difference but the quality of the worse image is sufficient to work with (grade 3), and large difference and the quality of the worse image is *not* sufficient to work with (grade 4). The counts of four grades are shown for all light fields together (left histogram), for the laboratory light fields only (middle histogram), and for the operating room light fields only (right histogram).

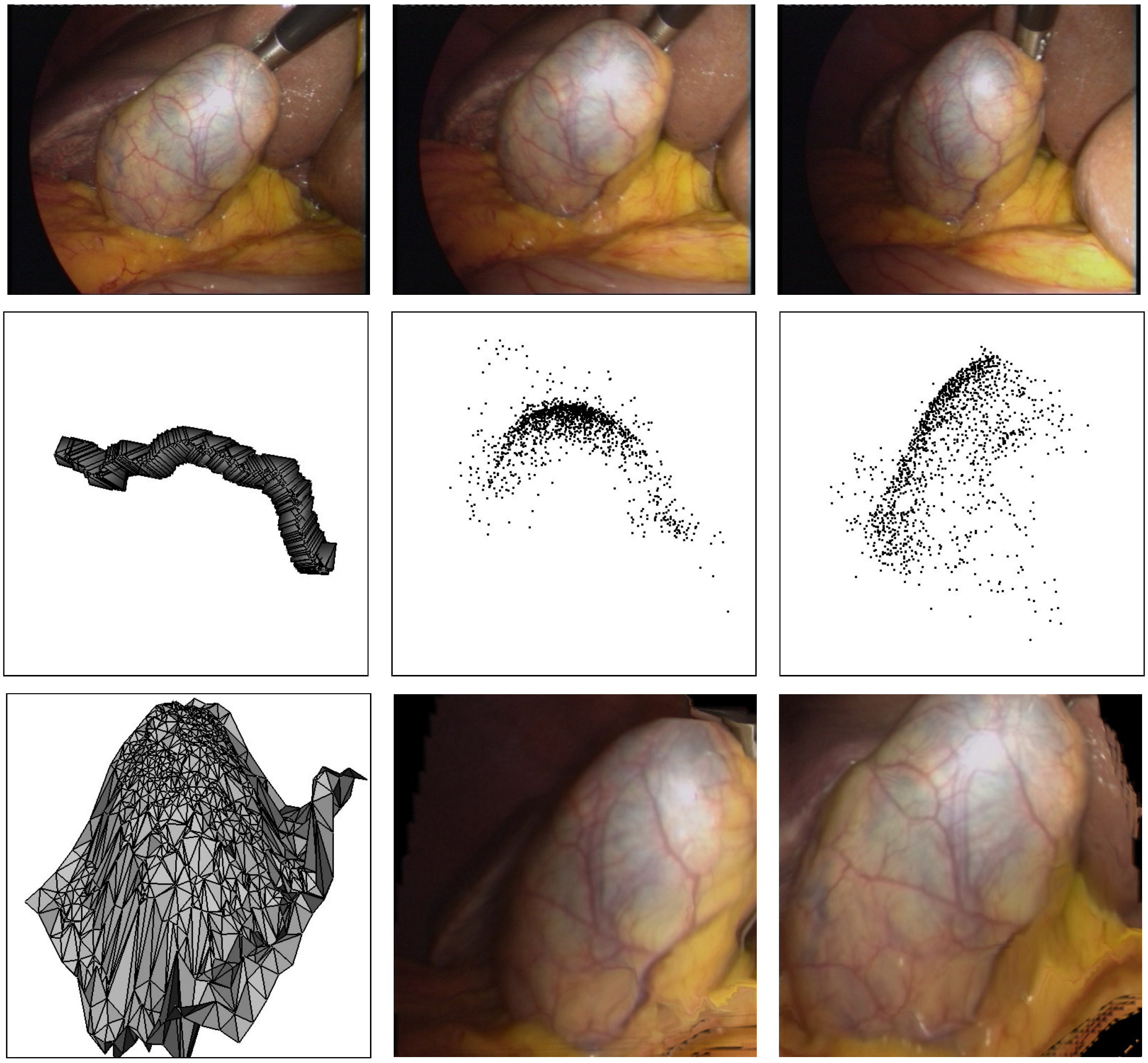


Figure 7.18: Structure-from-motion light field reconstruction of the sequence `Gall-20020708` (cholecystectomy, gall bladder). The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

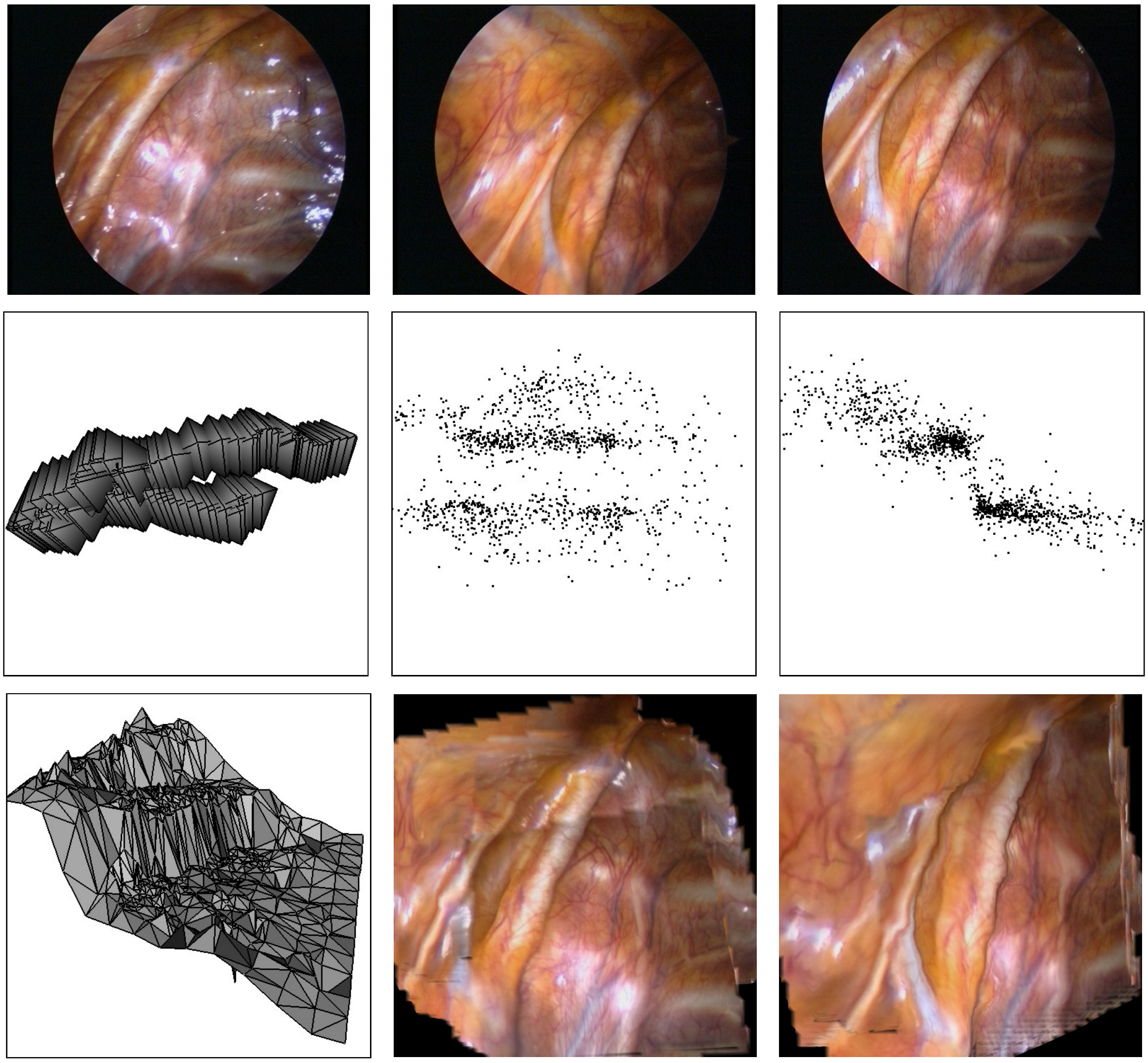


Figure 7.19: Structure-from-motion light field reconstruction of the sequence `Hyp-20010425` (thoracoscopic operation, thoracic cavity). The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

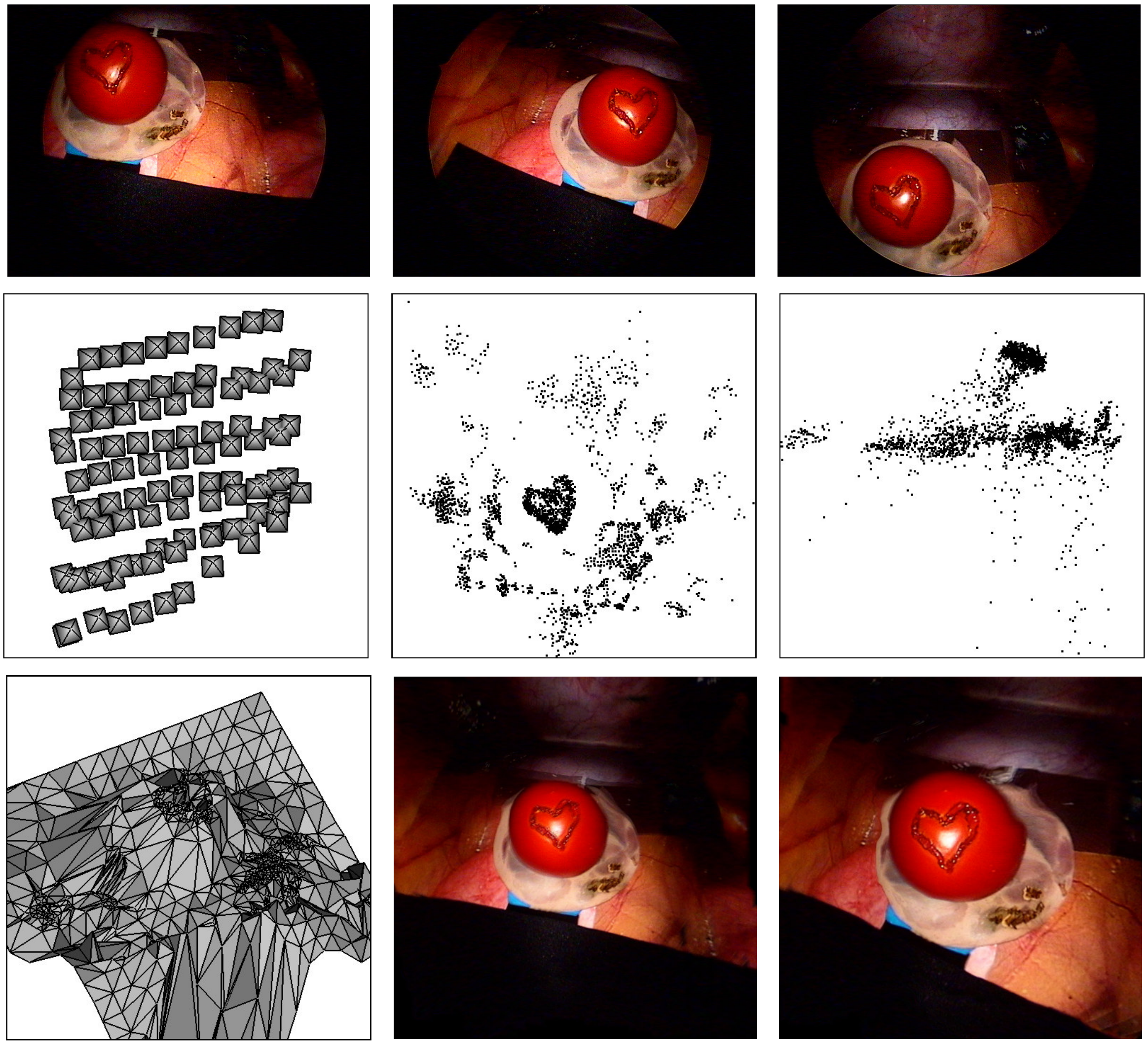


Figure 7.20: Light field reconstruction of the sequence `ALF 12` (laboratory, tomato) using AESOP. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

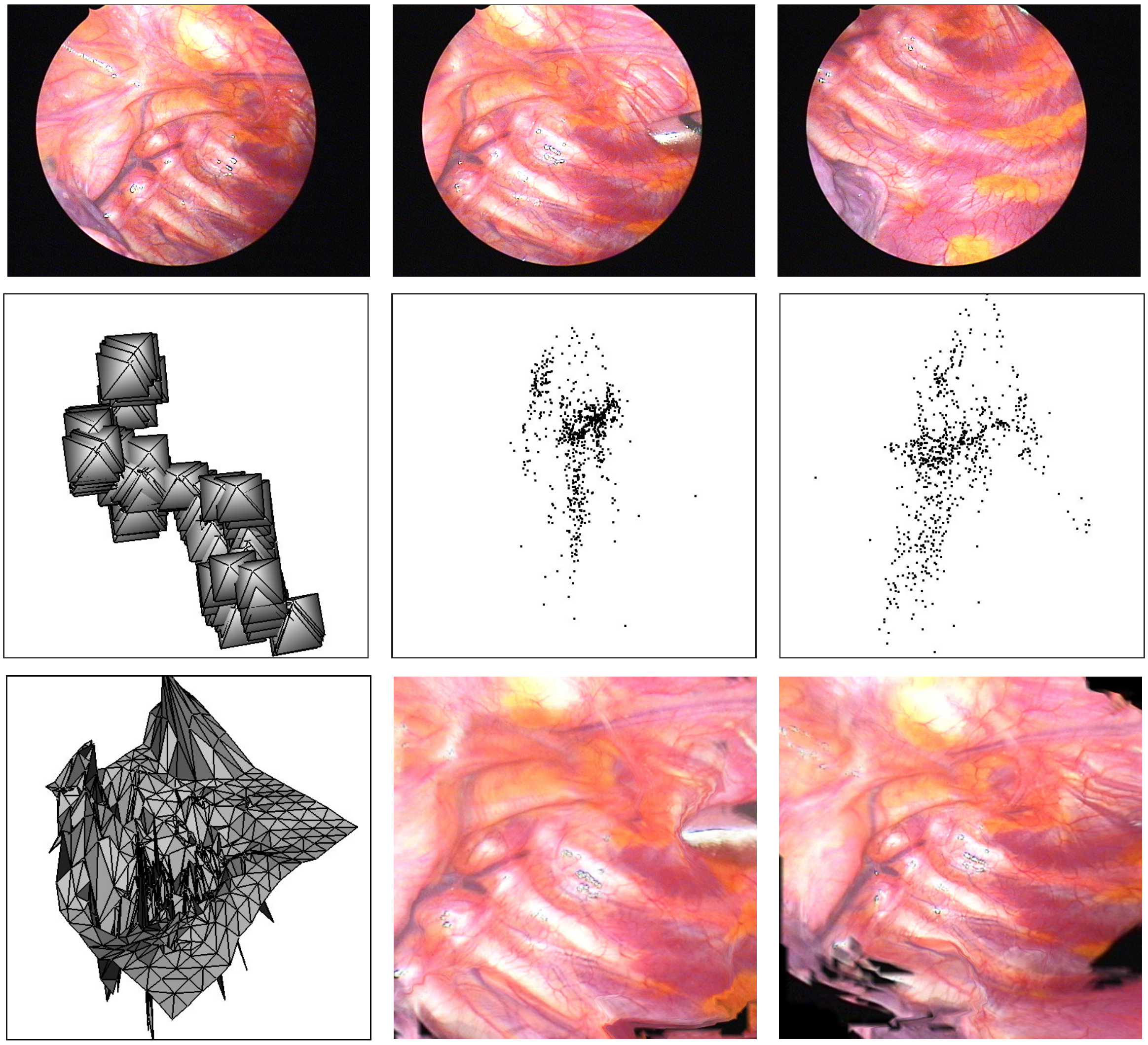


Figure 7.21: Light field reconstruction of the sequence `ALF 53` (thoracoscopic operation, thoracic cavity) using AESOP. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

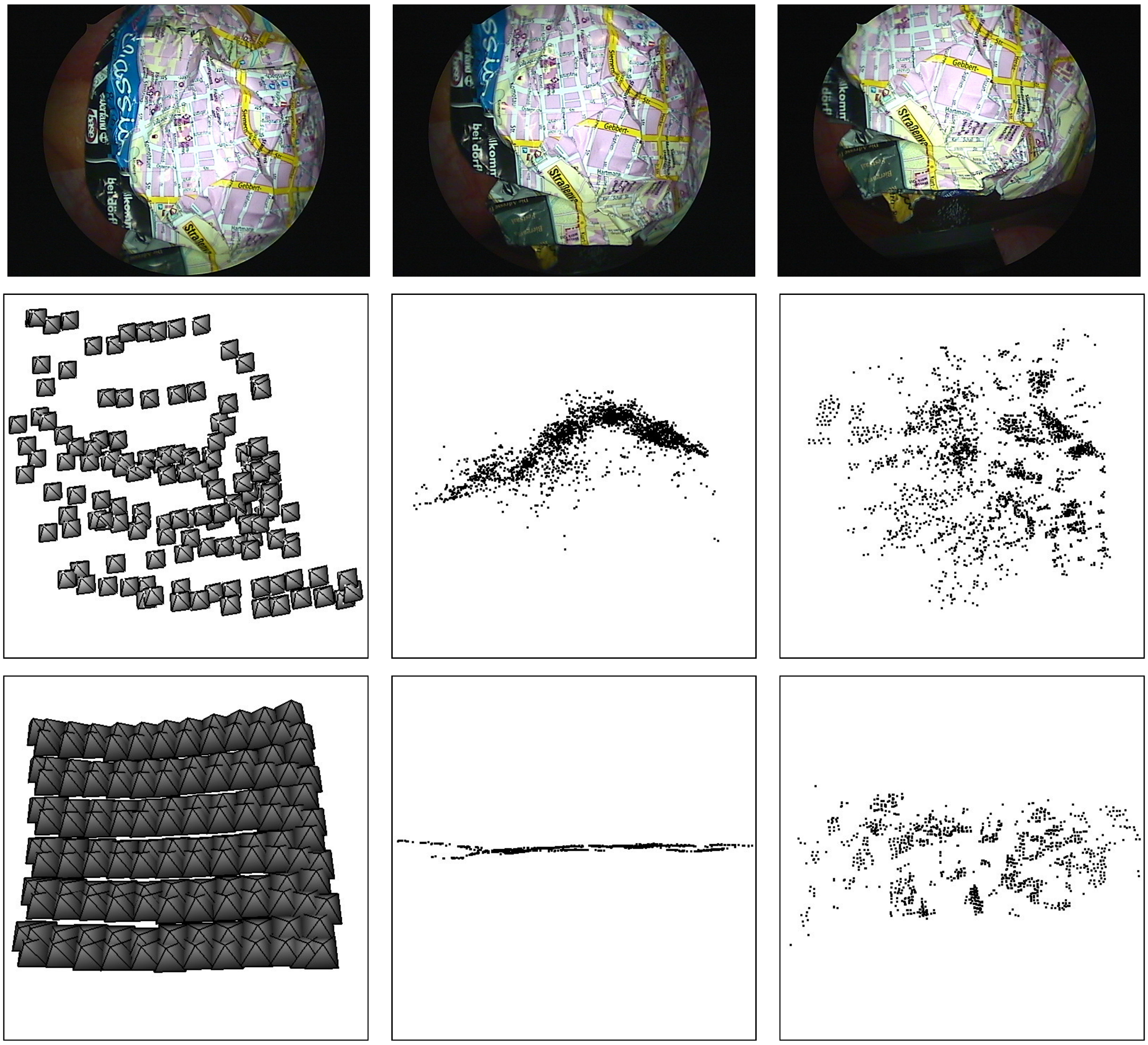


Figure 7.22: Comparison of light field reconstruction using structure-from-motion and AESOP. The top row shows three images of the sequence `ALF 67` (laboratory, city map paper ball). The middle row displays the camera path (left image), where each camera pose is visualized by a pyramid, and two views of the computed 3-D points using AESOP (middle and right image). The bottom row displays the same results using structure-from-motion techniques. The different appearance of the camera paths is due to the size of the pyramids. The base size was $3 \times 3\,\mathrm{mm}^2$ with a distance of 3 mm to the tip for the reconstruction using AESOP, whereas the size of the pyramids for the structure-from-motion approach was chosen heuristically since the camera positions are scaled arbitrarily.

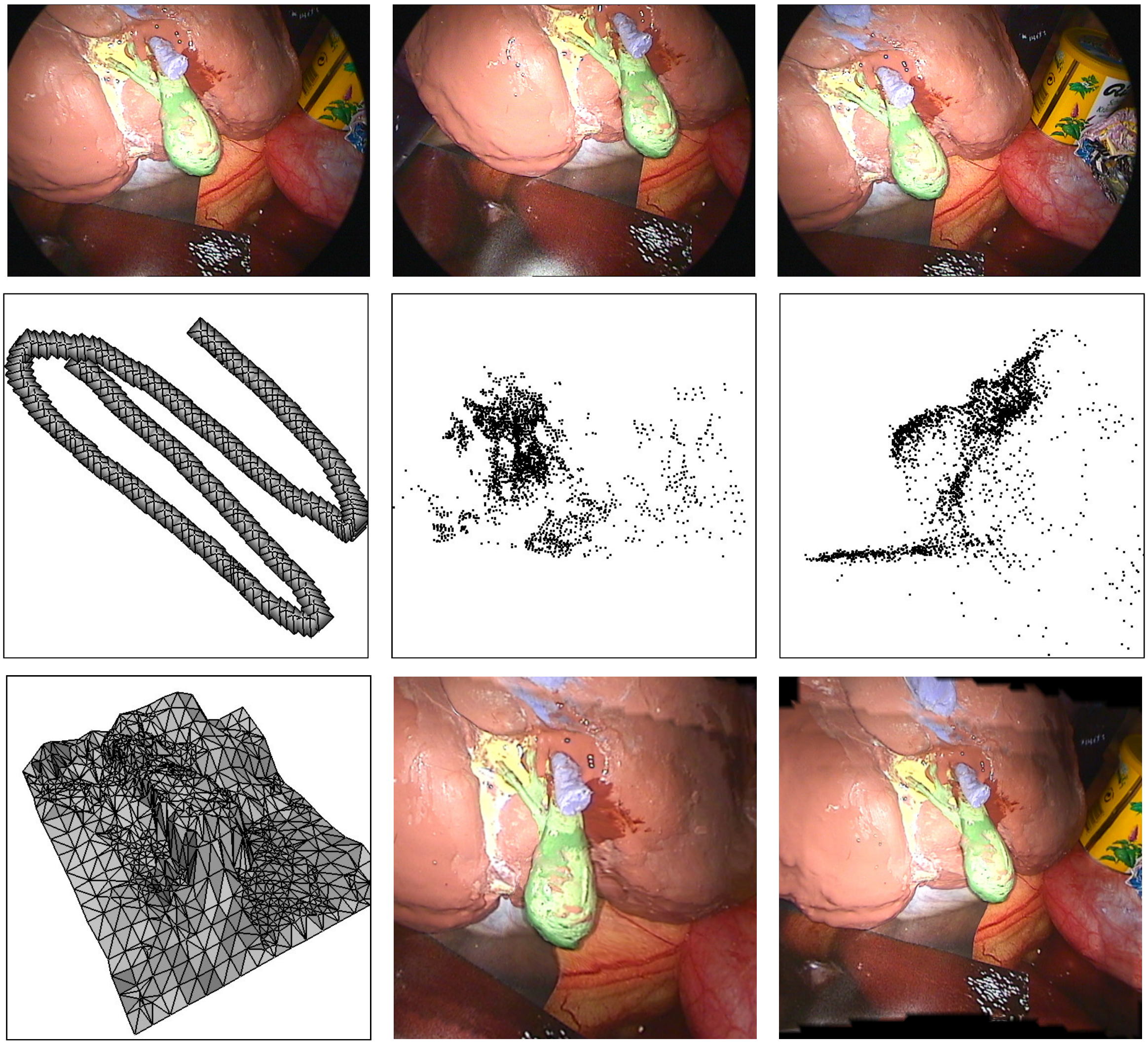


Figure 7.23: Light field reconstruction of the sequence `ART 22` (laboratory, liver/gall bladder model) using smARTtrack1. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

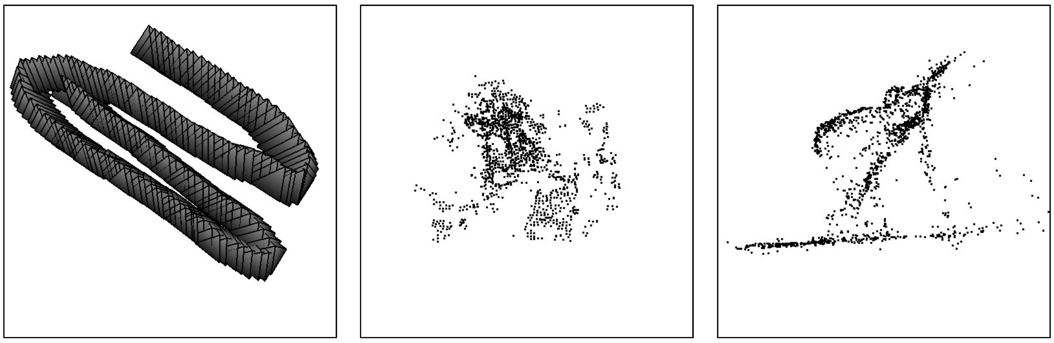


Figure 7.24: 3-D reconstruction result of the sequence `ART 22` using structure-from-motion. The reconstructed camera path (left image) and two views of the reconstructed 3-D points (middle and right image) are shown. The corresponding reconstruction results using smARTtrack1 are displayed in Figure 7.23, page 182 (middle row).

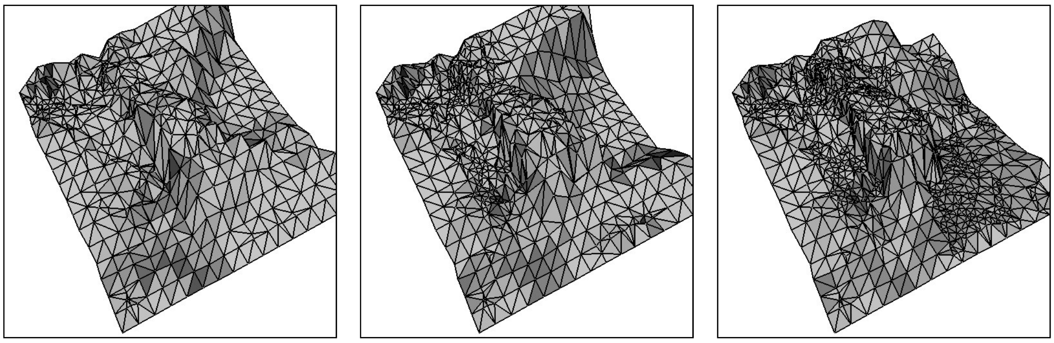


Figure 7.25: On the basis of one selected 3-D triangular mesh, the effects of applying the LMedS technique and non-linear optimization of the extrinsic camera parameters for depth computation are illustrated (sequence `ART 22`). The 3-D mesh contains less points without these techniques (left image). The benefit of LMedS is barely visible (middle image), whereas the benefit of non-linear optimization is clearly visible as the resulting 3-D mesh contains more and more accurate 3-D points (right image).

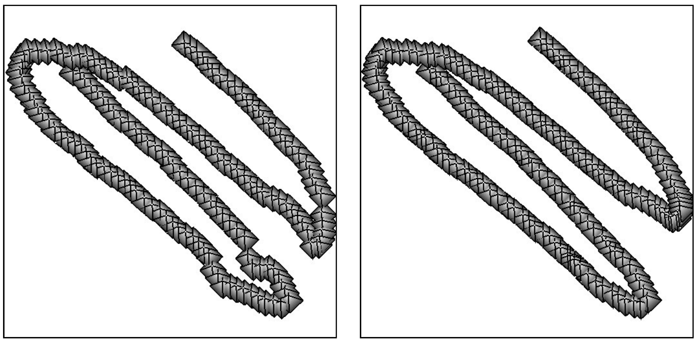


Figure 7.26: The image on the left shows the extrinsic camera parameters as computed by applying the estimated hand-eye transformation to the hand data provided by smARTtrack1 (sequence `ART 22`). Pyramids represent the extrinsic camera parameters, the tip being the camera center, the base being parallel to the image plane. The image on the right shows the non-linearly optimized extrinsic camera parameters.

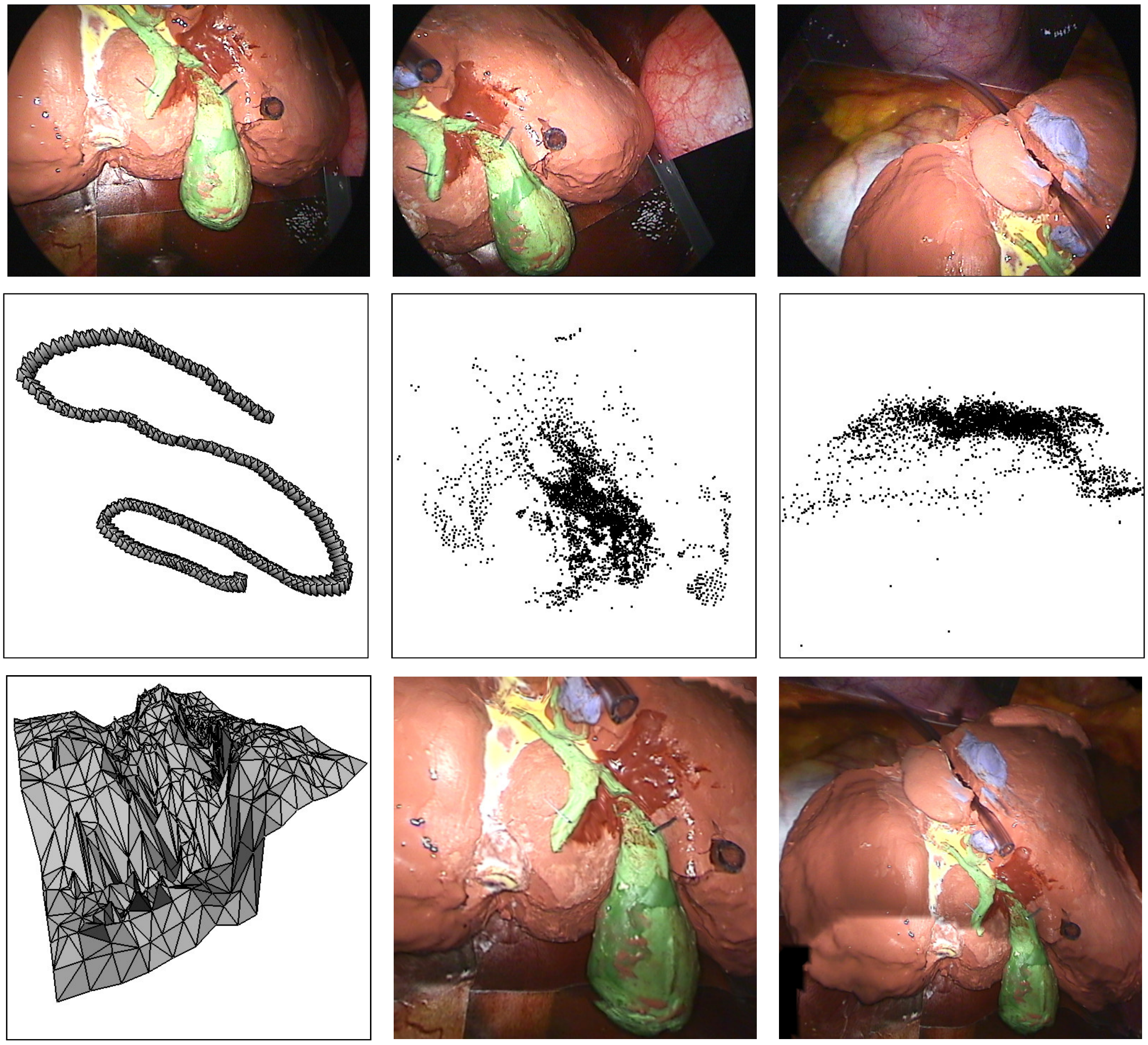


Figure 7.27: Light field reconstruction of the sequence `ART 75` (laboratory, liver/gall bladder model with tubes simulating vessels) using smARTtrack1. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

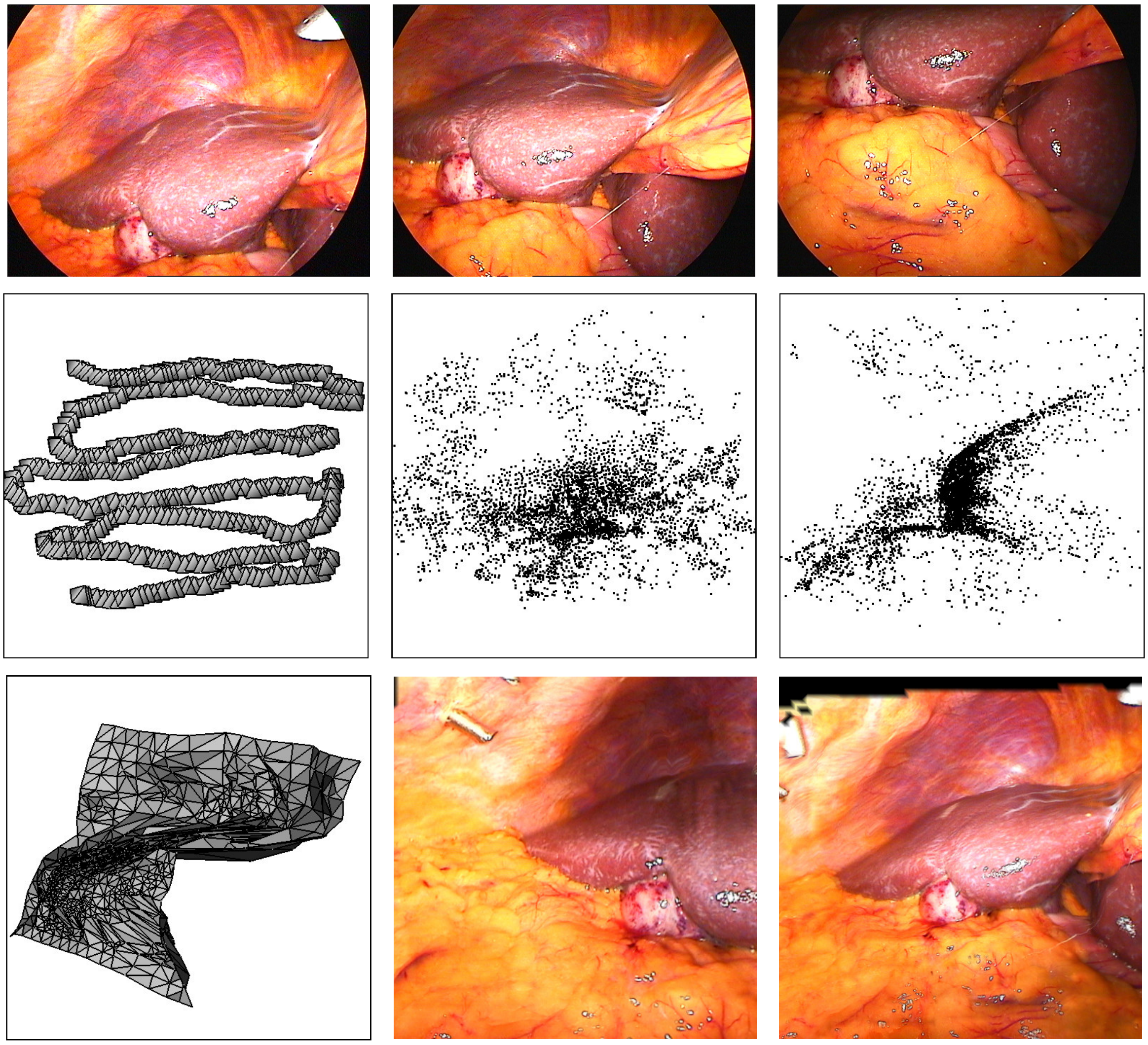


Figure 7.28: Light field reconstruction of the sequence `ART 93` (cholecystectomy) using smARTtrack1. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

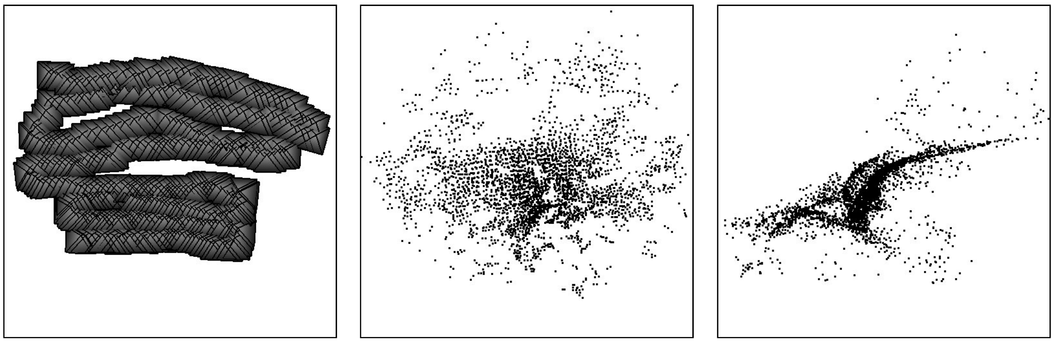


Figure 7.29: 3-D reconstruction result of the sequence `ART 93` using structure-from-motion. The reconstructed camera path (left image) and two views of the reconstructed 3-D points (middle and right image) are shown. The corresponding reconstruction results using smARTtrack1 are displayed in Figure 7.28, page 185 (middle row).

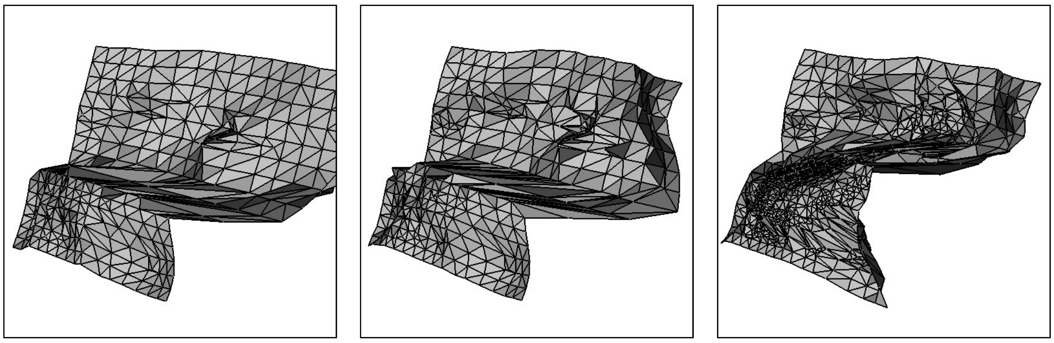


Figure 7.30: Illustration of the effects of LMedS and non-linear optimization of the extrinsic camera parameters for sequence `ART 93`: one selected 3-D triangular mesh without LMedS and non-linear optimization (left image), with LMedS (middle image), and with LMedS and non-linear optimization of the extrinsic camera parameters (right image). A clear benefit is visible for non-linear optimization together with LMedS, whereas the application of LMedS only barely influences the visible result.

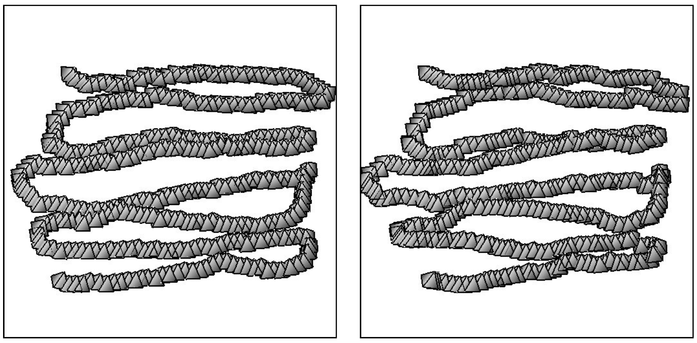


Figure 7.31: The image on the left shows the extrinsic camera parameters as computed by applying the estimated hand-eye transformation to the hand data provided by smARTtrack1 (sequence `ART 93`). Pyramids represent the extrinsic camera parameters, the tip being the camera center, the base being parallel to the image plane. The image on the right shows the non-linearly optimized extrinsic camera parameters.

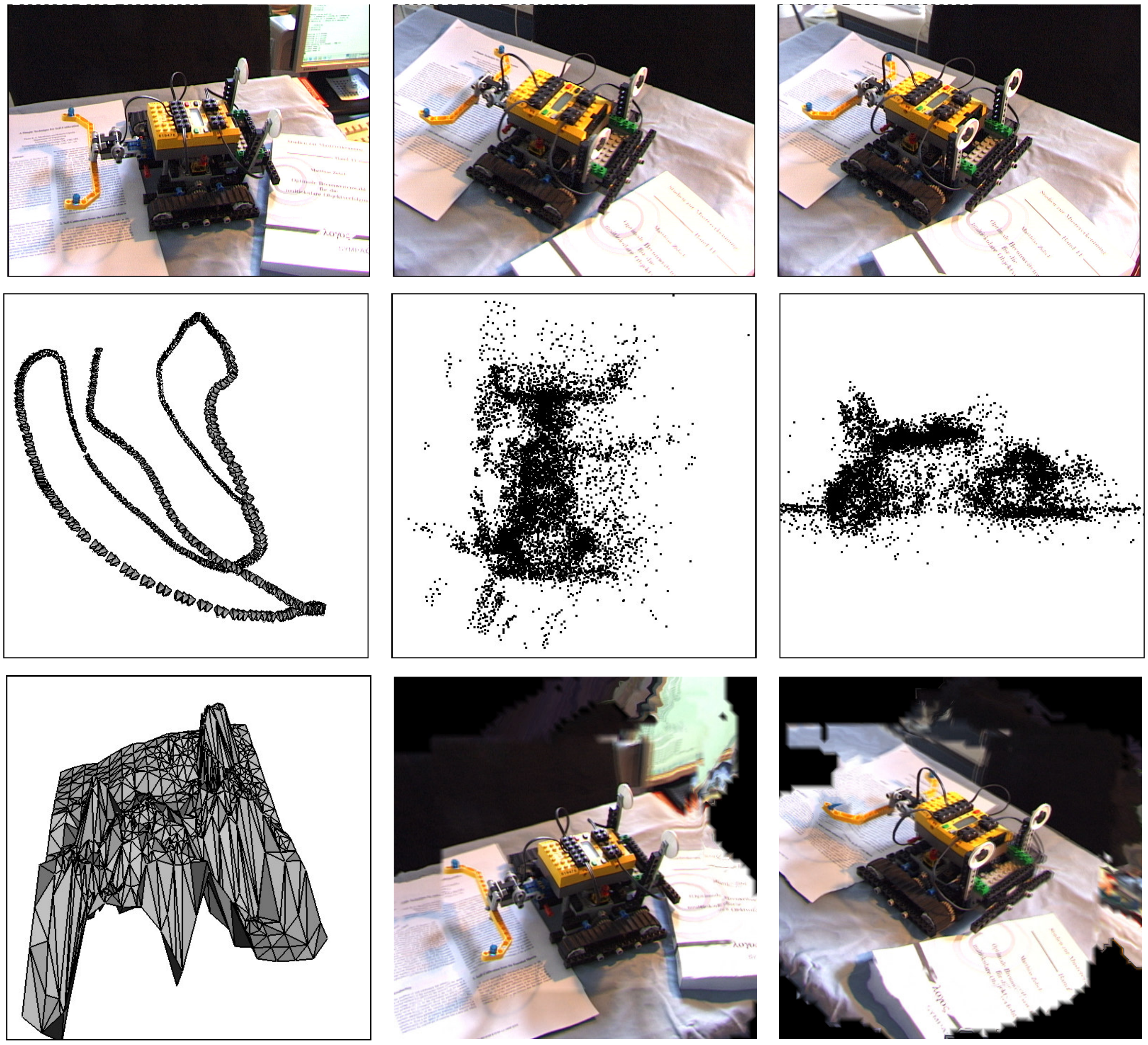


Figure 7.32: Light field reconstruction of the sequence `ART 118` (toy crawler, with a conventional video camera) using smARTtrack1. The top row shows three examples of the original sequence. The middle row displays the reconstructed camera path (left image), where each camera pose is visualized by a pyramid, and two views of the reconstructed 3-D points (middle and right image). The bottom row shows an example of a computed 3-D triangular mesh (left image) and two images rendered from the light field by using the unstructured lumigraph rendering approach (middle and right image).

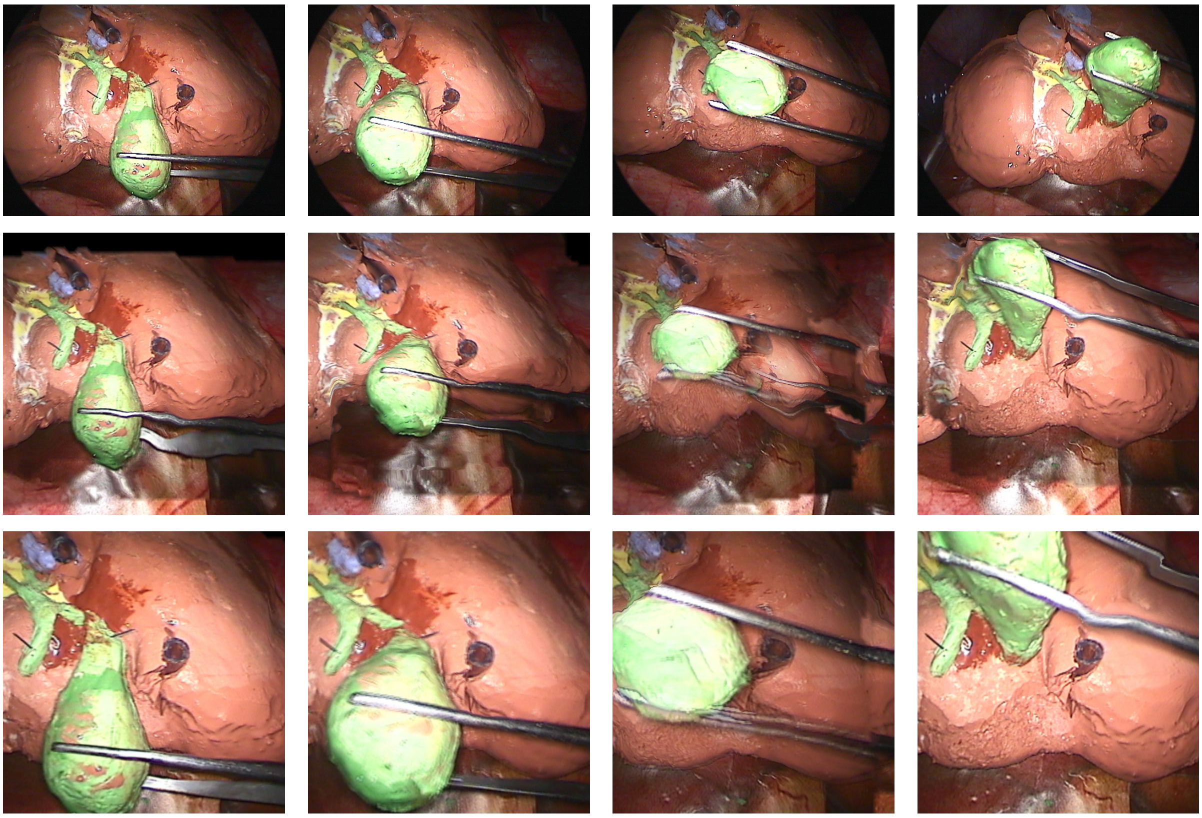


Figure 7.33: Dynamic light field reconstruction of the sequence `ART 82` (laboratory, liver/gall bladder model) using smARTtrack1. Four static light fields were reconstructed with 101, 106, 91, and 85 frames, respectively. The top row shows original images for each of the four time steps. The liver/gall bladder model was used and the gall was moved "upwards" by a pair of tweezers. This corresponds to the movement of the gall bladder during the operation. The middle and bottom row show rendered images from the dynamic light field. The camera parameters were the same for each image in a row, i. e., the dynamics visible in the original images should correspond to the dynamics in the rendered images.

7.3.4 Dynamic Light Fields Using smARTtrack1

This section presents two examples of dynamic light fields. In this thesis, dynamic light fields are obtained by reconstructing several static light fields for points in time where a static scene is assumed (cf. Section 3.4, page 53). As static light field reconstruction was examined extensively in the previous section, only examples of rendered images are displayed in this section. Images rendered from a dynamic light field reconstructed from the laboratory sequence `ART 82` are shown in Figure 7.33. The time steps were defined manually.

Several static light fields were reconstructed during a minimally invasive operation. These light fields implicitly define a dynamic light field. An example of such a dynamic light field with five time steps acquired during a cholecystectomy is shown in Figure 7.34. Dynamic light fields

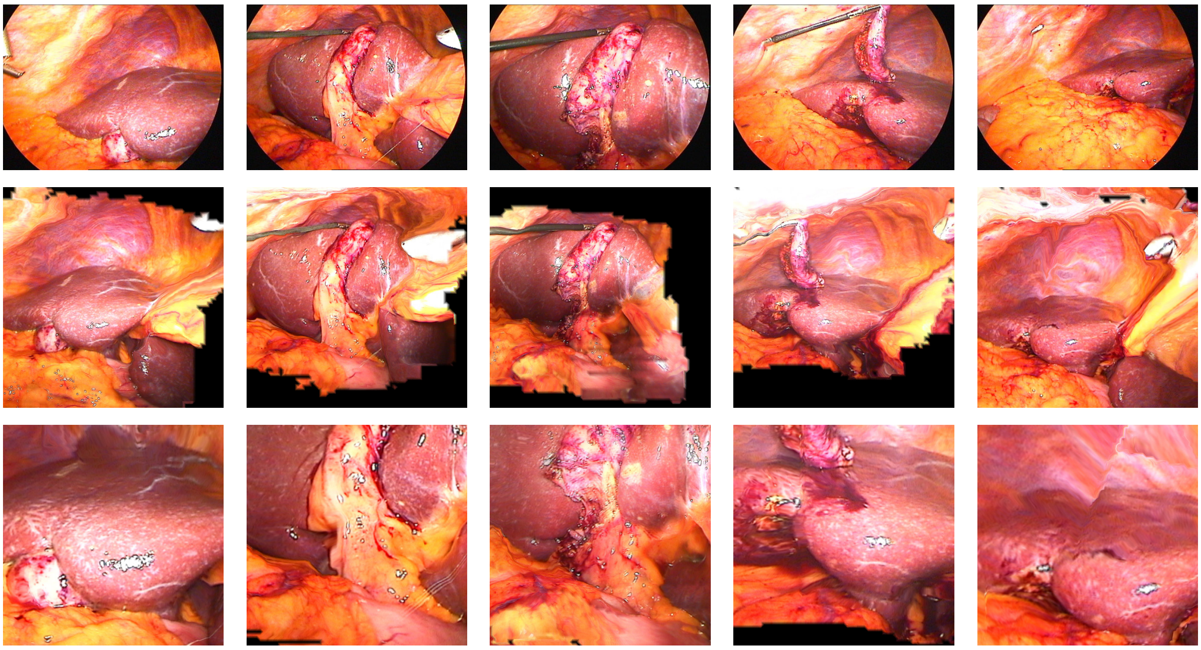


Figure 7.34: Dynamic light field reconstruction in the operating room using smARTrack1. Five sequences that were acquired during the operation are used for the dynamic light field (`ART 93`, `ART 94`, `ART 99`, `ART 103`, and `ART 104`). The number of frames were 342, 167, 251, 152, and 192, respectively. The top row shows original images for each of the five time steps. The first time step shows the untouched operating field with the gall bladder located beneath the liver. The second time step shows the situation shortly before the dissection of the gall bladder from the liver bed. The third time step shows the operating field after ligating and cutting the cystic duct. The fourth time step shows the gall bladder shortly before its removal. The fifth time step shows the operating field after the gall bladder has been removed. The middle and bottom row show rendered images from the dynamic light field. The camera parameters were the same for each image in a row, i. e., the dynamics visible in the original images should correspond to the dynamics in the rendered images.

allow the dynamics of the operation site to be viewed from an arbitrarily defined view point.

7.4 Image Enhancement by Light Fields

When a static light field of a scene is available, detectable image degradations that do not remain at the same position with respect to the scene can be reduced or even removed using the light field (cf. Section 4.6, page 77). Exemplarily, degradations caused by highlights are regarded. Additionally, results obtained with *simulated soilings* on the endoscope lens are presented.

First of all the degradations have to be detected. Figure 7.35 visualizes the results of two algorithms for highlight detection: color gradients and thresholds in *HSV* color space. For $H \in [0, 359]$, $S \in [0, 255]$ and $V \in [0, 255]$, the following thresholds were used: $0 \leq H \leq 359$ and

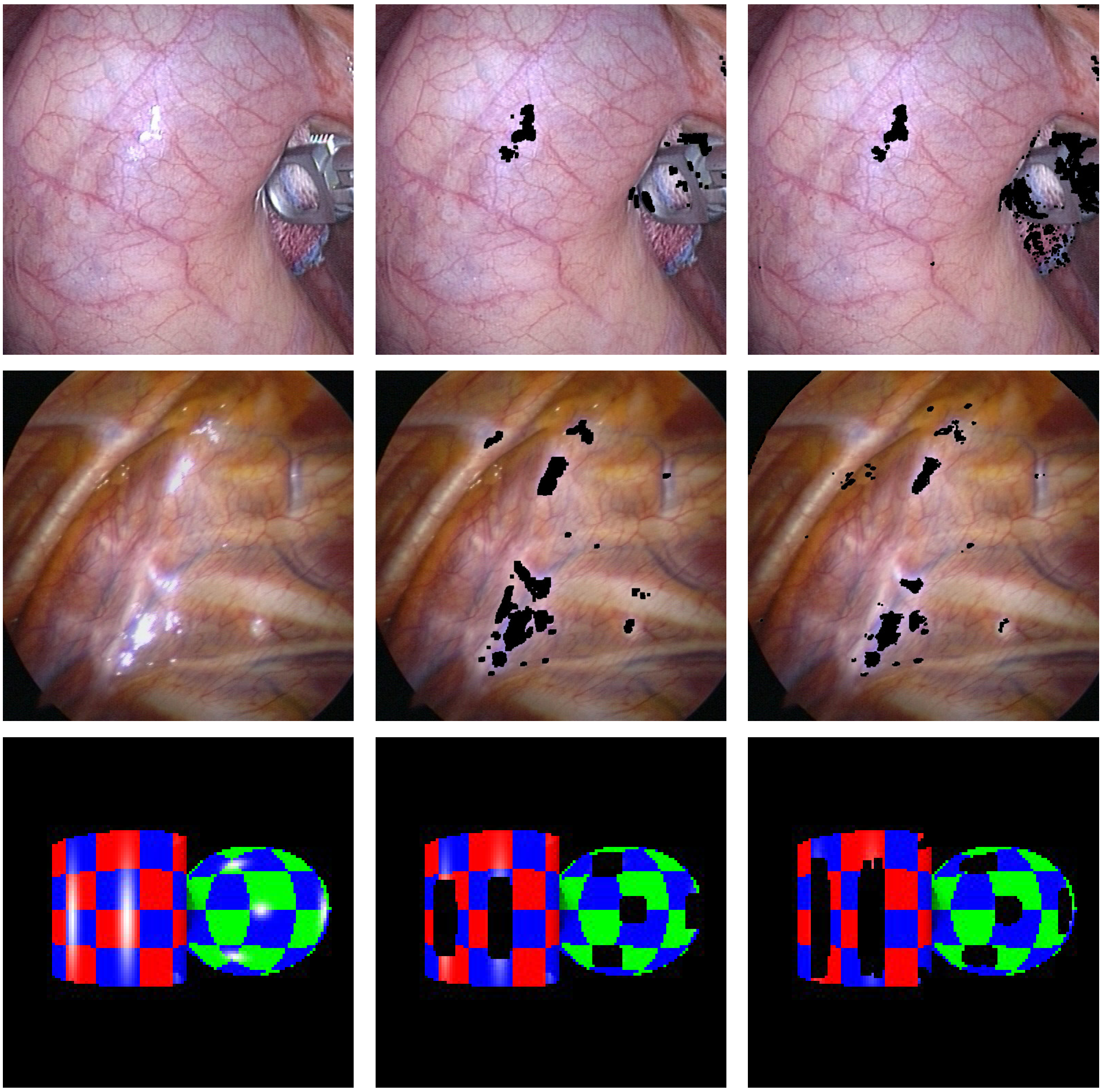


Figure 7.35: Examples of highlight detection. For each row, the original image (left) and the detected highlights (middle and right) are shown. Highlights are marked by black pixels. The middle image shows detected highlights by thresholds in *HSV* color space and the right image detected highlights by color gradients with subsequent region filling. The top row shows a gall bladder, the middle row a part of the thoracic cavity, and the bottom row an artificial image of a sphere and a cylinder.

$V \geq 200$ (all rows), $0 \leq S \leq 20$ (top row), $0 \leq S \leq 40$ (middle row), and $0 \leq S \leq 80$ (bottom row). The thresholds for the color gradients were: *RGB*-gradient > 70 (all rows), $c_1c_2c_3$-gradient > 0.1 (top row), $c_1c_2c_3$-gradient > 0.05 (middle and bottom row), $l_1l_2l_3$-gradient ≤ 0.9 (top row), $l_1l_2l_3$-gradient ≤ 0.6 (middle row), and $l_1l_2l_3$-gradient ≤ 0.3 (bottom row), where $c_1c_2c_3$ and $l_1l_2l_3$ are the names of the two computed color spaces (see [Gev99]). Since the difference

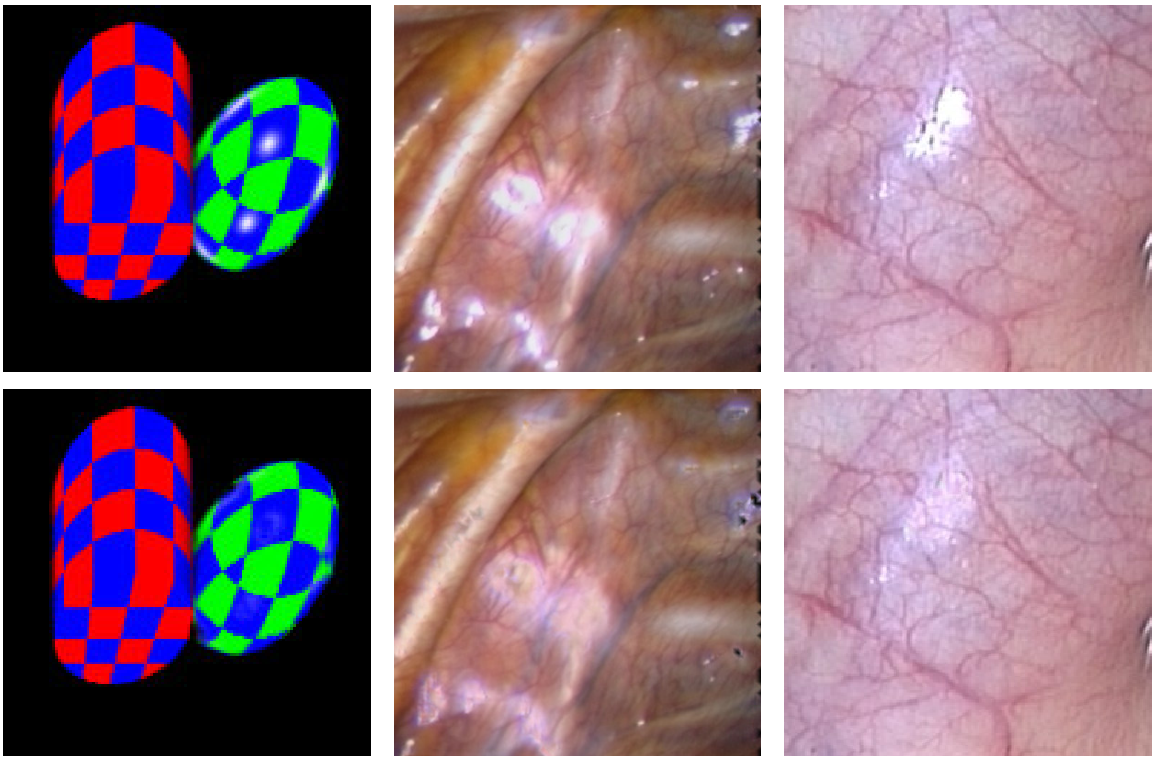


Figure 7.36: Examples of highlight substitution. The original image is always displayed on top of the processed/substituted image. The image pairs show a synthetic sphere and cylinder (left), a part of the thoracic cavity (middle), and a gall bladder (right).

between the results was small, *HSV*-thresholds were employed for highlight detection. The third method described in Section 4.6.1, page 79, which was proposed in [Pal99], did not yield usable results as the computed color of the light source was white. This does not allow determining the diffuse color of white highlights.

Results of highlight substitution are illustrated in Figure 7.36. Figure 7.37 displays results of the substitution of simulated soilings. Soilings on the endoscope lens were simulated by four circles with a diameter of 40 pixels, where the size of the cropped original image was 512×512 pixels. As the soilings were simulated they did not have to be detected. The confidence mask was known exactly in this case. Regarding Figures 7.36 and 7.37, it can be seen that highlights and simulated soilings could be removed almost completely.

The substitution of highlights and soilings was objectively evaluated using a synthetic sequence (see Figure 7.36, left) and simulated soilings for a laboratory sequence (see Figure 7.37). Thus, ground truth data were available for each sequence. The synthetic sequence was rendered once without highlights (ground truth) and once with highlights. The ground truth for the simulated soilings was the original sequence without soilings. Table 7.24 states the results for

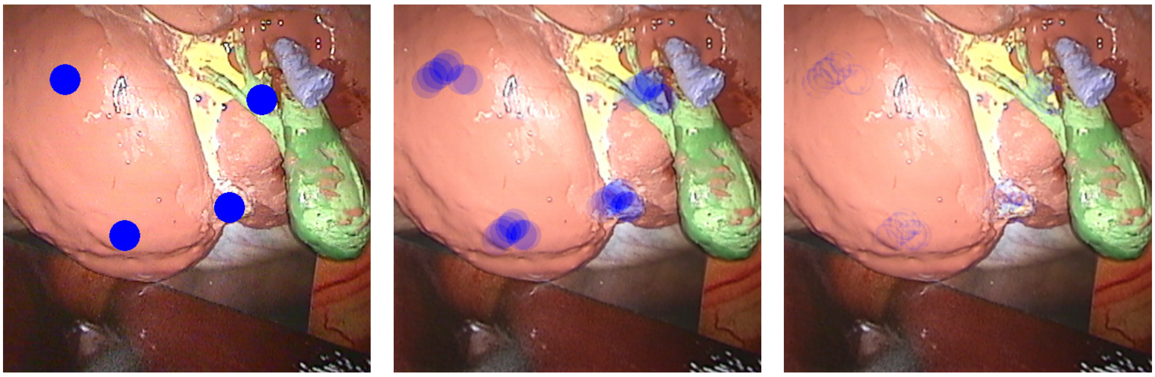


Figure 7.37: Examples of substituting simulated soilings on the camera lens. Four circles with a diameter of 40 pixels were overlaid over the original sequence (`ART 22`) of the liver/gall bladder model (left image). The middle image was rendered from the reconstructed light field without applying the substitution technique, for the right image the substitution technique was applied.

Comparison Method	LF - GT	LFsubst - GT	LF - LF GT	LFsubst - LF GT
$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	1.57	1.45	0.357	0.248
$\overline{Q}_{\mathrm{SNR}}$ [dB]	12.0	12.6	17.1	20.9
$\overline{Q}_{\mathrm{PSNR}}$ [dB]	26.8	27.4	32.4	36.1

Table 7.24: Objective Evaluation of highlight substitution. A synthetic sequence (sphere/cylinder) with 100 frames of size 256 × 256 pixels was used. The sequence was rendered once without highlights (ground truth) and once with highlights. For each ground truth image an image from the reconstructed light field without substitution was rendered and compared to the ground truth image (“LF - GT”). Secondly, the same image was rendered with substitution of the detected highlights (“LFsubst - GT”). Thirdly, the images rendered from the light field were compared to images rendered from a ground truth light field which was obtained by using the already computed camera parameters and depth information but exchanging the image data: instead of the highlight images the ground truth images were used for rendering (“LF - LF GT” and “LFsubst -LF GT”). $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ were computed for the 100 images.

highlight substitution and Table 7.25 states the results for the substitution of soilings. Firstly, for each original image an image from the light field without substitution was rendered and compared to the ground truth image (“LF - GT”). Secondly, the same image was rendered with substitution of the detected highlights/soilings (“LFsubst - GT”). Thirdly, the images rendered from the light field were compared to images rendered from a ground truth light field which was obtained by using the already computed camera parameters and depth information but exchanging the image data: instead of the disturbed images the ground truth sequence was used for rendering (“LF - LF GT” and “LFsubst -LF GT”). $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ were com-

Comparison Method	LF - GT	LFsubst - GT	LF - LF GT	LFsubst - LF GT
$\overline{Q}_{\mathrm{MAD}}$ [gray-values]	7.56	4.71	4.89	1.99
$\overline{Q}_{\mathrm{SNR}}$ [dB]	17.5	25.5	18.0	29.7
$\overline{Q}_{\mathrm{PSNR}}$ [dB]	22.3	30.3	22.7	34.4

Table 7.25: Objective evaluation of substituting soilings on the endoscope length. Four circles with a diameter of 40 pixels were overlaid over the "ground truth" sequence `ART 22` of the liver/gall bladder model (cf. Figure 7.37). This sequence contained 155 frames of size 512×512 pixels. For each ground truth image an image from the reconstructed light field without substitution was rendered and compared to the ground truth image ("LF - GT"). Secondly, the same image was rendered with substitution of the soilings ("LFsubst - GT"). Thirdly, the images rendered from the light field were compared to images rendered from a ground truth light field which was obtained by using the already computed camera parameters and depth information but exchanging the image data: instead of the disturbed images the ground truth images were used for rendering ("LF - LF GT" and "LFsubst -LF GT"). $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$ were computed for the 155 images.

puted, where the sphere/cylinder sequence with the synthetic highlights contained 100 frames of size 256×256 pixels and the `ART 22` sequence, that was used to simulate soilings, contained 155 frames of size 512×512 pixels. $\overline{Q}_{\mathrm{MAD}}$ was reduced by substituting the highlights but the quantitative difference was small as the highlight regions covered only a *small* part of the image and $\overline{Q}_{\mathrm{MAD}}$ is computed for the *whole* image. For the substitution of the simulated soilings, the quantitative difference was large: $\overline{Q}_{\mathrm{MAD}}$ was reduced by 38 % from 7.56 gray-values to 4.71 gray-values ("LF - GT" compared to "LFsubst - GT"). The improvement was even larger for the comparison to the ground truth light field: 4.89 gray-values compared to 1.99 gray-values (59 % reduction).

As the results of the objective evaluation were very clear, the subjective evaluation was performed by one surgeon only. 100 image pairs were rendered from two light fields (`Gall-Tape4` and `Hyp-20010425`), i. e., 50 pairs from each light field. The surgeon had to decide which image she prefers with respect to the disturbance by highlights. She preferred all 50 substituted images of the `Gall-Tape4` sequence and 45 substituted images of the `Hyp-20010425` sequence. For the five images of the `Hyp-20010425` sequence where the original image was preferred, too little information to substitute the highlights was available. The highlight regions were then substituted by black pixels which is obviously worse than doing nothing.

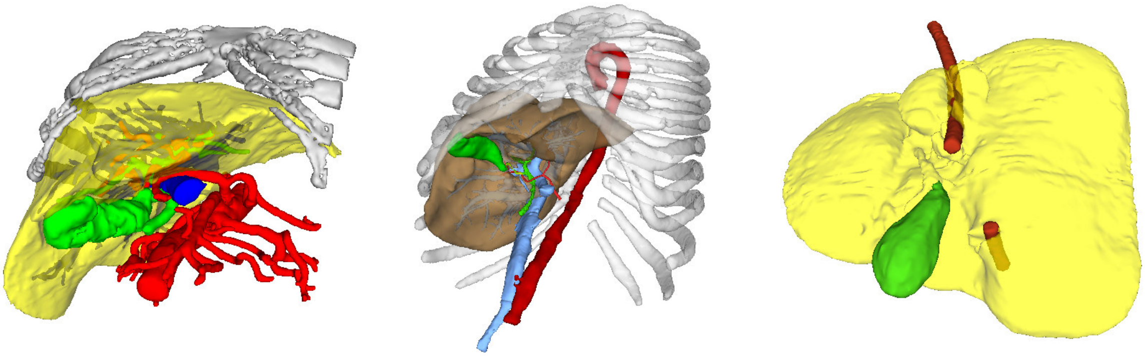


Figure 7.38: Examples of segmented CT datasets. The displayed triangular meshes were computed by the marching cubes algorithm based on the segmentation results. One of the datasets provided by MeVis (left), the "VOXEL-MAN" dataset (middle), and the liver/gall bladder model dataset (right) are displayed. The colors of the anatomical structures are: liver – yellow (left and right), liver – brown (middle), gall bladder and cystic duct – green, arteries – red, veins – blue, ribs and bones – gray, tubes simulating vessels – red (right).

7.5 Augmented Reality: Registration and Fusion with 3-D Data

Providing augmented reality during minimally invasive operations requires a registration of virtual data with the endoscope. Here, CT data were employed as virtual data for augmentation. The collected data, i. e., the anatomical database, consisted of nine datasets provided by MeVis as part of a research cooperation, four datasets from the Institute of Radiology, University of Erlangen-Nuremberg, and the "VOXEL-MAN". Additionally, a CT dataset of the liver/gall bladder model was acquired which was also part of the anatomical database.

The datasets were either pre-segmented (MeVis and "VOXEL-MAN") or segmented by using MeVisLab (liver/gall bladder model) or by a program developed at the Neurocenter of the University of Erlangen-Nuremberg (Institute of Radiology datasets). A segmentation of a liver and a gall bladder with the latter program took 25 hours, whereas a segmentation of a liver, a gall bladder and two tubes of the liver/gall bladder model took only three hours using MeVisLab. MeVisLab offers more sophisticated methods for semi-automatic segmentation. After segmentation, triangular meshes were computed by the marching cubes algorithm[7] based on the segmentation result (cf. Section 6.3.2, page 127). Examples of triangular meshes of the segmented datasets are shown in Figure 7.38. The resolutions and slice sizes of the datasets were:

- MeVis: Datasets with different slice sizes ranging from $0.62 \times 0.64 \times 1.25$ mm to $0.87 \times 0.87 \times 1.25$ mm and different resolutions, e. g., $437 \times 264 \times 142$ voxels and $366 \times 310 \times$

---

[7] The author thanks Marco Winter, Department of Computer Graphics, University of Erlangen-Nuremberg, for the cooperation.

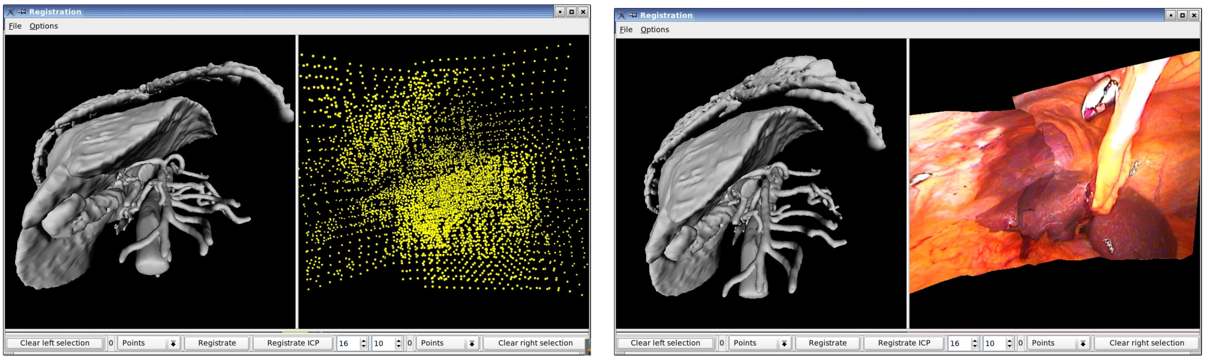


Figure 7.39: Screenshots of the program developed for point/triangle selection. At the beginning only the cloud of computed 3-D surface points was used to select points in the endoscope coordinate system (left): the 3-D point cloud is displayed to the right of the CT data. An improvement is achieved by additionally providing texture information (right): the textured 3-D triangular mesh instead of the 3-D point cloud is displayed to the right of the CT data. The identification of landmarks based only on the computed surface points is tedious compared to when using texture information.

152 voxels.

- Institute of Radiology, University of Erlangen-Nuremberg: Resolution $512 \times 512 \times 512$ voxels with a slice size of $0.74 \times 0.74 \times 0.5$ mm.
- “VOXEL-MAN”: Resolution $573 \times 330 \times 774$ voxels with a slice size of $1.0 \times 1.0 \times 1.0$ mm.
- Liver/gall bladder model: Resolution $512 \times 512 \times 375$ voxels with a slice size of $0.74 \times 0.74 \times 1.0$ mm.

Figure 7.39 displays screenshots of the program developed for selecting the 3-D point correspondences that are necessary for coarse registration. The determination of point correspondences based only on the computed 3-D surface points is tedious. The point selection step was therefore improved by additionally providing texture information based on the computed 3-D triangular mesh. The program allows selecting either points or triangles in 3-D with the mouse.

The points used for registration were selected in cooperation with a physician. It turned out that the costal arch could be used for registration, but also that the relative pose of the costal arch changes with respect to the liver and gall bladder due to the introduction of gas into the abdominal cavity. Thus, the costal arch was used but the results obtained had to be refined, either by applying the ICP algorithm or by manual interaction. For the ICP algorithm the maximal allowed distance of point pairs was set to 10 mm and usually 16 iterations were sufficient. Admittedly, it was faster and often more accurate to refine the coarse registration manually.

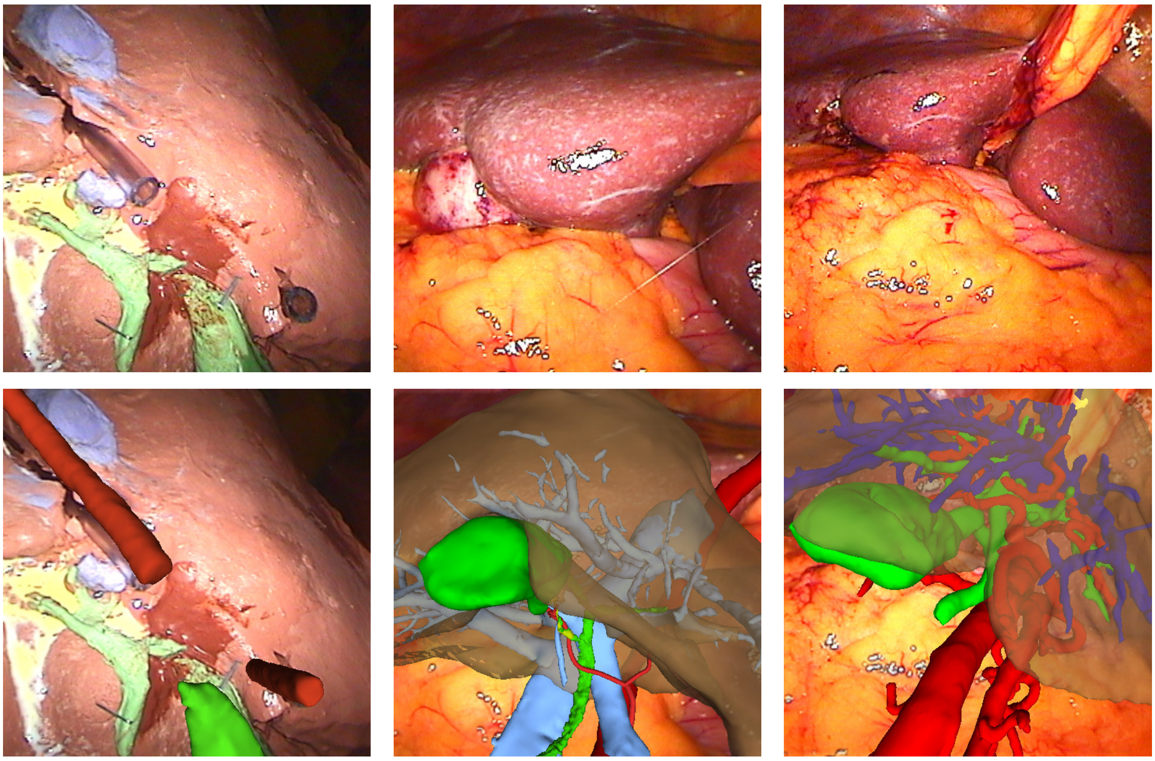


Figure 7.40: Examples of 2-D augmented reality: the live image (top row) is augmented by overlaying CT data (bottom row). The top left image shows the liver/gall bladder model (sequence `ART 75`) onto which the gall bladder and two tubes simulating vessels are overlaid (bottom left). The top middle image shows the liver *and* the gall bladder (sequence `ART 93`), whereas in the top right image the gall bladder was removed (sequence `ART 104`). Onto both images (bottom middle and right) the liver is overlaid transparently. Additionally, the gall bladder and the cystic duct and important arteries and veins are overlaid. The colors of the overlaid anatomical structures are: liver – brown (middle and right), gall bladder and cystic duct – green, arteries – red, veins – blue, tubes simulating vessels – red (left).

After registered CT data were available, 2-D and 3-D augmented reality could be provided. Either the 2-D live image or the light field was augmented. Examples of both types of augmented reality are presented: 2-D live augmented reality in Figure 7.40 and 3-D light field augmented reality in Figure 7.41. Exemplarily, augmented reality results of three light fields are shown: one of the liver/gall bladder model (`ART 75`) and two computed at the beginning and the end of a cholecystectomy (`ART 93` and `ART 104`). The benefit of additional information is clearly visible when comparing the augmented images to the non-augmented ones. Not or only partly visible vessels are completely visible in the augmented images (e. g., see Figures 7.40 and 7.41, bottom left). This is especially important for vessels located very close to the dissection area that must not be injured during the operation: aorta, vena cava inferior, hepatic artery, portal vein, and main bile duct. These vessels are only visible in the augmented images in Figures 7.40

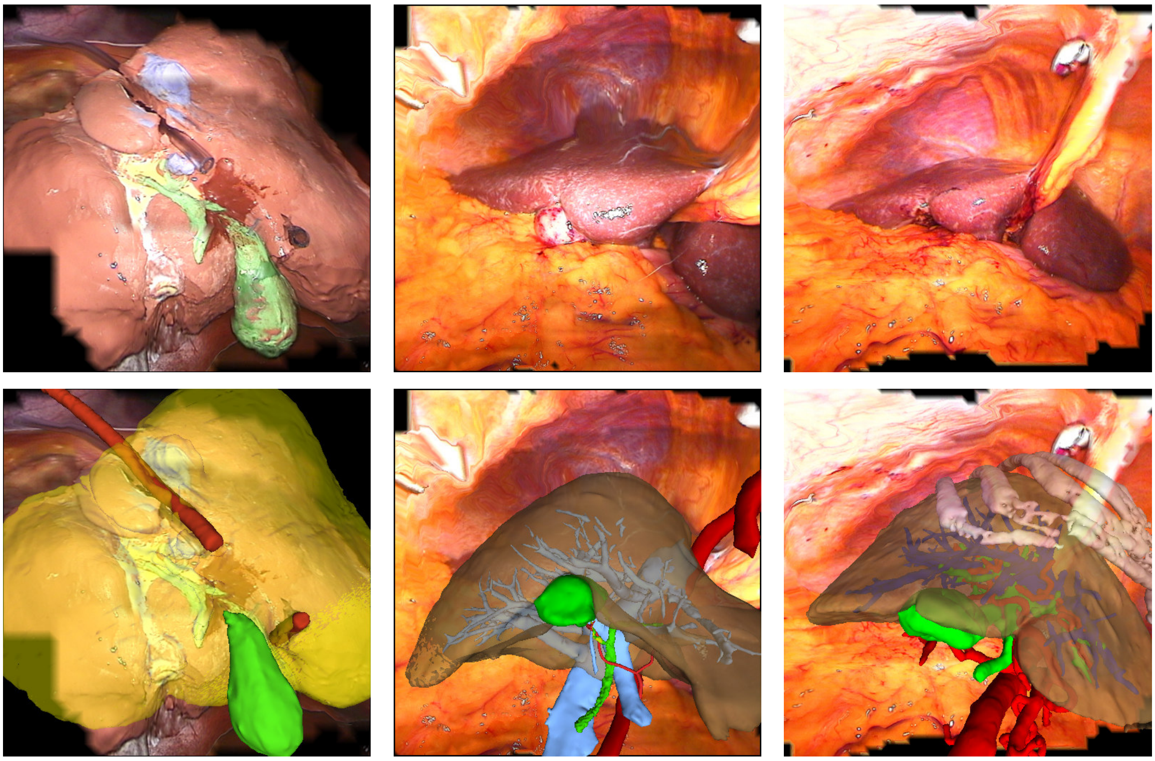


Figure 7.41: Examples of augmenting the light field. Three light fields were augmented: one of the liver/gall bladder model (left, sequence `ART 75`) and two computed at the beginning and the end of a cholecystectomy (sequence `ART 93`, middle, and sequence `ART 104`, right). The top row shows overview images rendered from the light fields (cf. Figure 7.40 for an example of an original image). The bottom rows displays the corresponding augmented images (liver, gall bladder and vessels). The colors of the overlaid anatomical structures are: liver – yellow (left), liver – brown (middle and right), gall bladder and cystic duct – green, arteries – red, veins – blue, ribs and bones – gray, tubes simulating vessels – red (left).

and 7.41 (middle and right). In order to simplify orientation, liver and gall bladder are overlaid additionally. First results of the proposed augmented reality visualization were also published in [Vog05b, Vog04b, Nie04, Nie03], yet without light fields computed during real operations and with only a part of the described anatomical database.

The benefit of augmented reality was only evaluated qualitatively: four augmented light fields were presented to 10 surgeons and they were asked what they think about this new possibility. The tenor was that especially for more complicated situations the augmentation by CT data is advantageous and improves the overview.

7.6 Discussion

The system developed for supporting the surgeon during endoscopic surgery provides real-time image enhancement, 3-D light field visualization of the operation site, and augmented reality. It was used in the operating room and evaluated subjectively by physicians as well as objectively. As already mentioned in Section 2.4, the first publication on real-time *distortion correction* was published in 2001 [Hel01]. In the same year first results of the work described in this thesis were published [Vog01a, Vog01b], including distortion correction, color normalization, and temporal filtering. The complete system for real-time endoscopic image enhancement was presented in 2003 together with a subjective evaluation of the image enhancement methods by physicians [Vog03a]. This was one year before the system of Fischer et al. [Fis04] was presented.

Distortion correction, color normalization, and temporal filtering can be applied in real-time on a modern PC (Pentium 4, 3.2 GHz). Even for the combination of all three methods 13 fps can be provided. An extensive subjective evaluation with 14 surgeons showed a statistically significant benefit of the processed images ($p \ll 10^{-4}$). In contrast to color normalization and distortion correction, it turned out that the benefit of temporal filtering is only visible when viewing image sequences and not single images. An objective evaluation was performed for temporal filtering which reinforced the improved image quality of a temporally filtered sequence.

Light field reconstructions in the laboratory and in the operating room by structure-from-motion techniques, by using the robot arm AESOP, and by using the optical tracking system smARTtrack1 were shown. The objective was to develop fast methods for light field reconstruction in the operating room using pose determination systems. This objective was reached: using AESOP or smARTtrack1 the reconstruction of a light field with 155 frames took approximately one minute (71 sec). A comparable reconstruction without using a pose determination system, i. e., based on the structure-from-motion algorithm, took almost ten minutes. Even for a larger sequence with 511 frames a light field could be reconstructed in five minutes, whereas it took 18 min using structure-from-motion. Apart from the computation time the main disadvantage of the structure-from-motion approach is its sensitivity to the input parameters: a small change of the tracking or reconstruction parameters may lead to a different and sometimes unusable result. Thus, the parameters often have to be adapted for each sequence which takes additional time and is not feasible in the operating room. In contrast to this, only one parameter, namely the threshold θ_{BPE} for the back-projection error, had to be adapted when using AESOP or smARTtrack1. When the reconstruction result was usable, the quality of the structure-from-motion light field was either better (`ALF 67`) or equal (`ART 22` and `ART 93`) to the quality of the light field reconstructed by using a pose determination system.

The achieved endoscope pose accuracy using AESOP and smARTtrack1 was determined. The relative errors using AESOP were about ten times larger than the errors using smARTtrack1:

- AESOP: $\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}} = 56\,\%$ (3.9 mm) and $\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}} = 26\,\%$ (1.9°).
- smARTtrack1 with the *DD* target: $\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}} = 3.8\,\%$ (1.5 mm) and $\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}} = 2.7\,\%$ (0.63°).

The endoscope plug of AESOP is the reason for the larger errors. It was not designed for high accuracy positioning of an endoscope. The slackness is too large: the tip of the endoscope can be moved several millimeters although the endoscope plug is kept fixed. The large errors of AESOP already indicate that the quality of light fields reconstructed using AESOP will be lower than the quality of a light field reconstructed using smARTtrack1. Regarding the hand-eye calibration errors, the endoscope pose errors, and the 3-D shape errors the *DD* target is the one that should be used for light field reconstruction with smARTtrack1.

The accuracy of the 3-D reconstruction using smARTtrack1 was evaluated additionally. Due to the large errors using AESOP such an evaluation did not make sense in this case. The surface of a sphere with known radius was reconstructed. The obtained errors using the *DD* target were $0.92\,\%$ (0.21 mm), where the shape error $\epsilon_{\mathrm{shape}}$ was 0.36 mm. This means that despite endoscope pose errors of $3\,\%$ to $4\,\%$, the error of the 3-D reconstruction of the scene surface was only $1\,\%$. An explanation for the smaller error is the use of non-linear optimization and LMedS for the computation of 3-D points. This is reinforced by comparing the quality of light field reconstructions with and without using these techniques. Usually a major improvement was achieved, especially for non-linear optimization. The additionally required computation time is justifiable with respect to the achieved improvement: it took only 17 sec longer using LMedS and non-linear optimization for a sequence of 155 frames and 75 sec longer for a sequence of 511 frames.

The quality of the reconstructed light fields was evaluated objectively and subjectively. The subjective evaluation of 10 physicians showed a clear difference between the quality of laboratory light fields (grade 2.5) and operating room light fields (grade 3.3). For the objective evaluation in terms of $\overline{Q}_{\mathrm{PSNR}}$ it is stated in [Wan02] for video compression that $\overline{Q}_{\mathrm{PSNR}} > 30\,\mathrm{dB}$ is very good, $20\,\mathrm{dB} \leq \overline{Q}_{\mathrm{PSNR}} \leq 30\,\mathrm{dB}$ is good, and $\overline{Q}_{\mathrm{PSNR}} < 20\,\mathrm{dB}$ is bad. According to this classification, AESOP light fields have an almost good quality ($16 \leq \overline{Q}_{\mathrm{PSNR}} \leq 20$), smARTtrack1 light fields have a good, sometimes almost very good quality ($24 \leq \overline{Q}_{\mathrm{PSNR}} \leq 30$), and structure-from-motion light fields have also a good and sometimes very good quality ($20 \leq \overline{Q}_{\mathrm{PSNR}} \leq 30$). For the light fields reconstructed in the operating room, $\overline{Q}_{\mathrm{PSNR}}$ was approximately 24 dB, whereas for laboratory light fields $\overline{Q}_{\mathrm{PSNR}}$ was almost 30 dB. Image enhancement by light fields was also evaluated subjectively and objectively. Both evaluations clearly showed an improvement; for instance, $\overline{Q}_{\mathrm{PSNR}}$ was increased from 22 dB to over 30 dB.

Two results of dynamic light fields were presented. Since dynamic light fields are modeled by several static light fields, all results obtained for static light fields apply to them as well.

The benefit of augmented reality, for which three examples were presented, was only evaluated qualitatively. The tenor of 10 surgeons was that especially for more complicated situations the augmentation by CT data is advantageous and improves the overview.

Finally, a remark about different processor types is necessary. It turned out that due to the use of IPP slightly different results were obtained for point tracking on different processors. As the methods for depth computation and extrinsic parameter estimation for structure-from-motion rely on point tracking, the computed results were also slightly different on different processors. Differences occurred especially for light field evaluations based on the quality criteria $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$, and $\overline{Q}_{\mathrm{PSNR}}$. The light field evaluation results presented here were computed on a Dell Latitude D800 notebook with an Intel Pentium M processor (2.0 GHz). The experiments were additionally performed on a PC with an Intel Pentium 4 processor (3.2 GHz) and on a PC with an AMD Athlon 2600+ processor (1.9 GHz). The results only differed slightly and did not lead to contradictory conclusions. Note that the computation times presented in this chapter were obtained on the PC with the 3.2 GHz Intel Pentium 4 processor. All other results were obtained on the Dell notebook.

Chapter 8

Summary and Outlook

This chapter summarizes the work in Section 8.1 and then concludes with an outlook in Section 8.2.

8.1 Summary

The tendency in the field of surgery is moving towards *minimally invasive* operations which traumatize the patient considerably less than conventional open surgery. The idea of minimally invasive surgery is to access the operation site through small “keyholes” with a diameter of about 1 to 2 cm. The image of the operation site is obtained by using an endoscope. This work focuses on those minimally invasive operations where rigid monocular endoscopes are utilized, e. g., the removal of the gall bladder (cholecystectomy). Compared to conventional surgery, several problems arise. In this thesis, techniques for reducing three of these problems, namely, *image degradations*, *limited vision*, and *loss of stereoscopic depth perception* have been developed. A complete system for usage in the operating room has been described. It provides

- real-time image enhancement,
- 3-D visualization of the operation site, and
- augmented reality.

Several image degradations can be reduced or even removed in real-time. A 3-D model of the operation site, namely, a light field can be reconstructed and regarded in 3-D from arbitrary positions, e. g., on a 3-D monitor. Either the 2-D live image or the light field can be augmented with CT data after registration based on the reconstructed 3-D information has been performed.

Concerning image degradations, most already published solutions were not developed for usage in the operating room and, except for the work presented in [Fis04], only solutions for single image degradations have been published. Furthermore, none of the proposed methods was evaluated by surgeons. Concerning light fields, other 3-D models have been used in minimally invasive surgery but light fields have not yet been examined. Several approaches for medical augmented reality systems exist, even for minimally invasive surgery. Here, intrinsic registration was employed whereas most current augmented reality approaches employ extrinsic registration based on some kind of markers.

Three enhancement methods for endoscopic image degradations were proposed: distortion correction, color normalization and temporal filtering. Image distortions are mainly due to lenses with small focal length. They can be corrected based on the intrinsic camera parameters which are determined by camera calibration. During a minimally invasive operation, the tissue of the operation site may be covered with blood. In this case, it is difficult to identify different tissue types. Color normalization reduces this problem and additionally provides illumination independent images. During the cutting of tissue with high frequency diathermy, smoke and small flying particles are generated. These disturbing degradations are reduced by temporal filtering. All three methods for image enhancement can be applied in real-time using a modern PC (Pentium 4, 3.2 GHz). Even for the combination of the three methods 13 frames per second can be provided. The evaluation of the methods by 14 surgeons showed a statistically significant benefit of the processed images ($p \ll 10^{-4}$). It turned out that the benefit of temporal filtering is only visible when viewing image sequences and not single images. Apart from image enhancement methods the system provides digital zooming and allows to keep the horizon steady when rotating endoscope optics and camera together. A technique for removing image degradations like highlights or soilings on the endoscope lens was also presented. This approach is based on a static light field and is capable of removing degradations that do not remain at the same position with respect to the scene while the endoscope is moved.

The challenges during the reconstruction of static light fields from endoscopic images are the determination of extrinsic camera parameters and the computation of depth information. Three possible solutions for the computation of the extrinsic camera parameters were examined:

- using structure-from-motion techniques,
- using the endoscope positioning robot AESOP, and
- using the optical tracking system smARTtrack1.

One assumption was made for all three methods, namely that the intrinsic camera parameters are constant for all captured images. These parameters are then estimated in advance by a camera calibration technique. Apart from the structure-from-motion approach, depth information in terms of 3-D points is computed by tracking 2-D points from image to image and triangulating 3-D points according to the known intrinsic and extrinsic camera parameters. A new representation for the depth information for light field rendering in terms of 3-D triangular meshes was introduced. This representation reduces the necessary time for depth computation and accelerates the rendering of images. An LMedS technique and non-linear optimization of the extrinsic camera parameters was proposed to improve the quality of the computed 3-D points.

Techniques for determining the hand-eye transformation for AESOP and smARTtrack1 were developed. Three targets were designed for usage with smARTtrack1 and their accuracy was examined. The *DD* target yielded the lowest endoscope pose errors: $\overline{\epsilon}_{\boldsymbol{t},\text{rel}} = 3.8\,\%$ $(1.5\,\text{mm})$ and $\overline{\epsilon}_{\boldsymbol{R},\text{rel}} = 2.7\,\%$ $(0.63°)$. The errors using AESOP were about ten times larger, where the slackness of the endoscope plug is the main reason for this extreme difference.

Light fields have been reconstructed by all three methods. Most light fields using one of the two pose determination systems were reconstructed in the laboratory, but each system was also used in the operating room. Both pose determination systems allow for the fast reconstruction of light fields in the operating room: the reconstruction of a light field with 155 frames took approximately one minute. In contrast to this, the corresponding reconstruction by applying the structure-from-motion algorithm took almost ten minutes. The quality of the reconstructed light fields was evaluated subjectively and objectively. The subjective evaluation by ten physicians showed a clear difference between the quality of laboratory light fields (grade 2.5) and operating room light fields (grade 3.3). The evaluation in terms of $\overline{Q}_{\text{MAD}}$, $\overline{Q}_{\text{SNR}}$, and $\overline{Q}_{\text{PSNR}}$ allows the quality of the light fields to be judged objectively. The objective evaluation showed the same difference between laboratory and operating room light fields. According to the grades defined in [Wan02], AESOP light fields have an almost good quality ($16 \leq \overline{Q}_{\text{PSNR}} \leq 20$), whereas smARTtrack1 light fields have a good and sometimes even very good quality ($24 \leq \overline{Q}_{\text{PSNR}} \leq 30$). The quality of the light fields reconstructed by structure-from-motion was comparable. In this work dynamic light fields are modeled by several static light fields. Thus, all results obtained for static light fields apply to them as well. Two results of dynamic light fields were presented, one was computed in the laboratory, the other during a cholecystectomy.

Apart from the computation time the main disadvantage of the structure-from-motion approach is its sensitivity to the input parameters: a small change of the parameters may lead to a different and sometimes unusable result. Thus, the parameters usually have to be adapted for each sequence which is not feasible in the operating room. In contrast to this, only the threshold θ_{BPE} for the back-projection error has to be adapted when using AESOP or smARTtrack1. The main disadvantage of AESOP is the large error of the computed endoscope pose. Regarding these drawbacks and the other advantages and disadvantages described in Section 5.8, page 118, it is proposed to use smARTtrack1 together with the *DD* target to reconstruct high quality light fields during minimally invasive operations. Furthermore, the light field quality can be improved by applying the proposed LMedS technique for triangulation and by non-linearly optimizing the extrinsic camera parameters.

The system provides augmented reality by overlaying CT data of an anatomical database either onto the rendered light field image or over the live image. In the first case the scene can be viewed in 3-D, in the second case only 2-D images can be viewed. In order to achieve real stereoscopic depth perception, a 3-D monitor is employed in the first case. A new method for intrinsic registration of CT data and endoscope has been developed. Based on the computed depth information, 3-D point correspondences for coarse registration can be selected by the surgeon. Afterwards, an iterative-closest-point (ICP) algorithm may be applied for fine registration. Additionally, the registration can be refined manually by the surgeon. Augmented reality enables the surgeon to “see” beyond the surface, through organs and tissue, e. g., important anatomical structures like vessels that must not be injured during an operation become completely visible even if the structures are not or only partly visible in the endoscopic image.

8.2 Outlook

A complete system for real-time image enhancement, 3-D visualization, and augmented reality for computer assisted endoscopic surgery was developed in this thesis. Although it provides several solutions for problems arising during minimally invasive surgery, extensions are possible. This section summarizes some of the ideas for extending the system.

Apart from further accelerating the already implemented algorithms for real-time image enhancement, new algorithms could be included. For instance, the problem of inhomogeneous illumination was not addressed here. A solution for this problem was presented recently in [Fis04] and could be integrated into the system. Additionally, currently only a very small fraction of the smoke resulting from cutting tissue with high frequency diathermy can be reduced by temporal filtering. This is due to the often slow movement of the “smoke clouds”. Nevertheless, provided that smoke clouds can be detected and assuming no movement inside the scene, a kind of temporal filtering could be used to substitute smoke pixels: the last valid color is used for the detected smoke pixel. Naturally, the benefit of the new methods would have to be evaluated.

The known camera parameters could also be used for accelerating and improving point tracking. Theoretically, the search range of a corresponding feature point in the next image can be restricted to the *epipolar line* which can be computed based on the known camera parameters. However, even the small pose errors when using smARTtrack1 are presumably too large to take advantage of this idea in practice. Another idea for improving point tracking based on the known camera parameters is the recovery of points that were lost because they were not visible for a certain time, e. g., because of occlusion. Currently, such a point will be lost and detected as a new point when it becomes visible again. Thus, two trails are obtained for this point yielding two 3-D points instead of one. This could be prevented by back-projecting the 3-D point computed from the first trail into all subsequent images and trying to track this back-projected point. This means that when the point becomes visible again it can be found and assigned to the “old” trail instead of beginning a new one. The pose errors should be small enough for this idea to be realized.

Non-linear optimization of the extrinsic camera parameters could be improved by using the Kalman filter for predicting and smoothing the camera pose in order to obtain better initialization values, especially for outliers.

Although prohibiting the rotation of endoscope optics with respect to the camera head is no drawback because the horizon can be kept steady using a pose determination system, this restriction could be waived by attaching a target to the camera head. However, realizing this is difficult due to the shape of the camera head and the fact that it is wrapped in a sterile foil.

Regarding augmented reality, it would be advantageous to extend the registration procedure by including algorithms for non-rigid registration. Thus, based on the reconstructed 3-D information, pre-operatively acquired 3-D data like CT could be registered even if the anatomical structures were moved or deformed. The proposed rigid registration procedure could be improved by including algorithms for automatic coarse registration. Alternatively to a complete algorithm for coarse registration, a small amount of 3-D points that are well suited for registration could be pre-selected automatically and highlighted for the surgeon.

Naturally, other modalities than CT, especially MRI and 3-D ultrasound, could be used for augmented reality. As devices for intra-operative 3-D ultrasound exist, matching 3-D information could be obtained during the operation. Additionally, the registration of intra-operative ultrasound and endoscopic images could be simplified by tracking the ultrasound transducer. For the new modalities as well as for CT data, more sophisticated methods for visualizing the overlaid structures have to be developed. For instance, an approach that employs the computed depth information to adapt the color of the overlaid structures according to the depth relative to the surface of the scene was presented in [Win05].

Future research should also take into account stereo endoscopes. The usage of stereo endoscopes would simplify depth computation because stereo algorithms could then be employed. Additionally, stereo endoscopes allow viewing the augmented live image in 3-D.

Finally, the benefit of the proposed system should be shown by an evaluation after it has been used during several minimally invasive operations. For this purpose, an appropriate interface would have to be developed, e. g., speech controlled navigation in 3-D as described in [Prü05]. Currently, the navigation in the light field and the control of the image enhancement methods is performed via the mouse, which can be wrapped in a sterile foil. However, this can only be seen as an intermediate solution since the operating surgeon usually needs both his hands during surgery. Additionally, a more detailed evaluation concerning augmented reality and light field quality, especially with respect to the stereoscopic 3-D impression and the benefit of augmented reality, would be useful as well.

Appendix A

Homogeneous coordinates

The elements of the n-dimensional projective space $\mathbb{P}^n$ are *homogeneous vectors*

$$\underline{\boldsymbol{x}} = (x_1, \ldots, x_{n+1})^{\mathrm{T}}, \tag{A.1}$$

where at least one of the x_i must be nonzero. Finite vectors are characterized by $x_{n+1} \neq 0$ and a finite homogeneous vector $\underline{\boldsymbol{x}}$ has a unique mapping to the Euclidean space. The coordinates of the corresponding Euclidean point $\boldsymbol{x}$ are

$$\boldsymbol{x} = \left(\frac{x_1}{x_{n+1}}, \ldots, \frac{x_n}{x_{n+1}} \right)^{\mathrm{T}}. \tag{A.2}$$

Therefore, two vectors $\underline{\boldsymbol{x}}_1$ and $\underline{\boldsymbol{x}}_2$ correspond to the same Euclidean point $\boldsymbol{x}$ if and only if there exists a nonzero scalar s such that $\underline{\boldsymbol{x}}_1 = s \cdot \underline{\boldsymbol{x}}_2$, since s is eliminated in equation (A.2). If $\underline{\boldsymbol{x}}_1 = s \cdot \underline{\boldsymbol{x}}_2$ the two vectors are *equal up to an unknown scalar* which is indicated by $\underline{\boldsymbol{x}}_1 \sim \underline{\boldsymbol{x}}_2$.

In addition to finite vectors, the projective space contains points lying at infinity, where a projective point $\underline{\boldsymbol{x}} = (x_1, \ldots, x_{n+1})^{\mathrm{T}}$ with $x_{n+1} = 0$ corresponds to an ideal point at infinity in the $(x_1, \ldots, x_n)^{\mathrm{T}}$ direction in Euclidean space [Moh96].

Appendix B

Singular Value Decomposition

Singular value decomposition (SVD) allows solving many tasks of linear algebra. Some of those solutions are described in this section (adapted from [Hei04]). The LAPACK implementation [And95] is used.

Let $\boldsymbol{X}$ be an $m \times n$ matrix with $m \geq n$. The singular value decomposition allows $\boldsymbol{X}$ to be decomposed into a product of three matrices:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}, \tag{B.1}$$

where the matrices $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$ have the following properties:

- $\boldsymbol{U}$ is a $m \times n$ matrix whose columns are orthonormal: $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_{n\times n}$
  (where $\boldsymbol{I}_{n\times n}$ is the $n \times n$ identity matrix).
- $\boldsymbol{V}$ is a $n \times n$ matrix whose columns are orthonormal: $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_{n\times n}$.
- $\boldsymbol{S}$ is a $n \times n$ diagonal matrix containing the so-called *singular values*:
  $\boldsymbol{S} = \mathrm{diag}(s_1, s_2, \ldots, s_n)$ with $s_i \geq 0\,,\ i = 1, \ldots, n$.

Without loss of generality it can be assumed that the singular values are ordered such that $s_1 \geq s_2 \geq \ldots \geq s_n$. The decomposition is unique if all singular values are different, i. e., $s_1 > s_2 > \ldots > s_n$. Otherwise, more than one decomposition fulfilling equation (B.1) and the properties above, exists. Finally, for $m < n$ the decomposition of $\boldsymbol{X}$ is obtained by decomposing $\boldsymbol{X}^{\mathrm{T}}$:

$$\boldsymbol{X} = \left(\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathrm{T}}. \tag{B.2}$$

The computational cost of the algorithm is $O\left(\max(m,n)^2 \cdot \min(m,n)\right)$.

The remaining paragraphs of this appendix describe selected applications of the singular value decomposition.

The rank of a matrix: Let k be the rank of matrix $\boldsymbol{X}$. According to the properties mentioned above, $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal and therefore have rank n. Hence, a reduction of the rank can only be caused by a reduced rank of $\boldsymbol{S}$. If $k < n$ it follows that $s_{k+1} = s_{k+2} = \ldots = s_n = 0$. This fact allows determining the rank of a matrix by testing how many singular values are larger than zero.

The null-space of a matrix (solution for homogeneous linear equation systems): Let matrix $\boldsymbol{A} \in \mathrm{I\!R}^{m\times n}$ with $m \geq n$ and rank k, $k < n$. Let $\boldsymbol{b} \in \mathrm{I\!R}^n$ and a homogeneous linear equation system be defined as $\boldsymbol{A}\boldsymbol{b} = \mathbf{0}_m$, where $\mathbf{0}_m = (\underbrace{0, \ldots, 0}_{m})^{\mathrm{T}} \in \mathrm{I\!R}^m$. The SVD of $\boldsymbol{A}$ leads to:

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{b} = \mathbf{0}_m\,. \tag{B.3}$$

Multiplication by $\boldsymbol{U}^{\mathrm{T}}$ from the left results in:

$$\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{b} = \mathbf{0}_n\,. \tag{B.4}$$

Equation (B.4) can only be valid if $\boldsymbol{b}$ is orthogonal to the first k row vectors of $\boldsymbol{V}^{\mathrm{T}}$. Since $\boldsymbol{V}^{\mathrm{T}}$ is orthonormal it spans the whole $\mathrm{I\!R}^{n\times n}$. Therefore, $\boldsymbol{b}$ must lie in the space spanned by the rows $k+1, \ldots, n$ of $\boldsymbol{V}^{\mathrm{T}}$, i.e., $\boldsymbol{b} = \sum_{i=k+1}^{n} \lambda_i \boldsymbol{v}_i$, where $\boldsymbol{v}_i$ is the i-th row of $\boldsymbol{V}^{\mathrm{T}}$ and $\lambda_i \in \mathrm{I\!R} \setminus 0$. These row vectors build an orthonormal basis of the null-space of $\boldsymbol{A}$ and define the solution of the homogeneous linear equation system $\boldsymbol{A}\boldsymbol{b} = \mathbf{0}_m$.

The inverse of a non-singular square matrix: Let $\boldsymbol{X} \in \mathrm{I\!R}^{n\times n}$ with rank $k = n$. Then the inverse of $\boldsymbol{X}$ can be obtained easily:

$$\boldsymbol{X}^{-1} = \left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right)^{-1} = \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{V}\mathrm{diag}\left(s_1^{-1}, s_2^{-1}, \ldots, s_n^{-1}\right)\boldsymbol{U}^{\mathrm{T}}\,. \tag{B.5}$$

Note that depending on the properties of $\boldsymbol{X}$, more efficient solutions to compute $\boldsymbol{X}^{-1}$ exist, e. g., for symmetric and positive definite matrices the inversion based on a Cholesky factorization is more efficient [And95].

The pseudo-inverse of a matrix: For $\boldsymbol{X} \in \mathrm{I\!R}^{m \times n}, m > n$, the pseudo-inverse $\boldsymbol{X}^{+}$ is defined by

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \,. \tag{B.6}$$

The inverse of the square matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is computed by equation (B.5). A problem occurs for singular values being zero or close to zero. In this case, the elements of $\boldsymbol{S}^{-1}$ are also set to zero [Tre97]. The singular value decomposition provides a numerically and computationally optimal solution in the sense of the Frobenius norm.

Orthogonalization of a matrix: For each orthonormal matrix $\boldsymbol{X} \in \mathrm{I\!R}^{m \times m}$ the following equation is valid:

$$\boldsymbol{I}_{m \times m} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{S}^{2}\boldsymbol{U}^{\mathrm{T}} \Leftrightarrow \boldsymbol{S} = \boldsymbol{I}_{m \times m} \tag{B.7}$$

This means that the singular values of an orthonormal matrix are all 1. An orthogonal approximation to an arbitrary matrix can therefore be computed by applying a singular value decomposition, setting all singular values to 1 and re-composing the matrix. This matrix will be the best approximation in the sense of the Frobenius norm.

Appendix C

Mathematical Symbols

In this appendix all mathematical symbols and notations used in this work are explained. First some general notations are given which are applicable to different symbols.

- Scalar values are denoted by italic letters like a, b, c.
- Vectors are denoted by bold italic letters like $\boldsymbol{x}$
- The i-th element of a vector $\boldsymbol{x}$ is denoted by x_i.
- Matrices are denoted by capital bold italic letters $\boldsymbol{X}$.
- The element at the i-th row and j-th column of a matrix $\boldsymbol{X}$ is denoted by x_{ij}.
- The transposed of a vector $\boldsymbol{x}$ and a matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{x}^{\mathrm{T}}$ and $\boldsymbol{X}^{\mathrm{T}}$.
- The inverse of a matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{X}^{-1}$.
- The Euclidean norm of a vector is denoted by $\|\boldsymbol{x}\|$.
- The Frobenius norm of a matrix is also denoted by $\|\boldsymbol{X}\|$.
- A homogeneous vector is denoted by underlining like $\underline{\boldsymbol{x}}$.
- An estimation to a value x is denoted by $\widehat{x}$.
- A matrix $\boldsymbol{Y}$ composed of a matrix $\boldsymbol{X}$ and vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$ is denoted by $\boldsymbol{Y} = [\boldsymbol{X}, \boldsymbol{x}_1, \boldsymbol{x}_2]$.

The following table lists the used symbols, their meaning, and the page of their first occurrence.

ψ	Plenoptic function	33
λ	Wavelength	33
τ	Continuous time value	33
$\boldsymbol{p}$	3-D point (lying on a light ray)	33
I	Intensity	33
$\boldsymbol{n}$	Viewing direction	33
$\boldsymbol{O}$	Origin	34
I_{r}	Discrete intensity for color red	34
I_{g}	Discrete intensity for color green	34
I_{b}	Discrete intensity for color blue	34
$\boldsymbol{\psi_6}$	Plenoptic function for static light source: 6-D parameter space	34
$\boldsymbol{\psi_5}$	Plenoptic function for static light source and scene: 5-D parameter space	35
$\boldsymbol{R}$	3×3 rotation matrix	36
$\boldsymbol{r}_{\mathrm{x}}$	First column of rotation matrix	36
$\boldsymbol{r}_{\mathrm{y}}$	Second column of rotation matrix	36
$\boldsymbol{r}_{\mathrm{z}}$	Third column of rotation matrix	36
$\boldsymbol{t}$	3-D translation vector	36
F	Focal length	36
F_{x}	Effective horizontal focal length	36
F_{y}	Effective vertical focal length	36
C_{x}	Horizontal coordinate of principal point	36
C_{y}	Vertical coordinate of principal point	36
dx	Size of a pixel on the sensor chip in x-direction	36

$\boldsymbol{w}$	3-D scene point	36
${}^{\mathrm{c}}\boldsymbol{w}$	3-D scene point, given in camera coordinates	36
$\boldsymbol{q}$	2-D image point vector	36
$\boldsymbol{K}$	Calibration matrix	37
$\underline{\boldsymbol{q}}$	Homogeneous image point vector	37
$\boldsymbol{P}$	Perspective projection matrix	37
$\underline{\boldsymbol{w}}$	Homogeneous scene point vector	37
$\boldsymbol{f}$	Discrete color image with three color channels: I_{r}, I_{g}, and I_{b}	38
d	Depth map	38
d_t	Depth map, storing distance to camera centre	38
d_z	Depth map, storing distance of projection to viewing direction	38
c	Confidence map	38
$\boldsymbol{\psi_{5,dc}}$	Extended plenoptic function (depth and confidence information)	38
$\boldsymbol{\psi_{4}}$	Plenoptic function with 4-D parameter space	40
N_{uv}	Number of subdivision of the uv-plane	40
N_{st}	Number of subdivision of the st-plane	40
B	Quadralinear basis function	40
$\boldsymbol{\psi}_{4,d}$	Discrete plenoptic function with 4-D parameter space	40
$\boldsymbol{\psi}_{4,r}$	Reconstructed plenoptic function	40
η	Image frequency	42
$\boldsymbol{H}$	Image homography	44
N_{v}	Number of contributing views	46
π	Penalty for unstructured lumigraph rendering	47
$\boldsymbol{G}$	Structure matrix	50
f_x	1. derivative of a gray-value image in x-direction	50
f_y	1. derivative of a gray-value image in y-direction	50

f	Discrete gray-value image	50
ϵ_{BPE}	Back-projection error	51
N_{w}	Number of world points	51
N_{f}	Number of (captured) frames/images	51
$\boldsymbol{\Gamma}$	Factorisation: measurement matrix	51
$\boldsymbol{\Psi}$	Factorisation: camera matrix	51
$\boldsymbol{\Phi}$	Factorisation: scene matrix	51
σ	Standard deviation	56
$\boldsymbol{\mu}$	Mean value (of a number of pixels)	56
$\mathcal{P}$	Set of pixel positions	57
$\sigma_{\mathcal{P}}$	Mean standard deviation for the set M	57
x_{ds}	Distorted x-coordinate in sensor coordinates	59
y_{ds}	Distorted y-coordinate in sensor coordinates	59
δ_x	Distortion term for x-coordinate	59
δ_y	Distortion term for y-coordinate	59
x_{us}	Undistorted x-coordinate in sensor coordinates	59
y_{us}	Undistorted y-coordinate in sensor coordinates	59
$\tilde{\delta}_x$	Distortion correction term for x-coordinate	59
$\tilde{\delta}_y$	Distortion correction term for y-coordinate	59
r	Distance of distorted point from center of sensor	59
x_{d}	Distorted x-coordinate of a pixel	60
y_{d}	Distorted y-coordinate of a pixel	60
x_{u}	Undistorted x-coordinate of a pixel	60
y_{u}	Undistorted y-coordinate of a pixel	60
dy	Size of a pixel on the sensor chip in y-direction	60
C	Circularity criterion	62

$\underline{\boldsymbol{w}}'$	Homogeneous scene point vector with w_z=0	63
$\overline{\boldsymbol{h}}$	Image homography, row vector	63
$\boldsymbol{U}$	Left matrix of singular value decomposition	64
$\boldsymbol{S}$	Middle matrix of singular value decomposition	64
$\boldsymbol{V}$	Right matrix of singular value decomposition	64
$\widetilde{\boldsymbol{K}}$	Calibration matrix of Zhang	65
$\widetilde{\boldsymbol{R}}$	Zhang's 3×3 rotation matrix	65
$\widetilde{\boldsymbol{t}}$	Zhang's 3-D translation vector	65
$\widetilde{\boldsymbol{r}}_{\mathrm{x}}$	First column of Zhang's rotation matrix	65
$\widetilde{\boldsymbol{r}}_{\mathrm{y}}$	Second column of Zhang's rotation matrix	65
$\widetilde{\boldsymbol{r}}_{\mathrm{z}}$	Third column of Zhang's rotation matrix	65
ϕ	Rotation angle	69
$\boldsymbol{r}$	Rotation axis	69
$\boldsymbol{R}_{\mathrm{C}}$	3×3 rotation matrix	69
N_{r}	Number of rows of an image	69
N_{c}	Number of columns of an image	69
$\boldsymbol{C}$	Color covariance matrix	69
$\boldsymbol{I}_{3\times3}$	The 3×3 identity matrix	70
Y	Reflecting light	79
L_{S}	Specular reflection	79
L_{D}	Diffuse reflection	79
$\boldsymbol{\theta}$	Capture properties	79
c_{S}	Weighting factor for specular reflection	79
c_{D}	Weighting factor for diffuse reflection	79
P_{S}	Specular reflection property	79
$E(\lambda)$	Spectrum of the light source	79

P_{D}	Diffuse reflection property	79
$\overline{\epsilon}_{\mathrm{BPE}}$	Mean back-projection error	86
$\boldsymbol{g}$	Warp function	91
$\boldsymbol{a}$	Affine warping parameters	91
0_3	The 3-D column null–vector	94
$\boldsymbol{T}$	4×4 transformation matrix	94
${}^{\mathrm{c}}\underline{\boldsymbol{w}}$	Homogeneous scene point vector given in camera coordinates	94
$\boldsymbol{T}_{\mathrm{B,H}}$	Transformation from base to hand	95
$\boldsymbol{T}_{\mathrm{H,E}}$	Transformation from hand to endoscope/eye	95
$\boldsymbol{R}_{\mathrm{B,H}}$	Rotation from base to hand	96
$\boldsymbol{t}_{\mathrm{B,H}}$	Translation from base to hand	96
l_{e}	AESOP kinematics: endoscope length	97
α_{plug}	AESOP kinematics: plug angle	97
α_{opt}	AESOP kinematics: side view optics angle	97
α_{c2o}	AESOP kinematics: camera head angle	97
$\boldsymbol{R}_{\mathrm{H,E}}$	Rotation from hand to endoscope/eye	98
$\boldsymbol{t}_{\mathrm{H,E}}$	Translation from hand to endoscope/eye	98
$\boldsymbol{T}_{\mathrm{H}}$	Endoscope plug (hand) pose	107
$\boldsymbol{R}_{\mathrm{H}}$	Endoscope plug (hand) rotation	107
$\boldsymbol{t}_{\mathrm{H}}$	Endoscope plug (hand) translation	107
$\boldsymbol{T}_{\mathrm{E}}$	Camera (eye) pose	107
$\boldsymbol{R}_{\mathrm{E}}$	Camera (eye) rotation	107
$\boldsymbol{t}_{\mathrm{E}}$	Camera (eye) translation	107
$\mathcal{M}$	Set of relative movements	108
θ_{angle}	Rotation angle threshold	109
$\mathcal{S}$	Set of movement pairs	109

N_{relmov}	Number of elements in the set of relative movements	109
$\boldsymbol{c}$	Cluster center	110
p_{out}	Assumed probability of LMedS outliers	113
θ_{LMedS}	Threshold for LMedS	113
θ_{BPE}	Threshold for back-projection error	113
$\boldsymbol{R}_{\text{reg}}$	Registration: 3×3 rotation matrix	126
$\boldsymbol{t}_{\text{reg}}$	Registration: 3-D translation vector	126
$\boldsymbol{x}_{\text{ct}}$	3-D point in CT coordinates	126
$\boldsymbol{x}_{\text{endo}}$	3-D point in endoscope coordinates	126
N_{iter}	Number of triangular mesh smoothing iterations	129
ρ	Triangular mesh smoothing coefficient	129
$\boldsymbol{x}_{\text{ct,reg}}$	Registered 3-D point of CT dataset	129
$\boldsymbol{R}_{\text{c}}$	Coarse registration: 3×3 rotation matrix	129
$\boldsymbol{t}_{\text{c}}$	Coarse registration: 3-D translation vector	129
$\widetilde{\boldsymbol{x}}_{\text{ct,reg}}$	Coarse registered 3-D point of CT dataset	129
$\boldsymbol{R}_{\text{f}}$	Fine registration: 3×3 rotation matrix	129
$\boldsymbol{t}_{\text{f}}$	Fine registration: 3-D translation vector	129
$\mathcal{C}$	Registration: set of 3-D points in CT coordinates	130
$\mathcal{E}$	Registration: set of 3-D points in endoscope coordinates	130
$\mathcal{C}_{\text{s}}$	Registration: selected set of 3-D points in CT coordinates	130
$\mathcal{E}_{\text{s}}$	Registration: selected set of 3-D points in endoscope coordinates	130
N_{corr}	Registration: number of correspondences for coarse registration	130
$\overline{\boldsymbol{x}}_{\text{ct}}$	Mean value of points in CT coordinates	130
$\overline{\boldsymbol{x}}_{\text{endo}}$	Mean value of points in endoscope coordinates	130
$\boldsymbol{A}$	Registration: correlation matrix	130
ϵ_{MSRE}	Mean square registration error (MSRE)	132

v	Evaluation value	140
α	Significance level for statistical tests	141
$\epsilon_{\boldsymbol{R}}$	Rotation error	143
$\epsilon_{\boldsymbol{t}}$	Translation error	143
$\epsilon_{\boldsymbol{t},\mathrm{rel}}$	Relative translation error	143
$\boldsymbol{R}_{\mathrm{diff}}$	Rotation matrix for the rotation between two rotation matrices	144
$\boldsymbol{r}_{\mathrm{diff}}$	Rotation axis of $\boldsymbol{R}_{\mathrm{diff}}$	144
ϕ_{diff}	Rotation angle of $\boldsymbol{R}_{\mathrm{diff}}$	144
α_{diff}	Cardan angle α of $\boldsymbol{R}_{\mathrm{diff}}$	144
β_{diff}	Cardan angle β of $\boldsymbol{R}_{\mathrm{diff}}$	144
γ_{diff}	Cardan angle γ of $\boldsymbol{R}_{\mathrm{diff}}$	144
$\epsilon_{\boldsymbol{R},\mathrm{rel}}$	Relative rotation error	144
$\epsilon_{\boldsymbol{R},\mathrm{C},\alpha}$	Rotation error using Cardan angles, x-axis	144
$\epsilon_{\boldsymbol{R},\mathrm{C},\beta}$	Rotation error using Cardan angles, y-axis	144
$\epsilon_{\boldsymbol{R},\mathrm{C},\gamma}$	Rotation error using Cardan angles, z-axis	144
$\epsilon_{\boldsymbol{R},\mathrm{rel},\mathrm{C}}$	Relative rotation error using Cardan angles	144
$\mathit{\Delta}_f$	Minimal frame distance	145
$\bar{\epsilon}_{\boldsymbol{t}}$	Mean translation error	145
$\bar{\epsilon}_{\boldsymbol{t},\mathrm{rel}}$	Mean relative translation error	145
$\bar{\epsilon}_{\boldsymbol{R}}$	Mean rotation error	145
$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel}}$	Mean relative rotation error	145
$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\alpha}$	Mean rotation error using Cardan angles, x-axis	145
$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\beta}$	Mean rotation error using Cardan angles, y-axis	145
$\bar{\epsilon}_{\boldsymbol{R},\mathrm{C},\gamma}$	Mean rotation error using Cardan angles, z-axis	145
$\bar{\epsilon}_{\boldsymbol{R},\mathrm{rel},\mathrm{C}}$	Mean relative rotation error using Cardan angles	145
Q_{MAD}	Mean absolute difference (MAD)	145

Q_{SNR}	Signal to noise ratio (SNR)	145
Q_{PSNR}	Peak signal to noise ratio (PSNR)	145
$\overline{Q}_{\mathrm{MAD}}$	Mean mean absolute difference)	146
$\overline{Q}_{\mathrm{SNR}}$	Mean signal to noise ratio	146
$\overline{Q}_{\mathrm{PSNR}}$	Mean peak signal to noise ratio	146
$\overline{v}_{\mathrm{All}}$	Mean evaluation value for all surgeons	157
$\overline{v}_{\mathrm{Exp}}$	Mean evaluation value for experienced surgeons	157
$\overline{v}_{\mathrm{Unexp}}$	Mean evaluation value for unexperienced surgeons	157
$\epsilon_{\mathrm{shape}}$	Shape error	169

Anhang D

German Title, Contents, Introduction, and Summary

Der deutsche Titel dieser Dissertation lautet:

Erweiterte Lichtfeld-Visualisierung und Echtzeit-Bildverbesserung für computerassistierte endoskopische Operationen

D.1 Inhaltsverzeichnis

1 Einleitung 1
- 1.1 Problembeschreibung und medizinische Relevanz 2
- 1.2 Beitrag dieser Arbeit 4
- 1.3 Bildmodalitäten 6
- 1.4 Datenfusion 8
- 1.5 Übersicht 10

2 Stand der Technik bei computerunterstützten endoskopischen Operationen 11
- 2.1 Herkömmliche Chirurgie 12
- 2.2 Minimal-invasive Chirurgie 14
- 2.3 Roboterassisierte Eingriffe 17
- 2.4 Bildverbesserung 20
- 2.5 Augmented Reality (AR) 24
  - 2.5.1 3-D-Visualisierung 25

2.5.2 Lagebestimmung . . . 27
2.5.3 Virtuelle Daten . . . 29
2.5.4 Registrierung und Fusion . . . 29
2.5.5 Vergleich medizinischer AR und 3-D-Visualisierungsverfahren . . . 30

3 Lichtfeld-Theorie 33
3.1 Definition und Konzept . . . 33
3.1.1 Samples der plenoptischen Funktion . . . 36
3.1.2 Tiefenkarten und Vertrauenskarten . . . 38
3.2 Lichtfeld-Visualisierung . . . 39
3.2.1 Zwei-Ebenen-Lichtfelder . . . 39
3.2.2 Freiform-Lichtfelder . . . 43
3.2.3 Einbeziehen von Vertrauenskarten . . . 48
3.3 Lichtfeld-Rekonstruktion . . . 48
3.3.1 Mechanische Kalibrierung . . . 48
3.3.2 Struktur-aus-Bewegung . . . 50
3.4 Dynamische Lichtfelder . . . 53

4 Bildverbesserung 55
4.1 Entzerrung . . . 57
4.1.1 Verzerrungsmodell . . . 59
4.1.2 Kamerakalibrierung . . . 61
4.2 Farbnormierung . . . 68
4.3 Zeitliche Filterung . . . 71
4.4 Filterung im interessierenden Gebiet . . . 75
4.5 Bildgeometrie-Transformationen . . . 75
4.6 Bildverbesserung durch Lichtfelder . . . 77
4.6.1 Glanzlichtdetektion . . . 79
4.6.2 Glanzlichtsubstitution . . . 80
4.7 Zusammenfassung . . . 81

5 Lichtfelder des Operationsgebiets 83
5.1 Vorverarbeitung . . . 84
5.1.1 De-Interlacing . . . 84
5.1.2 Entzerrrung . . . 86

5.1.3 Ausschneiden 87
5.2 Intrinsische Parameter 87
5.3 Extrinsische Parameter aus Punktkorrespondenzen 90
5.4 Extrinsische Parameter mit Hilfe eines Roboterarms 94
5.4.1 AESOPs Kinematik 95
5.4.2 Hand-Auge-Kalibrierung von AESOP 97
5.5 Extrinsische Parameter mit Hilfe eines optischen Trackingsystems 101
5.5.1 Target Design 103
5.5.2 Hand-Auge-Kalibrierung von smARTtrack1 106
5.6 Berechnung von Tiefen- und Vertrauenskarten 111
5.7 Lichtfeld-Visualisierung 117
5.8 Vergleich 118
5.9 Zusammenfassung 119

6 Erweiterte Realität: Registrierung und Fusion von 3-D-Daten mit Lichtfeldern 121
6.1 Wichtige anatomische Strukturen 122
6.2 Anatomische Datenbank 125
6.3 Bestimmung der Transformationsparameter zur Registrierung 126
6.3.1 Auswahl von 3-D-Punkten im Endoskop-Koordinatensystem 126
6.3.2 Auswahl von 3-D-Punkten im CT-Daten-Koordinatensystem 127
6.3.3 Berechnung der Transformationsparameter zur Registrierung 129
6.4 Fusion durch multimodale Visualisierung 133
6.5 Zusammenfassung 135

7 Experimente und Auswertungen 137
7.1 Methodik zur Auswertung 140
7.1.1 Echtzeit-Bildverbesserung 140
7.1.2 Lagefehler 142
7.1.3 Lichtfeldqualität und Bildqualität 145
7.2 Endoskopische Bildverbesserung 147
7.2.1 Kamerakalibrierung und Entzerrung 149
7.2.2 Farbnormierung 151
7.2.3 Zeitliche Filterung 152
7.2.4 Bildgeometrie-Transformationen 156
7.2.5 Auswertung 156

7.3 Lichtfeld-Rekonstruktion und -Visualisierung 161
7.3.1 Statische Lichtfelder mit Struktur-aus-Bewegung 161
7.3.2 Statische Lichtfelder mit AESOP 163
7.3.3 Statische Lichtfelder mit smARTtrack1 166
7.3.4 Dynamische Lichtfelder mit smARTtrack1 188
7.4 Bildverbesserung durch Lichtfelder 189
7.5 Erweiterte Realität: Registrierung und Fusion mit 3-D-Daten 194
7.6 Diskussion 198

8 Zusammenfassung und Ausblick **201**
8.1 Zusammenfassung 201
8.2 Ausblick 204

A Homogenen Koordinaten **207**

B Singulärwertzerlegung **209**

C Mathematische Symbole **213**

D Titel, Inhaltsverzeichnis, Einleitung und Zusammenfassung auf Deutsch **223**
D.1 Inhaltsverzeichnis 223
D.2 Einleitung 227
D.2.1 Problembeschreibung und medizinische Relevanz 228
D.2.2 Beitrag dieser Arbeit 230
D.2.3 Bildmodalitäten 232
D.2.4 Datenfusion 235
D.2.5 Übersicht 237
D.3 Zusammenfassung 239

Abbildungsverzeichnis **243**

Tabellenverzeichnis **247**

Literaturverzeichnis **249**

Index **273**

Lebenslauf **277**

D.2 Einleitung

Die Zufriedenheit der Menschen hängt hauptsächlich von ihrer Gesundheit ab. Daher gibt die Gesellschaft eine Menge Geld für Forschung zur Entwicklung optimaler Behandlungstechniken aus. Diese Doktorarbeit befasst sich mit Krankheiten, bei denen ein chirurgischer Eingriff derzeit als optimale Behandlungsmethode gesehen wird. Beispiele für solche Krankheiten sind Entzündungen der Gallenblase und des Wurmfortsatzes, die starke Schmerzen nach sich ziehen. Die Entfernung des betroffenen Körperteils ist oft die einzig mögliche Behandlung. Die Forschung in diesem Bereich hat zum Ziel, den Patienten so gut wie möglich zu behandeln, bei gleichzeitiger Verringerung seines Traumas. Das Trauma kann über die Schwere der Verletzung durch die Behandlung, den intra- und postoperativen Schmerz, die kosmetische Beeinträchtigung und die nötige Rekonvaleszenzzeit definiert werden. Die Entwicklung im Bereich der Chirurgie verläuft zu so genannten *minimal-invasiven* Operationen hin, die den Patienten deutlich weniger als herkömmliche Operationen belasten.

Die Grundidee bei minimal-invasiver Chirurgie ist den Zugang zum Operationsgebiet durch kleine „Schlüssellöcher“ herzustellen, für die nur ein kleiner Schnitt mit einem Durchmesser von ungefähr 1 bis 2 cm benötigt wird. Operiert wird mit speziellen chirurgischen Instrumenten, wobei das Bild des Operationsgebiets mit Hilfe eines Endoskops und einer Kamera zur Verfügung gestellt wird. Licht wird durch das Endoskop eingebracht, welches zusammen mit der Kamera das Bild, das auf einem Videomonitor angezeigt wird, erzeugt. Drei Begriffe werden für diese Art von Operation synonym gebraucht: *minimal-invasive Chirurgie*, *Schlüsselloch-Chirurgie* und *endoskopische Operation*.

Im Vergleich zur herkömmlichen Chirurgie erfordert die Durchführung einer minimal-invasiven Operation zusätzliches Training und es müssen eine Menge Nachteile in Kauf genommen werden: ungewöhnliche Instrumente, kein direkter Tastsinn bzw. nur über die chirurgischen Instrumente, eingeschränkte Bewegungsfreiheit, eingeschränkte Sicht, Beeinträchtigung des Bildes durch Glanzlichter, Rauch oder kleine umherfliegende Partikel sowie der Verlust der stereoskopischen Tiefenwahrnehmung auf Grund der Anzeige des Endoskopbilds auf einem Videomonitor. Das verringerte Trauma des Patienten rechtfertigt jedoch diesen menschlichen und technischen Aufwand. Immer mehr minimal-invasive Operationen ersetzen herkömmliche Operationen als „Gold Standard“, sie sind also die Behandlungsmethode, die derzeit als optimal betrachtet wird. Beispiele hierfür sind Cholecystektomie (Entfernung der Gallenblase), Appendektomie (Entfernung des Appendix), Leisten- und Zwerchfellhernie, gastro-esophagale Refluxkrankheit (engl. Abkürzung: GERD) und Darmchirurgie (Resektion des Darms im Falle von Entzündungen oder malignen Erkrankungen).

In dieser Doktorarbeit werden Verfahren zur Unterstützung des Chirurgen bei endoskopischen Operationen durch Methoden aus dem Bereich des Rechnersehens untersucht. In den folgenden Kapiteln wird eine detaillierte Beschreibung der Probleme, die bei minimal-invasiven Operationen auftreten (Kapitel D.2.1), sowie der Beitrag dieser Arbeit, um diese Probleme zu reduzieren (Kapitel D.2.2), gegeben. Die Beiträge werden außerdem in Beziehung gesetzt zu herkömmlichen bildgebenden Verfahren (Kapitel D.2.3) und Datenfusion (Kapitel D.2.4). Abschließend gibt Kapitel D.2.5 einen Überblick über die Gliederung der Arbeit.

D.2.1 Problembeschreibung und medizinische Relevanz

Im Vergleich zur herkömmlichen Chirurgie treten bei endoskopischen Operationen viele herausfordernde Probleme auf:

Beeinträchtigung des Bildes: Das Glasfaserbündel des Lichtleiters des Endoskops endet direkt neben der distalen Linse. Dies führt zu Glanzlichtern auf orthogonal zur Blickrichtung verlaufenden Gewebeoberflächen, insbesondere wenn das Gewebe feucht ist. Die Menge an Licht, die durch das Endoskop in Körperhöhlen eingebracht werden kann, ist beschränkt. Zu viel Licht würde zu viel Hitze erzeugen und dadurch das Gewebe nahe der Endoskopspitze verbrennen. Unter diesen Voraussetzungen können Körperhöhlen, vor allem große wie beispielsweise das Abdomen, nur inhomogen und in manchen Bereichen mit geringem Kontrast ausgeleuchtet werden. Außerdem kann es vorkommen, dass naheliegende Gewebeoberflächen auf Grund der Menge an Licht, die nötig ist um den hinteren Teil der Körperhöhle zu erleuchten, überbelichtet werden. Das Schneiden von Gewebe mit Hochfrequenz- oder Ultraschallschneidern führt zu Rauch, kleinen umherfliegenden Partikeln und einer Rotfärbung auf Grund von Einblutungen. Abschließend haben Endoskoplinsen eine sehr kleine Brennweite, beispielsweise 7 mm für ein $1/2\,''$ CCD Chip mit PAL Auflösung (768×576 Pixel), und herstellungsbedingt treten vor allem bei Linsen mit kleinen Brennweiten Verzerrungen des Bildes auf, insbesondere an den Rändern.

Eingeschränkte Sicht: Das Problem der eingeschränkten Sicht kann nachvollzogen werden, wenn man sich die Aufgabe vor Augen führt, einen klaren Eindruck eines Raumes zu erlangen, den man nur mit einer Kamera betrachten kann. Man würde wahrscheinlich die kleinste zur Verfügung stehende Brennweite wählen, in die Mitte des Raumes gehen und versuchen, sich umzusehen. Man stelle sich nun die gleiche Aufgabe vor, allerdings mit einer Kamera, die an einem langen Stab befestigt ist und die von Außen durch das Schlüsselloch der Tür bewegt werden muss. Falls die Brennweite der Kamera fest vorgegeben ist, was im Fall von endoskopischen Operationen üblich ist, muss die Kamera nahe an Objekte herangeführt werden, um diese detailliert zu betrachten. Dann ist jedoch nur ein sehr kleiner Teil des Raumes sichtbar und das Zurechtfinden im Raum wird schwierig.

Verlust der stereoskopischen Tiefenwahrnehmung: Die menschliche Tiefenwahrnehmung basiert hauptsächlich auf dem Vorhandensein und der Nutzung zweier Augen. Die Tiefeninformation eines Objekts wird aus den beiden Bildern entsprechend dessen Position im linken und rechten Bild extrahiert. Je größer der Unterschied der beiden Positionen, desto näher ist das Objekt. Beim Betrachten von projizierten Bildern, beispielsweise auf Fotos, im Fernsehen oder auf Computer- und Videomonitoren, ist diese Art von Tiefenwahrnehmung nicht möglich, da man ein flaches Bild betrachtet. Andere Hinweise, die mit der Tiefe eines Objekts korrelieren, werden dann benutzt: Verdeckung, Beleuchtung und Information über die Größe bei nicht bewegten Bildern sowie Geschwindigkeit der Objektbewegung im Verhältnis zu seiner Größe bei bewegten Bildern wie beispielsweise im Fernsehen. Die Wahrnehmung ist jedoch nicht die gleiche wie bei *echtem* stereoskopischem Sehen. Bei minimal-invasiven Operationen wird dieser Unterschied wichtig. Die einfache Aufgabe, ein Objekt mit einem Endo-Greifer zu greifen, veranschaulicht den Unterschied: während diese Aufgabe mit normalem Sehen einfach ist, wird sie extrem schwer, sobald ein Videobild verwendet werden muss.

Ungewöhnliche chirurgische Instrumente: Alle verwendeten chirurgischen Instrumente unterscheiden sich von den herkömmlichen, beispielsweise sind sie länger und schmaler, sodass deren Verwendung eingeübt werden muss.

Eingeschränkte Bewegungsfreiheit: Das Trauma des Patienten wird hauptsächlich durch den Zugang zum Operationsgebiet mittels kleinen „Schlüssellöchern“ reduziert. Dies führt zu einer Einschränkung der möglichen Bewegungen: jedes Instrument und das Endoskop muss durch solch ein „Schlüsselloch“ eingeführt und bewegt werden.

Eingeschränkter Tastsinn: Während einer herkömmlichen Operation kann der Chirurg den Operationssitus zusätzlich mit seinem Tastsinn untersuchen. Die bereits beschriebenen minimal-invasiven Techniken erlauben diese Palpation nicht. Es ist nur eine sehr eingeschränkte Tastwahrnehmung über die chirurgischen Instrumente möglich: die Elastizität von Gewebe kann untersucht werden, indem mit einem Instrument hineingedrückt wird.

Schwierige Hand-Auge-Koordination: Unter Hand-Auge-Koordination verseht man den Akt der Bewegung der Hand (die ein chirurgisches Instrument hält) an eine bestimmte Position.

> Im Fall von minimal-invasiven Operationen ist diese Aufgabe aus folgenden Gründen sehr schwer: außer dem Verlust der stereoskopischen Tiefenwahrnehmung kann es sein, dass die Blickrichtung des Endoskops nicht mit derjenigen des Chirurgen übereinstimmt; außerdem muss die Bewegung mit einem langen Instrument durch ein „Schlüsselloch“ durchgeführt werden. Beispielsweise bewegt sich die Instrumentspitze nach links wenn der Chirurg das Instrument nach rechts bewegt, da die Bewegung durch das Schlüsselloch hindurch erfolgt.

Im Hinblick auf das dargelegte Ziel der Entwicklung von optimalen Behandlungstechniken ist es wichtig, die erwähnten Probleme soweit wie möglich zu reduzieren. Dadurch verbessern sich die Bedingungen für den Chirurgen, beispielsweise, indem die Bildqualität oder die Übersichtlichkeit erhöht wird, was künftig eine verringerte Belastung für den Chirurgen nach sich zieht. Daraus resultiert eine verbesserte Durchführung der Operation sowie eine verringerte Operationszeit. Insgesamt führt dies zu einer Verringerung des Patiententraumas und der Rekonvaleszenzzeit.

In dieser Doktorarbeit werden Methoden zur Behebung oder Verringerung folgender Probleme vorgestellt: *Beeinträchtigung der Bilder*, *Eingeschränkte Sicht* und *Verlust der stereoskopischen Tiefenwahrnehmung*. Das folgende Kapitel beschreibt die Beiträge dieser Arbeit im Detail.

D.2.2 Beitrag dieser Arbeit

Um die Probleme bei endoskopischen Operationen zu reduzieren, wurde ein neues System entwickelt, das Bildverbesserung in Echtzeit, 3-D-Visualisierung des Operationsgebiets und Augmented Reality, d. h. Registrierung und Fusion mit CT/MRT Daten, zur Verfügung stellt. Es erlaubt das Entfernen bzw. die Reduktion unterschiedlicher Bildstörungen, die Rekonstruktion eines 3-D-Modells des Operationsgebiets (ein so genanntes *Lichtfeld*), welches dreidimensional aus beliebigen Positionen betrachtet werden kann, sowie die Erweiterung sowohl des 2-D-Livebildes als auch des Lichtfelds mit CT/MRT Daten, nachdem eine Registrierung basierend auf der rekonstruierten 3-D-Information durchgeführt wurde. Im Folgenden werden die Beiträge detaillierter beschrieben.

Beeinträchtigungen des Bildes Außer der vor kurzem in [Fis04] vorgestellten Arbeit wurden bisher lediglich Lösungen für einzelne Bildstörungen veröffentlicht, siehe z. B. [Grö01, Hel01, Mün04], doch die meisten Ansätze wurden nicht für den Einsatz im Operationssaal entwickelt. Daher wird zunächst ein System zusammengestellt, das das Verarbeiten und die Anzeige endoskopischer Bilder erlaubt. Die Hauptkomponente ist ein typisches Videoendoskopiesystem. Zur

Bildverbesserung in Echtzeit wird das System durch einen PC mit einer S-VHS Framegrabber Karte und einen zweiten Monitor erweitert. Dieses System erlaubt es, das Bild der Endoskopkamera aufzunehmen, es zu verarbeiten und anschließend auf dem zweiten Monitor anzuzeigen. Optimierte Algorithmen ermöglichen die Bildverarbeitung in Echtzeit. Verzerrungen des Bildes werden korrigiert, wobei die durch Kalibrierung berechneten intrinsischen Kameraparameter des Endoskops verwendet werden. Ein Farbnormierungsalgorithmus, der ursprünglich zur Verbesserung von Objektlokalisation und -klassifikation verwendet wurde, der *color cluster rotation* Algorithmus [Pau98], wird verwendet, um beleuchtungsunabhängige Bilder anzuzeigen, die es erlauben, verschiedene Gewebetypen auch in schwierigen Situation zu unterscheiden. Kleine umherfliegende Partikel und Rauch stören den Chirurgen während des Schneidens von Gewebe. Diese Beeinträchtigungen werden durch zeitliche Farbmedianfilterung reduziert. Eine Methode wird vorgestellt, die es erlaubt, schnelle *räumliche* Filter zur *zeitlichen* Filterung zu verwenden. Falls die Lage des Endoskops, d. h. dessen Position und Orientierung, bekannt ist, kann das Bild anhand eines vordefinierten Horizonts gedreht werden, wodurch der Horizont für fast beliebige Bewegungen des Endoskops konstant gehalten werden kann. Bisher wurde keine Evaluierung von Methoden zur Verbesserung endoskopischer Bilder veröffentlicht. Daher werden alle hier entwickelten Methoden durch Ärzte evaluiert.

Verlust der stereoskopischen Tiefenwahrnehmung und eingeschränkte Sicht Die vorgeschlagene Lösung für beide Probleme ist die Rekonstruktion eines Lichtfelds [Lev96, Gor96] des Operationsgebiets. Lichtfelder sind eine relativ neue bildbasierte Methode zur Modellierung und Visualisierung von 3-D-Szenen. Obwohl andere Techniken zur 3-D-Rekonstruktion des Operationsgebiets verwendet wurden [Tho02, Küb02, Dey02, Dev01], sind Lichtfelder bisher nicht untersucht worden. Die Hauptprobleme bei der Rekonstruktion von Lichtfeldern aus endoskopischen Bildern sind die Bestimmung der Lage des Endoskops und die Berechnung von 3-D-Szenengeometrie. Die Visualisierung von Lichtfeldern wäre auch ohne die Kenntnis der Szenengeometrie möglich, allerdings mit schlechter Qualität. Daher beinhalten die Lichtfelder, die in dieser Arbeit rekonstruiert werden, immer Szenengeometrie. Drei verschiedene Möglichkeiten zur Lichtfeldrekonstruktion werden vorgestellt: lediglich basierend auf den Eingabebildern, unter Verwendung eines Roboterarms, der das Endoskop bewegt und die Information über die Lage zur Verfügung stellt, sowie mit Hilfe eines optischen Trackingsystems zur Lagebestimmung des Endoskops. Jede Methode hat Vor- und Nachteile. Die Visualisierung des Operationsgebiets durch Lichtfelder erlaubt es, das Operationsgebiet dreidimensional zu betrachten, beispielsweise auf einem 3-D-Monitor oder einem Head-Mounted-Display (HMD). Die Betrachtungsposition ist dabei nicht auf die ursprünglichen Endoskoppositionen beschränkt, beispielsweise kann, falls ein Teil einer Szene aufgenommen wurde, indem das Endoskop sehr nahe herangeführt wurde, die Übersicht erhöht werden, indem die Brennweite virtuell verringert und das Endoskop rückwärts bewegt wird. Dies ist vor allem nützlich, um mit dem Problem der eingeschränkten Sicht zurechtzukommen. Da die 3-D-Szene im Computer repräsentiert wird, sind alle Bewegungen virtuell und machen keine Bewegung des echten Endoskops erforderlich. Abschließend wird eine Methode zur Substitution beliebiger Bildstörungen mit Hilfe von Lichtfeldern vorgestellt. Drei Voraussetzungen müssen erfüllt sein, um diese Methode anzuwenden: ein Lichtfeld der Szene muss vorhanden sein, die Störung darf nicht an der gleichen Position relativ zu der Szene bleiben während sich das Endoskop bewegt und die Störung muss im Bild detektierbar sein.

Augmented Reality Die in einem Lichtfeld enthaltene Information, namentlich die 3-D-Szenengeometrie, ermöglicht es, Augmented Reality während endoskopischen Operationen zur Verfügung zu stellen. Zu diesem Zwecke wird das Lichtfeld mit anderen 3-D-Daten wie beispielsweise MRT oder CT unter Verwendung anatomischer Landmarken, die in der Szene identifiziert werden können, registriert. Danach können CT/MRT Daten in die Lichtfeldvisualisierung eingeblendet werden, was es erlaubt hinter die Oberfläche, durch Organe und Gewebe hindurch zu „sehen“. Auf Grund der Szenengeometrie können anatomische Landmarken benutzt werden und Marker sind nicht erforderlich. Der Fokus liegt in dieser Arbeit auf der Berechnung der für die Rekonstruktion eines Lichtfeldes zusammen mit 3-D-Szenengeometrie nötigen Information und auf der Registrierung mit anderen 3-D-Daten. Rendering Techniken für das Lichtfeld sowie für Augmented Reality, d. h. die Überlagerung des Lichtfelds mit CT/MRT Daten, werden nicht untersucht. Bis jetzt existieren nur Augmented Reality Systeme, bei denen Marker zur Registrierung verwendet werden, siehe beispielsweise [Sch03a]: hier wird das registrierte CT-Bild in das 2-D-Monitorbild eingeblendet. Natürlich ist diese zweidimensionale Art der Augmented Reality auch möglich, nachdem das Lichtfeld mit den 3-D-Daten registriert wurde.

Nachdem die Beiträge dieser Arbeit beschrieben wurden, setzen die nächsten beiden Kapitel sie in Beziehung zu herkömmlichen bildgebenden Verfahren und zur Datenfusion.

D.2.3 Bildmodalitäten

In der modernen Medizin ist der Einsatz von bildgebenden Verfahren zur Erlangung von Informationen über das Innere des Patienten weit verbreitet. Im Vorfeld einer Operation wird beispielsweise fast immer Bildmaterial aquiriert, um detaillierte Informationen über die Anatomie und die Krankheit des Patienten zu erhalten. Falls eine exakte Diagnose der Krankheit nicht möglich ist, können vor allem bildgebende Verfahren zur Klärung der Diagnose beitragen. Außer Bildern, die dadurch entstehen, dass man mit Hilfe eines Endoskops *direkt* in den Patienten blickt, ermöglichte die Entdeckung der Röntgenstrahlung, Bilder der menschlichen Anatomie zu erzeugen ohne den Patienten zu verletzen bzw. lediglich mit der nur potentiell gefährlichen Röntgenstrahlung zu belasten. Wilhelm Röntgen entdeckte die Röntgenstrahlen im Jahre 1895. Die ersten Röntgengeräte waren kurze Zeit später verfügbar und wurden auf der ganzen Welt zu medizinischen Zwecken eingesetzt. Heutzutage gibt es eine große Zahl an etablierten bildgebenden Verfahren, z. B. Computertomographie (CT), Magnetresonanztomographie (MRT), funktionelle MRT (fMRT), Angiographie, Digitale Subtraktionsangiographie (DSA), Positronen-Emissions-Tomographie (PET), Single-Photon-Emissionscomputertomographie (SPECT) oder 3-D-Ultraschall. Die genannten bildgebenden Verfahren werden nach wie vor weiterentwickelt. In den folgenden Abschnitten werden CT, MRT und PET genauer erklärt.

Computertomographie basiert auf Röntgenstrahlung. Ein herkömmliches Röntgengerät besteht aus einer Röntgenquelle, die ein Bündel Röntgenstrahlen aussendet, und einem Röntgenfilm. Die Röntgenstrahlen durchdringen den Patienten. Unterschiedliche Gewebetypen absorbieren unterschiedlich viel Röntgenstrahlung. Der Röntgenfilm befindet sich gegenüber der Röntgenquelle, der Patient in der Mitte. Der „Schatten“ der ausgesendeten Röntgenstrahlung wird auf dem Film aufgezeichnet, d. h. je dichter eine anatomische Struktur, desto mehr werden die Röntgenstrahlen abgeschwächt und desto heller ist das Abbild auf dem Film. Da die Röntgenstrahlen üblicherweise mehrere Gewebetypen durchdringen ist das resultierende Bild die „Summe“ all dieser Gewebe. Die räumliche Auflösung von Röntgenbildern ist sehr hoch, Details mit einem Durchmesser von $0,1\,\mathrm{mm}$ können unterschieden werden. Die Idee der CT ist mit Hilfe von Röntgenstrahlung eine große Menge an Daten von jeder Seite des Patienten zu sammeln. Während der Untersuchung rotiert eine Röntgenquelle um den Patienten und sendet ein 1-D-Bündel Röntgenstrahlen aus. Entgegengesetzt von der Quelle befindet sich eine größere Anzahl von elektronischen Röntgendetektoren, die die Menge an Strahlung messen, die den Patienten durchdrungen hat. Die Quelle und der Detektor vollführen einen vollständigen Kreis um den Patienten, wobei die Detektoren mehrere tausend Röntgenstrahlen messen. Die Kenntnis der Röntgenphysik und der Aufnahmegeometrie ermöglicht die Rekonstruktion eines 2-D-Bildes (Schicht) aus den 1-D-Röntgenprojektionen. Ein 3-D-Volumen erhält man, indem mehrere 2-D-Schichten an unterschiedlichen Positionen aufgenommen werden. Nun ist es auch möglich 2-D-Schnitte mit beliebigem Winkel an einer beliebiger Position zu erzeugen (Multiplanare Rekonstruktion). Heutzutage beträgt die Auflösung von modernen CTs $0,2 \times 0,2 \times 0,4\,\mathrm{mm}^3$ $(x \times y \times z)$. CT- und herkömmliche Röntgengenbilder stellen Dichteunterschiede dar, d. h. morphologische

Unterschiede. Diese Techniken sind daher besonders geeignet um Brüche, Tumore, Atemwegserkrankungen wie beispielsweise Tuberkulose und andere Abnormalitäten, die mit einer Abweichung der Gewebedichte einhergehen, zu untersuchen.

Bei der Magnetresonanztomographie werden starke magnetische Felder benutzt, um 2-D-Schichtbilder des Patienten zu erzeugen. Die Spins aller Atomkerne in einer Schicht werden durch das magnetische Feld gleich ausgerichtet. Hochfrequenzsignale, die orthogonal zu der Schicht eingebracht werden, bringen einige der Wasserstoffatomkerne dazu, ihre Ausrichtung zu ändern. Sobald das Hochfrequenzsignal abgestellt wird geben die Wasserstoffatomkerne Radioenergie ab während sie ihre ursprüngliche Ausrichtung wieder annehmen. Detektoren, in diesem Fall Spulen, die um den Patienten herum angebracht sind, nehmen diese Hochfrequenzsignale auf. Wiederum erhält man aus mehreren 2-D-Schichten anatomische 3-D-Information. Die Auflösung moderner MRT Scanner beträgt $0,8 \times 0,8 \times 0,8\,\mathrm{mm}^3$ $(x \times y \times z)$. In der Medizin werden im Allgemeinen Wasserstoffatomkerne verwendet, allerdings könnten auch andere Atomkerne benutzt werden, was beispielsweise in der chemischen Forschung der Fall ist. Der Nutzen von MRT Aufnahmen liegt darin, anatomische Strukturen und Flüssigkeiten mit ähnlicher Dichte aber unterschiedlicher Anzahl an Wasserstoffatomkernen unterscheiden zu können, z. B. kann Fettgewebe mit wenig Wasseranteil von Blutgefäßen und anderen flüssigkeitsgefüllten Gebieten unterschieden werden.

Positronen-Emissionscomputertomographie ist ein medizinisches bildgebendes Verfahren, bei dem radioaktive Tracer in den Patienten injiziert werden. Ein Tracer besteht aus Positronen aussendenden Radionukliden, die zu gewöhnlichen Körperbestandteilen wie Glukose oder Wasser hinzugefügt werden. Der Tracer wird üblicherweise in den Blutkreislauf des Patienten injiziert. Die ausgesendeten Positronen treffen nach maximal einem Millimeter auf ein Elektron. Die Reaktion erzeugt ein Paar von Photonen (Gammastrahlung), die sich entgegengesetzt fortbewegen. Diese Gammastrahlung wird durch einen Detektorring aufgezeichnet und nur gleichzeitige Signale in entgegengesetzter Richtung werden weiterverarbeitet, die anderen werden als Rauschen betrachtet. Die Auflösung von PET Aufnahmen ist sehr gering: $4 \times 4 \times 6\,\mathrm{mm}$ $(x \times y \times z)$. Der Nutzen von PET Aufnahmen liegt darin, biochemische Prozesse studieren zu können, beispielsweise die Aktivität des Gehirns oder die Absorption von Glukose durch Gewebe, was auf einen Tumor hinweisen kann. PET ist daher ein funktionelles bildgebendes Verfahren.

Abbildung D.1 zeigt Beispielbilder der geschilderten Bildmodalitäten. Alle drei Modalitäten benötigen Computer, um die 3-D-Daten aus niedrigerdimensionalen Daten, die elektronisch mit einem Detektor aufgenommen werden, zu rekonstruieren.

Die Rekonstruktion von Lichtfeldern kann als ein neues Verfahren zur 3-D-Bildgebung be-

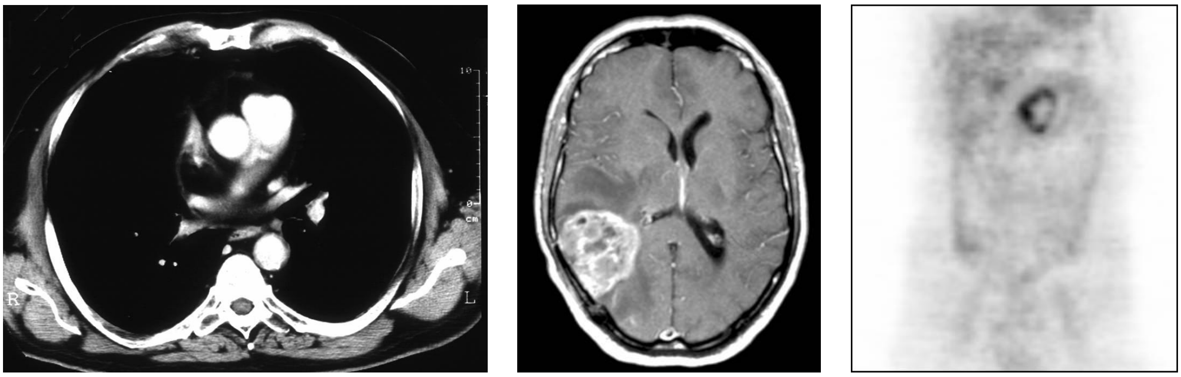


Abbildung D.1: Beispiele eines CT (links), MRT (mitte) und PET (rechts) Bildes. Das CT Bild zeigt einen Schnitt durch den Thorax (Lunge), das MRT Bild zeigt einen Schnitt durch den Kopf und das PET Bild zeigt einen „vertikalen“ (transversalen) Schnitt durch das Abdomen (Bilder mit freundlicher Genehmigung der Nuklearmedizinischen Klinik, Universität Erlangen-Nürnberg).

trachtet werden, wobei ähnlich zur CT 3-D-Information aus niedrigerdimensionalen Daten rekonstruiert wird. Das gemeinsame Prinzip ist die Rekonstruktion von n dimensionalen Daten aus $n-1$ dimensionalen Projektionen unter der Annahme, dass die Projektionsparameter bekannt sind: auf der einen Seite die Position der Röntgenquelle und des Detektors sowie die Gleichungen zur Beschreibung der Projektion von Röntgenstrahlung durch ein Objekt (Schwächungsgesetz), womit 2-D-Schichtbilder aus 1-D-Projektionen rekonstruiert werden können. Das 3-D-Volumen besteht aus aufeinanderfolgenden 2-D-Schichten. Auf der anderen Seite die Lage des Endoskops, die intrinsischen Kameraparameter und die Gleichung zur Beschreibung der optischen Projektion (Lochkameramodell, perspektivische Projektion), womit eine 3-D-Szene aus 2-D-Kamerabildern rekonstruiert werden kann.

D.2.4 Datenfusion

Wenn man sich den Nutzen von CT, MRT und PET vor Augen führt wird klar, dass kein perfektes bildgebendes Verfahren existiert. Jedes hat Vor- und Nachteile. CT Aufnahmen ermöglichen die Unterscheidung von Körperstrukturen mit unterschiedlicher Dichte. Vor allem kann die Anatomie von knöchernen Strukturen (hohe Dichte) beurteilt werden. MRT Aufnahmen sind für Weichgewebe gut geeignet. Ein hoher Kontrast ermöglicht die Erkennung pathologischer Abnormalitäten in Blutgefäßen und Organen, z. B. im Herz und der Prostata. Außerdem ist nach derzeitigem medizinischen Kenntnisstand eine MRT Untersuchung ungefährlich, da nichtionisierende Strahlung im Hochfrequenzbereich benutzt wird. Im Gegensatz zu CT und MRT, die beide anatomische Strukturen darstellen, veranschaulichen PET Aufnahmen biochemische Prozesse, wodurch Abnormalitäten erkannt werden können bevor sie im CT oder MRT Bild sichtbar werden. PET ist besonders zur Erkennung mehrerer Tumorarten und Metastasen geeignet, beispielsweise Tumore der Leber, der Lunge, der Brust und der Bauchspeicheldrüse. Der Nachteil der PET ist das Einbringen von radioaktivem Material in den Patienten und die niedrige Auflösung der erzeugten Bilder.

Die Motivation für Datenfusion entspringt dem Wunsch die Vorteile verschiedener Bildmodalitäten zu kombinieren. Sind zwei Bildmodalitäten und eine Anzahl von 2-D-Schichten für jede Modalität gegeben, stellt sich die Frage, welcher Voxel der ersten Modalität mit welchem Voxel der zweiten korrespondiert. Formaler ausgedrückt: eine Transformation, die das Koordinatensystem der ersten Modalität in das der zweiten Modalität abbildet, muss gefunden werden. Diesen Vorgang bezeichnet man als Registrierung. Ein Hauptproblem von Algorithmen zur Registrierung ist die Bewegung des Patienten und seine gewebedeformierenden Körperfunktionen wie beispielsweise die Atmung oder der Herzschlag. Vor allem Weichgewebe und deformierbare Organe finden sich nicht an der exakt gleichen Position wieder nachdem der Patient sich bewegt hat. Falls jedoch eine Transformation gefunden werden kann, ist es möglich die beiden Datensätze zu fusionieren, z. B. kann eine Schicht einer CT Untersuchung mit der korrespondierenden transformierten 2-D-Schicht einer PET Untersuchung überlagert werden und es ist möglich einen Tumor in der CT Aufnahme zu lokalisieren der nur bzw. deutlicher im PET Bild zu sehen ist (siehe Abbildung D.2). Natürlich können auch komplette 3-D-Datensätze fusioniert und gemeinsam dargestellt werden. Dann wird üblicherweise einer der beiden Datensätze semitransparent dargestellt. Eine schnelle Visualisierung wird durch Volume Rendering Algorithmen erreicht, welche die Graphikhardware ausnutzen.

Ähnlich zur Fusion von CT und PET kann ein Lichtfeld mit anderen zur Verfügung stehenden 3-D-Daten wie beispielsweise CT und MRT fusioniert werden. Der fusionierte Datensatz kann dann angezeigt werden um den Operateur während der Operation durch Augmented Reality zu unterstützen.

Fasst man die beiden letzten Kaptitel zusammen, kann man Lichtfelder als eine neue Art von 3-D bildgegbendem Verfahren, das direkt im Operationssaal unter Verwendung von 2-D-Endoskopiebildern erzeugt wird, betrachten. Außerdem können Lichtfelder mit 3-D-Daten von herkömmlichen bildgebenden Verfahren fusioniert werden. Abschließend ist zu bemerken, dass die Rekonstruktion von Lichtfeldern nicht nur während einer Operation sondern auch zur Diagnose durchgeführt werden kann.

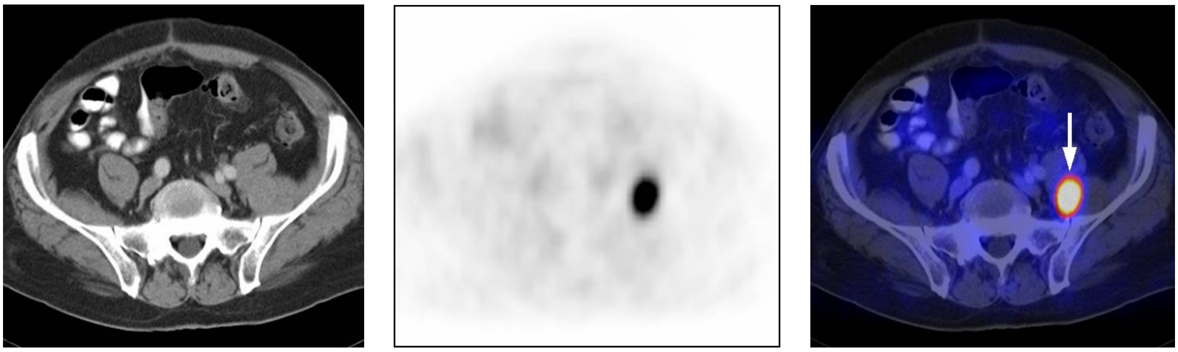


Abbildung D.2: Der Vorteil unterschiedlicher Bildmodalitäten ist sichtbar: der Tumor (eine Metastase im Abdomen), der nicht – oder nur für Experten – im CT Bild (links) sichtbar ist, ist im PET Bild als schwarzer Fleck deutlich sichtbar (mitte). Die Fusion beider Modalitäten (rechts) ermöglicht die exakte Lokalisierung des Tumors im CT Bild (Bilder mit freundlicher Genehmigung der Nuklearmedizinischen Klinik, Universität Erlangen-Nürnberg).

D.2.5 Übersicht

Diese Doktorarbeit ist wie folgt gegliedert: Kapitel 2 beschreibt den Stand der Technik im Bereich computerunterstützter endoskopischer Operationen; im Speziellen wird die Entwicklung von der herkömmlichen zur minimal-invasiven Chirurgie dargestellt. Dieses Kapitel fasst auch die neuesten Entwicklungen in diesem Bereich zusammen: roboterunterstütze Eingriffe, endoskopische Bildverbesserungsmethoden und medizinische Augmented Reality Systeme.

Die Lichtfeldtheorie wird in Kapitel 3 eingeführt. Nach der Definition eines Lichtfeldes werden bekannte Rekonstruktions- und Visualisierungstechniken zusammengefasst.

In Kapitel 4 werden Lösungen zur Reduktion störender Beeinträchtigungen der Bilder, die während endoskopischer Operationen auftreten, vorgestellt. Ein Gesamtsystem zur Bildverbesserung in Echtzeit wird beschrieben. Zusätzlich wird eine Methode zur Bildverbesserung mit Hilfe von Lichtfeldern dargestellt.

In Kapitel 5 werden drei unterschiedliche Wege zur Lichtfeldrekonstruktion beschrieben. Lagebestimmungssysteme werden verwendet um die Berechnungszeit zu reduzieren und die Robustheit zu erhöhen. Zwei Lagebestimmungssysteme werden untersucht: ein Roboterarm und ein optisches Trackingsystem. Zusätzlich wird die Rekonstruktion von Lichtfeldern lediglich aus dem Eingabe-Videostrom beschrieben und mit den anderen beiden Methoden verglichen.

Zur Bereitstellung von 3-D (und 2-D) Augmented Reality im Operationssaal wird das Lichtfeld mit CT Daten registriert und fusioniert. Die entwickelten Methoden werden in Kapitel 6 beschrieben. Zunächst werden wichtige anatomische Strukturen identifiziert, segmentiert und in einer Datenbank hinterlegt. Danach werden die Registrierparameter geschätzt und die fusionierten Daten visualisiert.

Experimente und Evaluationen der Methoden, die in den Kapiteln 4 bis 6 entwickelt wurden, werden in Kapitel 7 gezeigt. Die Doktorarbeit wird in Kapitel 8 zusammengefasst und mit einem Ausblick beendet.

D.3 Zusammenfassung

Die Entwicklung im Bereich der Chirurgie verläuft zu so genannten *minimal-invasiven* Operationen hin, die den Patienten deutlich weniger als die herkömmliche offene Chirurgie belasten. Die Grundidee bei minimal-invasiver Chirurgie ist, den Zugang zum Operationsgebiet durch kleine „Schlüssellöcher“ mit einem Durchmesser von 1 bis 2 cm herzustellen. Das Bild des Operationsgebiets wird mit Hilfe eines Endoskops dargestellt. Diese Doktorarbeit befasst sich mit solchen minimal-invasiven Operationen, bei denen starre monokulare Endoskope verwendet werden, beispielsweise die Entfernung der Gallenblase (Cholecystektomie). Im Vergleich zur herkömmlichen Chirurgie treten dabei mehrere Probleme auf. In dieser Doktorarbeit wurden Verfahren zur Reduzierung von dreien dieser Probleme, namentlich *Beeinträchtigungen des Bildes, Eingeschränkte Sicht* und *Verlust der stereoskopischen Tiefenwahrnehmung*, entwickelt. Ein Gesamtsystem, das im Operationssaal verwendet werden kann, wurde beschrieben. Es ermöglicht

- Bildverbesserung in Echtzeit,
- 3-D-Visualisierung des Operationsgebiets und
- Augmented Reality.

Mehrere Bildstörungen können in Echtzeit reduziert oder sogar behoben werden. Ein 3-D-Modell des Operationsgebiets, namentlich ein Lichtfeld, kann rekonstruiert und dreidimensional aus beliebigen Positionen betrachtet werden, beispielsweise auf einem 3-D-Monitor. Sowohl das 2-D-Livebild als auch das Lichtfeld kann durch CT-Daten überlagert werden, nachdem eine Registrierung basierend auf der rekonstruierten 3-D-Information durchgeführt wurde.

Die meisten bereits veröffentlichten Lösungen zur Behebung von Bildstörungen wurden nicht für den Einsatz im Operationssaal entwickelt und außer der in [Fis04] vorgestellten Arbeit wurden bisher lediglich Lösungen für einzelne Bildstörungen veröffentlicht. Außerdem wurde keine dieser Methoden von Ärzten evaluiert. Andere 3-D-Modelle wurden in der minimal-invasiven Chirurgie benutzt, jedoch wurden Lichtfelder bisher nicht untersucht. Es existieren mehrere Ansätze für medizinische Augmented Reality Systeme, sogar für minimal-invasive Operationen. In dieser Doktorarbeit wird eine intrinische Registrierung durchgeführt, wohingegen die meisten aktuellen Augmented Realtiy Ansätze extrinsische Registrierung basierend auf Markern verwenden.

Drei Verbesserungsmethoden für endoskopische Bilder wurden vorgeschlagen: Entzerrung, Farbnormierung und zeitliche Filterung. Verzerrungen des Bildes resultieren hauptsächlich aus der Verwendung von Linsen mit kleiner Brennweite. Sie können mit Hilfe der intrinsischen Kameraparameter, die durch Kamerakalibrierung bestimmt werden, korrigiert werden. Während einer minimal-invasiven Operation kann es vorkommen, dass das Gewebe im Operationsgebiet mit Blut bedeckt ist. In diesem Fall ist es schwer, verschiedene Gewebetypen zu unterscheiden. Eine Farbnormierung reduziert dieses Problem und erzeugt zusätzlich beleuchtungsunabhängige Bilder. Während des Schneidens von Gewebe mit Hochfrequenzschneidern entstehen Rauch und kleine umherfliegende Partikel. Diese Störungen des Bildes werden durch zeitliche Filterung reduziert. Alle drei Methoden zur Bildverbesserung können in Echtzeit auf einem modernen PC (Pentium 4, 3,2 GHz) angewandt werden. Sogar mit der Kombination aller drei Methoden ist es möglich, 13 Bilder pro Sekunde zu verarbeiten. Die Evaluation der Methoden durch 14 Ärzte ergab einen statistisch signifikanten Nutzen durch das verarbeitete Bild ($p \ll 10^{-4}$). Es stellte sich heraus, dass der Nutzen der zeitlichen Filterung nur sichtbar ist, wenn Bildsequenzen und nicht Einzelbilder betrachtet werden. Außer den Bildverbesserungsmethoden ermöglicht das System digitalen Zoom und das Konstanthalten des Horizonts während die Endoskopoptik samt Kamera gedreht wird. Ein Verfahren zur Entfernung von Bildstörungen wie beispielsweise Glanzlichter oder Verschmutzungen auf der Endoskoplinse wurde auch vorgestellt. Dieser Ansatz basiert auf einem statischen Lichtfeld und er ermöglicht die Entfernung von Bildstörungen, die nicht an der gleichen Position relativ zur Szene bleiben während sich das Endoskop bewegt.

Die Herausforderungen bei der Rekonstruktion von statischen Lichtfeldern aus endoskopischen Bildern sind die Bestimmung der extrinsischen Kameraparameter und die Berechnung von Tiefeninformation. Drei Möglichkeiten für die Berechnung der extrinsischen Kameraparameter wurden untersucht:

- Verwendung von Struktur-aus-Bewegung Verfahren,
- Verwendung des Endoskoppositionierroboters AESOP und
- Verwendung des optischen Trackingsystems smARTtrack1.

Für alle drei Methoden wurde angenommen, dass die intrinsischen Kameraparameter für alle aufgenommenen Bilder konstant sind. Diese Parameter werden dann im voraus durch ein Kamerakalibrierverfahren bestimmt. Abgesehen von dem Strukur-aus-Bewegung Ansatz wird Tiefeninformation in Form von 3-D-Punkten durch 2-D-Punktverfolgung von Bild zu Bild und anschließender Triangulierung von 3-D-Punkten anhand der bekannten intrinsischen und extrinsischen Kameraparameter, berechnet. Eine neue Repräsentation der Tiefeninformation für das Lichtfeldrendering in Form von 3-D-Dreiecksnetzen wurde eingeführt. Diese Repräsentation reduziert die benötigte Zeit für die Tiefenberechnung und beschleunigt das Rendering von Bildern.

Um die Qualität der berechneten 3-D-Punkte zu erhöhen, wurde ein LMedS Verfahren und nichtlineare Optimierung der extrinsischen Kameraparameter vorgeschlagen.

Es wurden Verfahren zur Bestimmung der Hand-Auge-Transformation von AESOP und smARTtrack1 entwickelt. Drei Targets zur Benutzung mit smARTtrack1 wurden entworfen und ihre Genauigkeit untersucht. Das *DD* Target ergab die kleinsten Endoskoplagefehler: $\overline{\epsilon}_{\boldsymbol{t},\mathrm{rel}} = 3,8\,\%$ $(1,5\,\mathrm{mm})$ und $\overline{\epsilon}_{\boldsymbol{R},\mathrm{rel}} = 2,7\,\%$ $(0,63°)$. Die Fehler bei Benutzung von AESOP waren ungefähr zehn Mal höher, wobei das Spiel der Endoskophalterung der Hauptgrund für diesen extremen Unterschied ist.

Mit allen drei Methoden wurden Lichtfelder rekonstruiert. Die meisten mit einem der beiden Lagebestimmungssysteme erstellten Lichtfelder wurden im Labor rekonstruiert, aber jedes System wurde auch im Operationssaal eingesetzt. Beide Lagebestimmungssysteme ermöglichen eine schnelle Rekonstruktion von Lichtfeldern im Operationssaal: die Rekonstruktion eines Lichtfeldes mit 155 Bildern benötigte ungefähr eine Minute. In Gegensatz hierzu dauerte die entsprechende Rekonstruktion mit dem Struktur-aus-Bewegung Ansatz fast zehn Minuten. Die Qualität der rekonstruierten Lichtfelder wurde subjektiv und objektiv evaluiert. Die subjektive Evaluation durch zehn Ärzte ergab einen deutlichen Unterschied zwischen der Qualität von Laborlichtfeldern (Note $2,5$) und Operationssaallichtfeldern (Note $3,3$). Die Evaluation in Form von $\overline{Q}_{\mathrm{MAD}}$, $\overline{Q}_{\mathrm{SNR}}$ und $\overline{Q}_{\mathrm{PSNR}}$ ermöglicht die objektive Bewertung der Lichtfelder. Die objektive Evaluation ergab den gleichen Unterschied zwischen Labor- und Operationssaallichtfeldern. Entsprechend der Einteilung von [Wan02] ist die Qualität von AESOP Lichtfeldern annähernd gut $(16 \leq \overline{Q}_{\mathrm{PSNR}} \leq 20)$, wohingegen die Qualität von smARTtrack1 Lichtfeldern gut und manchmal sogar sehr gut ist $(24 \leq \overline{Q}_{\mathrm{PSNR}} \leq 30)$. Die Qualität der Lichtfelder, die mit Struktur-aus-Bewegung rekonstruiert wurden, war vergleichbar. In dieser Doktorarbeit werden dynamische Lichtfelder durch mehrere statische Lichtfelder modelliert. Dementsprechend treffen alle Resultate für statische Lichtfelder auch auf sie zu. Zwei Ergebnisse dynamischer Lichtfelder wurden gezeigt, eines wurde im Labor berechnet, das andere während einer Cholecystektomie.

Abgesehen von der Berechnungszeit ist der hauptsächliche Nachteil des Struktur-aus-Bewegung Ansatzes seine Sensitivität bezüglich der Eingabeparameter: eine geringe Änderung der Parameter kann zu einem unterschiedlichen und manchmal unbrauchbaren Ergebnis führen. Daher müssen die Parameter normalerweise für jede Sequenz angepasst werden, was im Operationssaal nicht durchführbar ist. Im Gegensatz hierzu muss lediglich die Schwelle θ_{BPE} für den Rückprojektionsfehler bei Verwenden von AESOP oder smARTtrack1 angepasst werden. Der hauptsächliche Nachteil von AESOP ist der große Fehler der berechneten Endoskoplage. Zieht man diese Nachteile und die anderen Vor- und Nachteile aus Kapitel 5.8, Seite 118, in Betracht, wird vorgeschlagen, smARTtrack1 zusammen mit dem *DD* Target zur Rekonstruktion qualitativ hochwertiger Lichtfelder bei minimal-invasiven Operationen zu verwenden. Die Qualität des Lichtfelds kann außerdem verbessert werden, indem das vorgeschlagene LMedS Verfahren zur Triangulierung und die nicht-lineare Optimierung der extrinsischen Kameraparameter verwendet werden.

Das System stellt Augmented Reality zur Verfügung, indem CT Daten aus einer anatomischen Datenbank entweder über das gerenderte Lichtfeldbild oder über das Livebild überlagert werden. Im ersten Fall kann die Szene dreidimensional betrachtet werden, im zweiten Fall können lediglich 2-D-Bilder betrachtet werden. Um echte stereoskopische Tiefenwahrnehmung zu ermöglichen wird im ersten Fall ein 3-D-Monitor benutzt. Eine neue Methode zur intrinsischen Registrierung von CT Daten und Endoskop wurde entwickelt. Basierend auf der berechneten Tiefeninformation, können 3-D-Punktkorrespondenzen zur Grobregistrierung vom Chirurgen ausgewählt werden. Danach kann der Iterative-Closest-Point (ICP) Algorithmus zur Feinregistrierung benutzt werden. Zusätzlich kann die Registrierung durch den Chirurgen manuell verfeinert werden. Augmented Reality ermöglicht dem Chirurgen, hinter die Oberfläche, durch Organe und Gewebe hindurch, zu „sehen“, beispielsweise werden wichtige anatomische Strukturen wie z. B. Gefäße, die während der Operation nicht verletzt werden dürfen, sogar dann komplett sichtbar, wenn die Struktur nicht oder nur teilweise im Endoskopbild zu sehen ist.

List of Figures

1.1 CT, MRI, and PET Image Examples 8
1.2 Fused CT and PET Image 9

2.1 Conventional Open Cholecystectomy 13
2.2 Endoscopes 14
2.3 Laparoscopic Cholecystectomy 16
2.4 Operation Situation with an AESOP robot arm 20
2.5 Degradations in Endoscopic Images 21

3.1 The Plenoptic Function 34
3.2 Plenoptic Function Sample 35
3.3 Two-Plane Parameterization of Light Rays 39
3.4 Rendering of a Two-Plane Light Field 41
3.5 Use of Depth Information 42
3.6 Ghosting Artefacts 43
3.7 Free Form Light Field Visualization 44

4.1 A Real-Time Endoscopic Image Enhancement System 56
4.2 Examples of Endoscopic Image Enhancement 58
4.3 Distortion Examples 59
4.4 Calibration Patterns 61
4.5 Projective 2-D Mapping 64
4.6 Color Cluster Rotation 69
4.7 Temporal Filter Method by One Large Image 73
4.8 Temporal Filter Method by Several Small Images 73
4.9 Side View Endoscope 76
4.10 Advantage of Side View Endoscopes 76

4.11 Degradation Reduction by Light Fields 78
4.12 Di-Chromatic Reflectance Model 79
4.13 Example of Highlight Substitution 81

5.1 Preprocessing for Light Field Reconstruction 88
5.2 Endoscope vs. Camera Head Rotation 89
5.3 AESOP 3000 Kinematics 97
5.4 AESOP 3000 Extended Kinematics 98
5.5 Notch Detection 99
5.6 Optical Tracking System *smARTtrack1* 101
5.7 Targets and Calibration Kit for smARTtrack1 102
5.8 Synchronization of smARTtrack1 with Video-Endoscopic System 103
5.9 Target Adapter and *Epee* Target 104
5.10 *Epee* Target with Occlusion 105
5.11 *Double Decker* Target 105
5.12 *Double Decker 2z* Target 106
5.13 Triangulation From Two Views 113
5.14 3-D Triangular Mesh 115
5.15 3-D Triangular Mesh with Grid Points 116
5.16 Comparison of Depth Maps 116

6.1 Abdominal Anatomy 123
6.2 Costal Arch 124
6.3 Marching Cubes Algorithm 128
6.4 Augmented Reality Example 134

7.1 Experimental Setup in the Laboratory 137
7.2 Artificial Patient and Liver/Gall Bladder Model 138
7.3 Light Field Reconstruction in the Operating Room Using AESOP 138
7.4 Light Field Reconstruction in the Operating Room Using smARTtrack1 139
7.5 Hand-Eye Calibration in the Operating Room 140
7.6 Screenshot of *EvaMedIm* 141
7.7 Pose Error Measurement 143
7.8 Steps of the Camera Calibration Algorithm 150
7.9 Examples of Distortion Correction 152
7.10 Examples of Color Normalization I 153

7.11 Examples of Color Normalization II 154
7.12 Examples of Temporal Filtering 155
7.13 Examples of Horizon Equalization 156
7.14 Example of Digital Zoom 157
7.15 Calibration Sphere and Computed 3-D Points 169
7.16 Target for the Sony Camera 171
7.17 Subjective Light Field Evaluation 176
7.18 Light Field Reconstruction `Gall-20020708` 177
7.19 Light Field Reconstruction `Hyp-20010425` 178
7.20 Light Field `ALF 12` 179
7.21 Light Field `ALF 53` 180
7.22 Light Field `ALF 67` Using Structure-From-Motion and AESOP 181
7.23 Light Field `ART 22` 182
7.24 3-D Reconstruction of sequence `ART 22` Using Structure-From-Motion 183
7.25 LMedS and Non-Linear Optimization of Sequence `ART 22` 183
7.26 Non-Linear Optimization of the Extrinsic Camera Parameters of Sequence `ART 22` 183
7.27 Light Field `ART 75` 184
7.28 Light Field `ART 93` 185
7.29 3-D Reconstruction of sequence `ART 93` Using Structure-From-Motion 186
7.30 LMedS and Non-Linear Optimization of Sequence `ART 93` 186
7.31 Non-Linear Optimization of the Extrinsic Camera Parameters of Sequence `ART 93` 186
7.32 Light Field `ART 118` 187
7.33 Dynamic Light Field (Laboratory) 188
7.34 Dynamic Light Field (OR) 189
7.35 Examples of Highlight Detection 190
7.36 Examples of Highlight Substitution 191
7.37 Examples of Substituting Soilings on the Endoscope Lens 192
7.38 Examples of CT Data 194
7.39 Registration by Selecting Corresponding Points/Triangles 195
7.40 Examples of Augmented Reality in the Live-Image 196
7.41 Examples of Augmenting the Light Field 197

D.1 CT, MRT und PET Beispielbilder 235

D.2 Fusion eines CT und PET Bildes 237

List of Tables

2.1 Comparison of Robot and Human Characteristics 17
2.2 Degradations in Endoscopic Images 22
2.3 Comparison of Medical AR and 3-D Visualization Approaches 31

7.1 Computation Times of Image Enhancement Methods 148
7.2 Sensor Noise 148
7.3 Camera Calibration Results Using a 5 mm Endoscope 149
7.4 Camera Calibration Results Using a 10 mm Endoscope 150
7.5 Color Normalization 153
7.6 Computation Times of Some IPL Filters 153
7.7 Comparison of Spatial Color Median Filters 155
7.8 Subjective Evaluation of Image Enhancement Methods 158
7.9 Subjective Evaluation of Image Enhancement Methods, p-values 159
7.10 Objective Evaluation of Temporal Filtering 160
7.11 Light Field Reconstruction Using Structure-From-Motion 161
7.12 Endoscope Pose Accuracy Using AESOP 164
7.13 Light Field Reconstruction Using AESOP 164
7.14 AESOP LMedS, Non-Linear Optimization, and Structure-From-Motion 166
7.15 smARTtrack1 Hand-Eye Calibration Accuracy 167
7.16 Data Selection for Hand-Eye Calibration of smARTtrack1 168
7.17 Endoscope Pose Accuracy Using smARTtrack1 168
7.18 smARTtrack1 3-D Accuracy 169
7.19 Light Field Reconstruction Using smARTtrack1 170
7.20 smARTtrack1 LMeds, Non-Linear Optimization, and Structure-From-Motion . . 172
7.21 Computation Times for Structure-From-Motion Light Field Reconstruction . . . 173
7.22 Computation Times of Depth Computation Approaches 173

7.23 Influence of Hand-Eye Calibration on the Light Field Quality 174
7.24 Objective Evaluation of Highlight Substitution 192
7.25 Objective Evaluation of Substituting Soilings on the Endoscope Lens 193

Bibliography

[Abd98] G. Abdoulaev, S. Cadeddu, G. Delussu, M. Donizelli, L. Formaggia, A. Giachetti, E. Gobbetti, A. Leone, C. Manzi, P. Pili, and A. Scheinine. ViVa: The Virtual Vascular Project. *IEEE Transactions on Information Technology in Biomedicine*, 2(4):268–274, 1998.

[Abl02] V. Ablavsky, M. Snorrason, and C. J. Taylor. Real-time Autonomous Video Enhancement System (RAVE). In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 317–320, Rochester, USA, 2002. IEEE Computer Society Press, Los Alamitos.

[Ade91] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In M. Landy and J. A. Movshon, editors, *Computational Models of Visual Processing*, chapter 1. MIT Press, Cambridge, 1991.

[Adv05] Advanced Realtime Tracking GmbH. http://www.ar-tracking.com, last visited January 2005.

[Aio02] S. Aiono, J. M. Gilbert, B. Soin, P. A. Finlay, and A. Gordan. Controlled trial of the introduction of a robotic camera assistant (EndoAssist) for laparoscopic cholecystectomy. *Surgical Endoscopy*, 16(9):1267–1270, 2002.

[All98] M. E. Allaf, S. V. Jackman, P. G. Schulam, J. A. Cadeddu, B. R. Lee, R. G. Moore, and L. R. Kavoussi. Laparoscopic visual field - voice vs. foot pedal interfaces for control of the AESOP robot. *Surgical Endoscopy*, 12:1415–1418, 1998.

[Ami05] Software Amira. http://www.amiravis.com, last visited January 2005.

[And95] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorenson. *LAPACK Users'*

Guide. Society for Industrial and Applied Mathematics (SIAM) Publications, release 2.0, 2nd edition, 1995.

[Arb96] K. Arbter and G.-Q. Wei. Verfahren zum Nachführen eines Stereo-Laparoskopes in der minimalinvasiven Chirurgie. Deutsche Patent Nr. 195 29 950, 14. Nov., 1996.

[Aru87] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

[Asc05] Ascension Technology Corp. http://www.ascension-tech.com, last visited January 2005.

[Aue99] T. Auer, S. Brantner, and A. Pinz. The integration of optical and magnetic tracking for multi-user augmented reality. In M. Gervaut, D. Schmalstieg, and A. Hildebrand, editors, *Virtual Environments '99. Proceedings of the Eurographics Workshop*, pages 43–52, Vienna, Austria, 1999. Springer, Berlin, Heidelberg, New York.

[Bac97] I. Baca. Roboterarm in der laparoskopischen Chirurgie. *Der Chirurg*, 68:837–839, 1997.

[Bak01] S. Baker and I. Matthews. Equivalence and Efficiency of Image Alignment Algorithms. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1090–1097, Kauai, USA, 2001. IEEE Computer Society Press, Los Alamitos.

[Bal02] G. Ballantyne. Robotic surgery, telerobotic surgery, telepresence, and telementoring. Review of early clinical results. *Surgical Endoscopy*, 16(10):1389–1402, 2002.

[Bar98] W. L. Bargar, A. Bauer, and M. Börner. Primary and Revision Total Hip Replacement Using the ROBODOC system. *Clinical Orthopaedics & Related Research*, 354:82–91, 1998.

[Bes92] P. J. Besl and N. D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[Bop99] S. A. Boppart, T. F. Deutsch, and D. W. Rattner. Optical imaging technology in minimally invasive surgery. *Surgical Endoscopy*, 13:718–722, 1999.

[Bor02] R. Bornard, E. Lecan, L. Laborelli, and J.-H. Chenot. Missing Data Correction in Still Images and Image Sequences. In *Proceedings of ACM Multimedia*, pages 355–361, Juan-les-Pins, France, 2002. ACM Press, New York.

[Büh01] C. Bühler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen. Unstructured Lumigraph Rendering. In E. Fiume, editor, *Proceedings of ACM SIGGRAPH*, pages 425–432, New York, USA, 2001. ACM Press, New York.

[Cab04] J. J. Caban. Reconstruction and Enhancement in Monocular Laparoscopic Imagery. In J. D. Westwood, H. M. Hoffmann, G. T. Mogel, D. Stredney, and R. A. Robb, editors, *Proceedings of the 12th Annual Medicine Meets Virtual Reality Conference (MMVR)*, pages 37–39, Newport Beach, USA, 2004. IEEE Computer Society Press, Los Alamitos.

[Cam99] S. Campagna. *Polygonreduktion zur effizienten Speicherung, Übertragung und Darstellung komplexer polygonaler Modelle*. PhD thesis, University Erlangen-Nuremberg, Germany, 1999.

[Cha00] J.-X. Chai, X. Tong, S.-C. Chand, and H.-Y. Shum. Plenoptic Sampling. In K. Akeley, editor, *Proceedings of ACM SIGGRAPH*, pages 307–318, New Orleans, USA, 2000. ACM Press, New York.

[Che91] H. Chen. A Screw Motion Approach to Uniqueness Analysis of Head-Eye Geometry. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 145–151, Maui, Hawaii, 1991. IEEE Computer Society Press, Los Alamitos.

[Che92] Y. Chen and G. Medioni. Object Modelling by Registration of Multiple Range Images. *Image and Vision Computing*, 10(3):145–155, 1992.

[Cho91] J. C. K. Chou and M. Kamel. Finding the Position and Orientation of a Sensor on a Robot Manipulator Using Quaternions. *International Journal of Robotics Research*, 10(3):240–254, 1991.

[Com05] Computer Motion Inc. http://www.computermotion.com, last visited January 2005.

[Coo92] A. M. Cooperaman. *Laparoscopic Cholecystectomy - Difficult Cases & Creative Solutions*. Quality Medical Publishing Inc., St. Louis, 1992.

[Cor00] J. Cortadellas, G. Bellaire, and G. Graschew. New Concepts for Intraoperative Navigation: Calibration of a 3-D Laparoscope. In A. Horsch and T. Lehmann, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 158–162, Munich, Germany, 2000. Springer, Berlin, Heidelberg, New York.

[Dan99] K. Daniilidis. Hand-Eye Calibration Using Dual Quaternions. *International Journal of Robotics Research*, 18:286–298, 1999.

[Dan01] K. Daniilidis. Using the Algebra of Dual Quaternions for Motion Alignment. In G. Sommer, editor, *Geometric Computing with Clifford Algebras*, chapter 20, pages 489–500. Springer, Berlin, Heidelberg, New York, 2001.

[DB01] S. De Buck, J. V. Cleynenbreugel, I. Geys, T. Koninckx, P. R. Koninckx, and P. Suetens. A system to support laparoscopic surgery by augmented reality visualization. In W. J. Niessen and M. A. Viergever, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 691–698, Utrecht, Netherlands, 2001. Springer, Berlin, Heidelberg, New York.

[Den83] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, New Jersey, 1983.

[Dev01] F. Devernay, F. Mourgues, and È. Coste-Manière. Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In *Proceedings of the International Workshop on Medical Imaging and Augmented Reality (MIAR)*, pages 16–20, Hong Kong, China, 2001. IEEE Computer Society Press, Los Alamitos.

[Dey00] D. Dey, P. J. Slomka, D. G. Gobbi, and T. M. Peters. Mixed Reality Merging of Endoscopic Images and 3-D Surfaces. In S. L. Delp, A. M. DiGioia, and B. Jaramaz, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 796–803, Pittsburgh, USA, 2000. Springer, Berlin, Heidelberg, New York.

[Dey02] D. Dey, D. G. Gobbi, P. J. Slomka, K. J. M. Surry, and T. M. Peters. Automatic Fusion of Freehand Endoscopic Brain Images to Three-Dimensional Surfaces: Creating Stereoscopic Panoramas. *IEEE Transactions on Medical Imaging*, 21(1):23–30, 2002.

[Dod95] N. A. Dodgson, N. E. Wiseman, S. R. Lang, D. C. Dunn, and A. R. L. Travis. Autostereoscopic 3D display in laparoscopic surgery. In H. Lemke, K. Inamura, C. Jaffe, and M. Vannier, editors, *Proceedings of the Computer Assisted Radiology Conference (CAR)*, pages 1139–1144, Berlin, Germany, 1995. Springer, Berlin, Heidelberg, New York.

[Dod00] N. A. Dodgson, J. R. Moore, S. R. Lang, G. Martin, and P. Canepa. Time-sequential multi-projector autostereoscopic 3D display. *Journal of the Society for Information Display*, 8(2):169–176, 2000.

[Ell03] J. Ellsmere, J. Stoll, D. Rattner, D. Brooks, R. Kane, W. Wells, R. Kikinis, and K. Vosburgh. A Navigation System for Augmenting Laparoscopic Ultrasound. In R. E. Ellis and T. M. Peters, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 184–191, Montreal, Canada, 2003. Springer, Berlin, Heidelberg, New York.

[Fau93] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, 1993.

[Fed01] P. A. Federspil, J. Stallkamp, and P. K. Plinkert. Robotik - Eine neue Dimension in der HNO-Heilkunde. *HNO*, 7:505–513, 2001.

[Feu05] M. Feuerstein, S. M. Wildhirt, R. Bauernschmitt, and N. Navab. Automatic Patient Registration for Port Placement in Minimally Invasive Endoscopic Surgery. In J. Duncan and G. Gerig, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 287–294, Palm Springs, USA, 2005. Springer, Berlin, Heidelberg, New York.

[Fin95] G. D. Finlayson. Color Consistancy in Diagonal Chromaticity Space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 218–223, Massachusetts, USA, 1995. IEEE Computer Society Press, Los Alamitos.

[Fin98] G. D. Finlayson, B. Schiele, and J. L. Crowley. Comprehensive colour image normalization. In H. Burkhard and B. Neumann, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 475–490, Freiburg, Germany, 1998. Springer, Berlin, Heidelberg, New York.

[Fis04] B. Fischer, B. Vaessen, T. M. Lehmann, and K. Spitzer. Bildverbesserung endoskopischer Videosequenzen in Echtzeit. In T. Tolxdorff, J. Braun, H. Handels,

A. Horsch, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 386–389, Berlin, Germany, 2004. Springer, Berlin, Heidelberg, New York.

[Fit95] A. W. Fitzgibbon and R. B. Fischer. A Buyer's Guide to Conic Fitting. In *Proceedings of the 5th British Machine Vision Conference*, pages 513–522, Birmingham, UK, 1995. BMVA Press, Surrey.

[Fuc98] H. Fuchs, M. A. Livingston, R. Raskar, D. Colucci, K. Keller, A. State, J. R. Crawford, P. Rademacher, S. H. Drake, and A. A. Meyer. Augmented Reality Visualization for Laparoscopic Surgery. In W. M. Wells, A. Colchester, and S. Delp, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 934–943, Cambridge, USA, 1998. Springer, Berlin, Heidelberg, New York.

[Fus99] A. Fusiello, E. Trucco, T. Tommasini, and V. Roberto. Improving Feature Tracking with Robust Statistics. *Pattern Analysis and Application*, 2(4):312–320, 1999.

[Gal99] K. Galousi, S. A. Karkanis, and D. E. Maroulis. Classification of Endoscopic Images Based on Texture Spectrum. In *Proceedings of the Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence (ACAI)*, pages 63–69, Chania, Greece, 1999. Technical Report, ACAI-99 Organization Committee.

[Get02] M. Gettman, R. Peschel, R. Neururer, and G. Bartsch. A Comparison of Laparoscopic Pyeloplasty Performed with the daVinci Robotic System versus Standard Laparoscopic Techniques: Initial Clinical Results. *Eur Urol*, 42(5):453–458, 2002.

[Gev99] T. Gevers and A. W. M. Smeulders. Color Based Object Recognition. *Pattern Recognition*, 32:453–464, 1999.

[Gev00] T. Gevers and H. M. G. Stokman. Classifying Color Transitions into Shadow-Geometry, Illumination Highlight or Material Edges. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 521–524, Vancouver, Canada, 2000. IEEE Computer Society Press, Los Alamitos.

[Gib03] J. D. Gibbons. *Nonparametric Statistical Inference*. Marcel Dekker, New York, 4th edition, 2003.

[Gir00] B. Girod, G. Greiner, and H. Niemann, editors. *Principles of 3D Image Analysis and Synthesis*. Kluwer Academic Publishers, Dordrecht, 2000.

[Gla95] D. Glauser, H. Frankhauser, M. Epitaux, J.-L. Hefti, and A. Jaccottet. Neurosurgical Robot Minerva: first results and current developments. *Journal Image Guided Surgery*, 1:266–272, 1995.

[Gol02] B. Goldlücke, M. Magnor, and B. Wilburn. Hardware-Accelerated Dynamic Light Field Rendering. In G. Greiner, H. Niemann, T. Ertl, B. Girod, and H.-P. Seidel, editors, *Proceedings of the Workshop Vision, Modeling and Visualization (VMV)*, pages 455–461, Erlangen, Germany, 2002. Aka GmbH, Berlin.

[Gor96] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The Lumigraph. In H. Rushmeier, editor, *Proceedings of ACM SIGGRAPH*, pages 43–54, New Orleans, USA, 1996. ACM Press, New York.

[Grä03] C. Gräßl, T. Zinßer, and H. Niemann. Illumination Insensitive Template Matching with Hyperplanes. In B. Michaelis and G. Krell, editors, *Proceedings of the Conference on Pattern Recognition, 25th DAGM Symposium*, pages 273–280, Magdeburg, Germany, 2003. Springer, Berlin, Heidelberg, New York.

[Grö01] M. Gröger, W. Sepp, T. Ortmaier, and G. Hirzinger. Reconstruction of Image Structure in Presence of Specular Reflections. In B. Radig and S. Florczyk, editors, *Proceedings of the Conference on Pattern Recognition, 23th DAGM Symposium*, pages 53–60, Munich, Germany, 2001. Springer, Berlin, Heidelberg, New York.

[Haj01] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes. *Medical Image Registration*. CRC Press, London, 2001.

[Har94] R. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem. *International Journal of Computer Vision*, 13(3):331–356, 1994.

[Har97] R. I. Hartley. In Defense of the Eight-Point Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997.

[Har03] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2nd edition, 2003.

[Has99] P. Hastreiter. *Registrierung und Visualisierung medizinischer Bilddaten unterschiedlicher Modalitäten*. PhD thesis, University Erlangen-Nuremberg, Germany, 1999.

[Hay01] M. Hayashibe and Y. Nakamura. Laser-Pointing Endoscope System for Intra-Operative 3D Geometric Registration. In *Proceedings of the IEEE International Conference on Robotics & Automation*, pages 1543–1548, Seoul, Korea, 2001. IEEE Computer Society Press, Los Alamitos.

[Hei99] B. Heigl, R. Koch, M. Pollefeys, J. Denzler, and L. van Gool. Plenoptic Modeling and Rendering from Image Sequences Taken by a Hand-Held Camera. In W. Förstner, J. M. Buhmann, A. Faber, and P. Faber, editors, *Mustererkennung, 21th DAGM Symposium*, pages 94–101, Bonn, Germany, 1999. Springer, Berlin, Heidelberg, New York.

[Hei04] B. Heigl. *Plenoptic Scene Modeling from Uncalibrated Image Sequences*. ibidem, Hannover, 2004.

[Hel01] J. P. Helferty, C. Zhang, G. McLennan, and W. E. Higgins. Videoendoscopic Distortion Correction and its Application to Virtual Guidance of Endoscopy. *IEEE Transactions on Medical Imaging*, 20(7):605–617, 2001.

[Hof02] M. Hoffmann. Endoskopische 3D-Vermessung biologischer Oberflächen. *Biomedizinische Technik*, 47(1):674–677, 2002.

[Höh00] K. H. Höhne, S. Gehrmann, S. Noster, B. Pflesser, A. Pommert, M. Riemer, T. Schiemann, R. Schubert, U. Schumacher, and U. Tiede. A realistic 3D atlas of the inner organs based on the Visible Human data. In H. U. Lemke, M. W. Vannier, K. Inamura, A. G. Farman, and K. Doi, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 14th International Congress and Exhibition*, pages 625–628, San Francisco, USA, 2000. Elsevier, Amsterdam.

[Hol01] K. Holl, W. Wies, and G. Wurm. Clinical Benefit of Image Guidance in Neurosurgery. In H. U. Lempke, M. W. Vannier, K. Inamura, A. G. Farman, and K. Doi, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 15th International Congress and Exhibition*, pages 99–101, Berlin, Germany, 2001. Elsevier, Amsterdam.

[Hon03] M. Honl, O. Dierk, C. Gauck, V. Carrero, F. Lampe, S. Dries, M. Quante, K. Schwieger, E. Hille, and M. M. Morlock. Comparison of robotic-assisted and manual implantation of a primary total hip replacement. A prospective study. *J Bone Joint Surg Am*, 85-A(8):1470–1478, 2003.

[Hor95] R. Horaud and F. Dornaika. Hand-Eye Calibration. *International Journal of Robotics Research*, 14(3):195–210, 1995.

[How99] R. D. Howe and Y. Matsuoka. Robotics for Surgery. *Annu. Rev. Biomed. Eng.*, 1:211–240, 1999.

[Hub03] D. Huber and M. Hebert. Fully Automatic Registration of Multiple 3D Data Sets. *Image and Vision Computing*, 21(7):637–650, 2003.

[Hum02] J. Hummel, M. Figl, C. Kollmann, and H. Bergmann. Evaluation of a miniature electromagnetic position tracker. *Med Phys*, 29(10):2205–2212, 2002.

[Ind05] Indeed-Visual Concepts GmbH. http://www.indeed3d.com, last visited January 2005.

[Int05a] Intel Integrated Performance Primitives. http://www.intel.com/software/products/ipp, last visited January 2005.

[Int05b] Intuitive Surgical Inc. http://www.intuitivesurgical.com, last visited January 2005.

[Jac97] L. Jacobs, V. Shayani, and J. Sackier. Determination of the learning curve of the AESOP robot. *Surgical Endoscopy*, 11:54–55, 1997.

[Jin01] H. Jin, P. Favaro, and S. Soatto. Real-Time Feature Tracking and Outlier Rejection with Changes in Illumination. In *Proceedings of the International Conference On Computer Vision (ICCV)*, pages 684–689, Vancouver, Canada, 2001. IEEE Computer Society Press, Los Alamitos.

[Joh97] A. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, USA, 1997.

[Kar03] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras. Computer-Aided Tumor Detection in Endoscopic Video Using Color Wavelet Features. *IEEE Transactions on Information Technology in Biomedicine*, 7(3):141–149, 2003.

[Koc99a] R. Koch, B. Heigl, M. Pollefeys, L. van Gool, and H. Niemann. A Geometric Approach to Lightfield Calibration. In F. Solina and A. Leonardis, editors, *Proceedings of the Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 596–603, Ljubliana, Slovenia, 1999. Springer, Berlin, Heidelberg, New York.

[Koc99b] R. Koch, M. Pollefeys, B. Heigl, L. van Gool, and H. Niemann. Calibration of Hand-held Camera Sequences for Plenoptic Modeling. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 585–591, Corfu, Greece, 1999. IEEE Computer Society Press, Los Alamitos.

[Koc01] R. Koch, B. Heigl, and M. Pollefeys. Image-based Rendering from Uncalibrated Lightfields with Scaleable Geometry. In R. Klette, T. Huang, and G. Gimel'farb, editors, *Multi-Image Search and Analysis, Proceedings of the Conference on Theoretical Foundations of Computer Vision*, pages 51–66, Dagstuhl, Germany, 2001. Springer, Berlin, Heidelberg, New York.

[Kon98] W. Konen, M. Scholz, and S. Tombrock. The VN project: endoscopic image processing for neurosurgery. *Computer Aided Surgery*, 3(3):145–148, 1998.

[Kop01] D. Koppel, Y.-F. Wang, and H. Lee. Automated Image Rectification in Video-Endoscopy. In W. J. Niessen and M. A. Viergever, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 1412–1414, Utrecht, Netherlands, 2001. Springer, Berlin, Heidelberg, New York.

[Kop04] D. Koppel, Y.-F. Wang, and H. Lee. Image-based Rendering and Modelling in Video-Endoscopy. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 269–275, Arlington, USA, 2004. IEEE Computer Society Press, Los Alamitos.

[Kor04] W. Korb, T. Bodenmüller, G. Eggers, T. Ortmaier, M. Schneberger, M. Suppa, J. Wiechnik, R. Marmulla, and S. Hassfeld. Surface-based Image-to-Patient Registration using a Hand-guided Laser-range Scanner System. In H. U. Lempke, K. Inamura, K. Doi, M. W. Vannier, A. G. Farman, and J. H. C. Reiber, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 18th International Congress and Exhibition*, page 1328, Chicago, USA, 2004. Elsevier, Amsterdam.

[Krü03a] S. Krüger, F. Vogt, W. Hohenberger, D. Paulus, H. Niemann, and C. H. Schick. Evaluation der rechnergestützten Bildverbesserung in der Videoendoskopie von Körperhöhlen. In T. Wittenberg, P. Hastreiter, U. Hoppe, H. Handels, A. Horsch, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 293–297, Erlangen, Germany, 2003. Springer, Berlin, Heidelberg, New York.

[Krü03b] S. Krüger, F. Vogt, D. Paulus, H. Niemann, W. Hohenberger, and C. H. Schick. Computer Improved Reality - Rechnergestützte Bildverbesserung in der Videoendoskopie von Körperhöhlen. In N. Haas and H. Bauer, editors, *Zurück in die Zukunft, 120. Deutscher Chirurgenkongress (DGCH) 2003*, pages 53–54. Springer, Berlin, Heidelberg, New York, 2003.

[Krü04] S. Krüger, F. Vogt, W. Hohenberger, D. Paulus, H. Niemann, and C. H. Schick. Evaluation of Computer-assisted Image Enhancement in Minimal Invasive Endoscopic Surgery. *Methods of Information in Medicine*, 43:362–366, 2004.

[Küb02] C. Kübler, J. Raczkowsky, and H. Wörn. Rekonstruktion eines 3D-Modells aus endoskopischen Bildfolgen. In M. Meiler, H. Handels, F. Kruggel, T. Lehmann, and D. Saupe, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 211–214, Leipzig, Germany, 2002. Springer, Berlin, Heidelberg, New York.

[Küh00] U. Kühnapfel, H. Cakmak, and H. Maaß. Endoscopic surgery training using virtual reality and deformable tissue simulation. *Computers & Graphics*, 24:671–682, 2000.

[Lee80] D. T. Lee and B. J. Schachter. Two Algorithms for Constructing a Delaunay Triangulation. *International Journal Computer and Information Sciences*, 9(3):219–242, 1980.

[Leh99] T. M. Lehmann, C. Gönner, and K. Spitzer. Survery: Interpolation Methods in Medical Image Processing. *IEEE Transactions on Medical Imaging*, 18(11):1049–1075, 1999.

[Lev96] M. Levoy and P. Hanrahan. Light Field Rendering. In H. Rushmeier, editor, *Proceedings of ACM SIGGRAPH*, pages 31–42, New Orleans, USA, 1996. ACM Press, New York.

[LH81] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[Li98] W. Li, Q. Ke, X. Huang, and N. Zheng. Light Field Rendering of Dynamic Scene. *Machine Graphics and Vision*, 7(3):551–563, 1998.

[Lié01] M. Liévin and E. Keeve. Stereoscopic Augmented Reality System for Computer Assisted Surgery. In H. U. Lempke, M. W. Vannier, K. Inamura, A. G. Farman, and K. Doi, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 15th International Congress and Exhibition*, pages 108–112, Berlin, Germany, 2001. Elsevier, Amsterdam.

[Lil67] H. W. Lilliefor. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association (JASA)*, 62:399–402, 1967.

[Lin80] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[Lor87] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Proceedings of the 14th annual conference on computer graphics and interactive techniques*, pages 163–169, Anaheim, USA, 1987. ACM Press, New York.

[Lor95] A. Lorusso, D. W. Eggert, and R. B. Fisher. A comparison of four algorithms for estimating 3-D rigid transformations. In *Proceedings of the British conference on machine vision (BMVC)*, pages 237–246, Birmingham, UK, 1995. BMVA Press, Surrey.

[Mai98] J. Maintz and M. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[Mar03a] D. E. Maroulis, D. K. Iakovidis, S. A. Karkanis, and D. A. Karras. ColD: a versatile detection system for colorectal lesions in endoscopy video-frames. *Computer Methods and Programs in Biomedicine*, 70:151–166, 2003.

[Mar03b] S. Martelli, S. Bignozzi, M. Bontempi, S. Zaffagnini, and L. Garcia. Comparison of an Optical and a Mechanical Navigation System. In R. E. Ellis and T. M. Peters, editors, *Proceedings of the International Conference on Medical Image Computing* *and Computer-Assisted Intervention (MICCAI)*, pages 303–310, Montreal, Canada, 2003. Springer, Berlin, Heidelberg, New York.

[Maz04] F. Mazoochian, C. Pellengahr, A. Huber, J. Kircher, H. J. Refior, and V. Jansson. Low accuracy of stem implementation in THR using the CASPAR-system. *Acta Orthopaedica Scandinavica*, 75(3):261–264, 2004.

[McC76] C. S. McCamy, H. Marcus, and J. G. Davidson. A color-rendition chart. *J. App. Photog. Eng.*, 2:95–99, 1976.

[McK91] P. J. McKerrow. *Introduction to Robotics*. Addison-Wesley Publishing Company, Sydney, 1991.

[Met98] L. Mettler, M. Ibrahim, and W. Jonat. One Year of Experience Working with the Aid of a Robotic Assistant the Voice-controlled optic holder AESOP in Gynaecological Endoscopic Surgery. *Human Reproduction*, 13(10):2748–2750, 1998.

[MeV05a] Center for Medical Diagnostic Systems and Visualization. http://www.mevis.de, last visited January 2005.

[MeV05b] Software MeVisLab. http://www.mevislab.de, last visited January 2005.

[Mil99] P. Milgram and H. Colquhoun Jr. A Taxonomy of Real and Virtual World Display Integration. In Y. Ohta and H. Tamura, editors, *Mixed Reality — Merging Real and Virtual Worlds*, chapter 1, pages 5–30. Springer, Berlin, Heidelberg, New York, 1999.

[Moh96] R. Mohr and B. Triggs. Projective geometry for image analysis. In *International Symposium on Photogrammetry and Remote Sensing*, Vienna, Austria, 1996. RICS Books, London. Tutorial.

[Mor77] J. J. More. The Levenberg-Marquardt Algorithm, Implementation, and Theory. In G.A. Watson, editor, *Numerical Analysis, Lecture Notes in Mathematics*, volume 630, pages 105–116. Springer, Berlin, Heidelberg, New York, 1977.

[Mou01] F. Mourgues, F. Devernay, and È. Coste-Manière. 3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery. In *Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR)*, pages 191–192, New York, USA, 2001. IEEE Computer Society Press, Los Alamitos.

[Mül02] A. Müller, M. Schubert, and E. Beleites. Noncontact Three-Dimensional Laser Measuring Device for Tracheoscopy. *Annals of Otology, Rhinology & Laryngology*, 111(9):821–827, 2002.

[Mün03] C. Münzenmayer, F. Naujokat, S. Mühldorfer, and T. Wittenberg. Enhancing Texture Analysis by Color Shading Correction. In *Proceedings of the Workshop Farbbildverarbeitung*, pages 35–42, Esslingen, Germany, 2003. Zentrum für Bild- und Signalverarbeitung e.V., Ilmenau.

[Mün04] C. Münzenmayer, F. Naujokat, S. Mühldorfer, B. Mayinger, and T. Wittenberg. Lineare Farbkorrektur zur automatischen Gewebeerkennung in der Endoskopie des Ösophagus. In T. Tolxdorff, J. Braun, H. Handels, A. Horsch, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 15–19, Berlin, Germany, 2004. Springer, Berlin, Heidelberg, New York.

[Neb03] P. B. Nebot, Y. Jain, K. Haylett, R. Stone, and R. McCloy. Comparison of task performance of the camera-holder robots EndoAssist and AESOP. *Surg Laparosc Endosc Percutan Tech*, 13(5):334–338, 2003.

[Nie03] H. Niemann, J. Denzler, B. Heigl, F. Vogt, C. Schick, S. Krüger, and W. Hohenberger. Image-Based Modeling and its Application in Image Processing. In *6th German-Russian Workshop Pattern Recognition and Image Understanding*, pages 14–17, Katun, Russian Federation, 2003. Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences, Novosibirsk.

[Nie04] H. Niemann, J. Denzler, B. Heigl, F. Vogt, C. Schick, S. Krüger, and W. Hohenberger. Image-Based Modeling and its Application in Image Processing. *Pattern Recognition and Image Analysis*, 14(2):184–189, 2004.

[Nie05] H. Niemann and I. Scholz. Evaluating the Quality of Light Fields Computed from Hand-held Camera Images. *Pattern Recognition Letters*, 26(3):239–249, 2005.

[Nim04] C. Nimsky, J. Rachinger, H. Iro, and R. Fahlbusch. Adaptation of a hexapod-based robotic system for extended endoscope-assisted transsphenoidal skull base surgery. *Minimal Invasive Neurosurgery*, 47(1):41–46, 2004.

[Nio01] D. Nio, W. Bemelman, K. den Boer, M. Dunker, D. Gouma, and T. van Gulik. Efficiency of manual versus robotical (ZEUS) assisted laparoscopic surgery in the performance of standardized tasks. *Surgical Endoscopy*, 16(3):412–415, 2001.

[Nor05] Northern Digital Inc. http://www.ndigital.com, http://www.ndieurope.com, last visited January 2005.

[Oja84] E. Oja and J. Parkkinen. On Subspace Clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 692–695, San Diego, USA, 1984. IEEE Computer Society Press, Los Alamitos.

[Olb05] B. Olbrich, J. Traub, S. Wiesner, A. Wiechert, H. Feußner, and N. Navab. Respiratory Motion Analysis: Towards Gated Augmentation of the Liver. In H. U. Lempke, K. Inamura, K. Doi, M. W. Vannier, and A. G. Farman, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 19th International Congress and Exhibition*, pages 248–253, Berlin, Germany, 2005. Elsevier, Amsterdam.

[Omo99] K. Omote, A. Ungeheuer, H. Feussner, K. Arbter, G.-Q. Wei, J. R. Siewert, and G. Hirzinger. Self-guided robotic camera control for laparoscopic surgery compared with human camera control. *American Journal of Surgery*, 177:321–324, 1999.

[Ope05] Open Computer Vision Library. http://sourceforge.net/projects/opencvlibrary, last visited January 2005.

[Pal99] C. Palm, T. Lehmann, and K. Spitzer. Bestimmung der Lichtquellenfarbe bei der Endoskopie makrotexturierter Oberflächen des Kehlkopfs. In K.-H. Franke, editor, *Proceedings of the Workshop Farbbildverarbeitung*, pages 3–10, Berlin, Germany, 1999. Schriftenreihe des Zentrums für Bild- und Signalverarbeitung e.V., Ilmenau.

[Pau98] D. Paulus, L. Csink, and H. Niemann. Color Cluster Rotation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 161–165, Chicago, USA, 1998. IEEE Computer Society Press, Los Alamitos.

[Pet00] J. Petermann, R. Kober, R. Heinze, J. J. Fröhlich, P. F. Heeckt, and L. Gotzen. Computer-assisted planning and robot-assisted surgery in anterior cruciate ligament reconstruction. *Op Tech Orthop*, 10:50–55, 2000.

[Plu03] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[Pra02] S. Prasad, H. Maniar, N. Soper, R. Damiano, and M. Klingensmith. The effect of robotic assistance on learning curves for basic laparoscopic skills. *Am J Surg*, 183(6):702–707, 2002.

[Prü05] M. Prümmer, E. Nöth, J. Hornegger, and A. Horndasch. Mensch-Maschine Interaktion für den interventionellen Einsatz. In H.-P. Meinzer, H. Handels, A. Horsch, and T. Tolxdorff, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 485–489, Heidelberg, Germany, 2005. Springer, Berlin, Heidelberg, New York.

[Psc98] *Pschyrembel Klinisches Wörterbuch.* Walter de Gruyter, Berlin, 258. edition, 1998.

[Reh98] V. Rehrmann and L. Priese. Fast and Robust Segmentation of Natural Color Scenes. In R. T. Chin and T.-C. Pong, editors, *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 598–606, Hong Kong, China, 1998. Springer, Berlin, Heidelberg, New York.

[Reu98] M. A. Reuter. *Geschichte der Endoskopie.* Karl Krämer Verlag, Stuttgart, 1998.

[Ric05] Richard Wolf GmbH. http://www.richard-wolf.com, last visited January 2005.

[Rou87] P. Roussseeuw and A. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, 1987.

[Rus01] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. In D. C. Young, editor, *Proceedings of the International Conference on 3D Digital Imaging and Modeling*, pages 145–152, Quebec City, Canada, 2001. IEEE Computer Society Press, Los Alamitos.

[Sal01] T. Salb, O. Burgert, T. Gockel, B. Giesler, and R. Dillmann. Comparison of Tracking Techniques for Intraoperative Presentation of Medical Data using a See-Through Head-Mounted Display. In J. D. Westwood, H. M. Hoffmann, G. T. Mogel, D. Stredney, and R. A. Robb, editors, *Proceedings of the 9th Annual Medicine Meets Virtual Reality Conference (MMVR)*, pages 443–445, Newport Beach, USA, 2001. IOS Press, Amsterdam.

[Sal02] J. Salvi, X. Armangue, and J. Batlle. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition*, 35(7):1617–1635, 2002.

[Sch98] M. Scholz, W. Konen, S. Tombrock, B. Fricke, L. Adams, M. von Düring, A. Hentsch, L. Heuser, and A. G. Harders. Development of an Endoscopic Navigation System Based on Digital Image Processing. *Computer Aided Surgery*, 3(3):134–143, 1998.

[Sch00] K. Schlüns and A. Koschan. Global and local highlight analysis in color images. In *Proceedings of the First International Conference on Color in Graphics and Image Processing (CGIP)*, pages 147–151, St. Etienne, France, 2000. Cepadues editions, Toulouse.

[Sch01a] M. Scheuering, C. Rezk-Salama, H. Barfuss, K. Barth, A. Schneider, G. Greiner, G. Wessels, and H. Feussner. Intra-operative Augmented Reality With Magnetic Navigation And Multi-texture Based Volume Rendering For Minimally Invasive Surgery. *Rechner- und Sensorgestützte Chirurgie*, pages 83–91, 2001.

[Sch01b] H. Schirmacher, C. Vogelgsang, H.-P. Seidel, and G. Greiner. Efficient free form light field rendering. In T. Ertl, B. Girod, G. Greiner, H. Niemann, and H.-P. Seidel, editors, *Proceedings of the Workshop Vision, Modeling, and Visualization (VMV)*, pages 249–256, Stuttgart, Germany, 2001. Infix, St. Augustin.

[Sch01c] A. Schneider, J. Hams, M. Scheuering, F. Härtl, H. Feussner, and G. Wessels. Navigation in der starren und flexiblen Endoskopie. *Biomedizinische Technik*, 46(1):420–421, 2001.

[Sch02a] C. H. Schick, T. Horbach, H. Weber, F. Vogt, G. Greiner, D. Paulus, and W. Hohenberger. Rechnergetütze Endoskopie des Bauchraums. In J. R. Siewert and W. Hartel, editors, *Digitale Revolution in der Chirurgie, 119. Deutscher Chirurgenkongress (DGCH) 2002*, page 964, Berlin, Germany, 2002. Springer Berlin, Heidelberg, New York.

[Sch02b] J. Schmidt, F. Vogt, and H. Niemann. Nonlinear Refinement of Camera Parameters using an Endoscopic Surgery Robot. In K. Ikeuchi, editor, *Proceedings of the Workshop on Machine Vision Applications (MVA)*, pages 40–43, Nara, Japan, 2002. IAPR MVA Organizing Committee.

[Sch03a] M. Scheuering. *Fusion of Medical Video Images and Tomographic Volumes*. PhD thesis, Institute of Computer Science, University Erlangen-Nuremberg, Germany, 2003.

[Sch03b] J. Schmidt, F. Vogt, and H. Niemann. Robust Hand-Eye Calibration of an Endoscopic Surgery Robot Using Dual Quaternions. In B. Michaelis and G. Krell, editors, *Proceedings of the Conference on Pattern Recognition, 25th DAGM Symposium*, pages 548–556, Magdeburg, Germany, 2003. Springer, Berlin, Heidelberg, New York.

[Sch04a] J. Schmidt, F. Vogt, and H. Niemann. Vector Quantization Based Data Selection for Hand-Eye Calibration. In B. Girod, M. Magnor, and H.-P. Seidel, editors, *Proceedings of the Workshop Vision, Modeling, and Visualization (VMV)*, pages 21–28, Stanford, USA, 2004. Infix, St. Augustin.

[Sch04b] I. Scholz, J. Denzler, and H. Niemann. Calibration of Real Scenes for the Reconstruction of Dynamic Light Fields. *IEICE Transactions on Information and Systems*, E87-D(1):42–49, 2004.

[Sch05] I. Scholz, C. Vogelgsang, J. Denzler, and H. Niemann. Dynamic Light Field Reconstruction and Rendering for Multiple Moving Objects. In K. Ikeuchi, editor, *Proceedings of the Workshop on Machine Vision Applications (MVA)*, pages 184–188, Tsukuba Science City, Japan, 2005. IAPR MVA Organizing Committee.

[Sch06] J. Schmidt. *3-D Reconstruction and Stereo Self-Calibration for Augmented Reality*. PhD thesis, Institute of Computer Science, University Erlangen-Nuremberg, Germany, 2006. to appear.

[See05] SeeReal Technologies GmbH. http://www.seereal.com, last visited January 2005.

[Sel00] M. Selig, H. Fischer, L. Gumb, H. Schäf, R. Ullrich, B. Vogel, R. Cichon, M. Cornelius, U. Kappert, and S. Schüle. Minimal invasive Herzchirurgie. *Nachrichten - Forschungszentrum Karlsruhe*, 32(1-2):55–60, 2000.

[Sha85] S. A. Shafer. Using Color to Separate Reflection Components. *COLOR research and application*, 10(4):210–218, 1985.

[Sha99] G. C. Sharp, S. W. Lee, and D. K. Wehe. Invariant Features and the Registration of Rigid Bodies. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 932–937, Detroit, USA, 1999. IEEE Computer Society Press, Los Alamitos.

[She97] J. Shewchuk. *Delaunay Refinement Mesh Generation.* PhD thesis, School of Computer Science, Carnegie Mellon University, USA, 1997. Technical Report CMU-CS-97-137.

[She02] J. Shewchuk. Delaunay Refinement Algorithms for Triangular Mesh Generation. *Computational Geometry: Theory and Applications*, 22(1-3):86–95, 2002.

[Shi89] Y. C. Shiu and S. Ahmad. Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX = XB. *IEEE Transactions on Robotics and Automation*, 5(3):16–29, 1989.

[Shi94] J. Shi and C. Tomasi. Good Features to Track. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 593–600, Seattle, USA, 1994. IEEE Computer Society Press, Los Alamitos.

[Sla05] G. Slabaugh. Computing Euler Angles from a Rotation Matrix. http://home.comcast.net/~greg_slabaugh/publications/euler.pdf, last visited June 2005.

[Sob04] J. Sobotta. *Anatomie des Menschen.* Urban & Fischer, München, 21. edition, 2004.

[Sop92] N. J. Soper, P. T. Stockmann, D. L. Dunnegan, and S. W. Ashley. Laparoscopic cholecystectomy. The new gold standard? *Arch Surg*, 127(8):921–923, 1992.

[Stö00] M. Störring, E. Granum, and H. J. Andersen. Estimation of the illuminant color using highlights from human skin. In *Proceedings of the First International Conference on Color in Graphics and Image Processing (CGIP)*, pages 45–50, St. Etienne, France, 2000. Cepadues editions, Toulouse.

[Sto05] Karl Storz GmbH & Co. KG. http://www.karlstorz.com, last visited January 2005.

[Suz85] S. Suzuki and K. Abe. Topological Structural Analysis of Digitized Binary Images by Border Following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.

[Sze93] R. Szeliski. Rapid octree construction from image sequences. *CVGIP: Image Understanding*, 58(1):23–32, 1993.

[Tho02] T. Thormählen, H. Broszio, and P. N. Meier. Automatische 3D-Rekonstruktion aus endoskopischen Bildfolgen. In M. Meiler, H. Handels, F. Kruggel, T. Lehmann, and

D. Saupe, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 207–210, Leipzig, Germany, 2002. Springer, Berlin, Heidelberg, New York.

[Tom91] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, USA, 1991.

[Tra04] J. Traub, M. Feuerstein, M. Bauer, E. U. Schirmbeck, H. Najafi, R. Bauernschmitt, and G. Klinker. Augmented Reality for Port Placement and Navigation in Robotically Assisted Minimally Invasive Cardiovascular Surgery. In H. U. Lempke, K. Inamura, K. Doi, M. W. Vannier, A. G. Farman, and J. H. C. Reiber, editors, *Computer Assisted Radiology and Surgery (CARS), Proceedings of the 18th International Congress and Exhibition*, pages 735–740, Chicago, USA, 2004. Elsevier, Amsterdam.

[Tre97] L. Trefethen and D. Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM) Publications, Philadephia, 1997.

[Tru98] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, New York, 1998.

[Tru99] E. Trucco, A. Fusiello, and V. Roberto. Robust Motion and Correspondence of Noisy 3-D Point Sets with Missing Data. *Pattern Recognition Letters*, 20(8):889–898, 1999.

[Tsa87] R. Y. Tsai. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology using Off-the-Shelf TV Cameras and Lenses. *IEEE Journal of Robotics and Automation*, Ra-3(3):323–344, 1987.

[Tsa89] R. Y. Tsai and R. K. Lenz. A New Technique for Fully Autonomous and Efficient 3D Robotics Hand/Eye Calibration. *IEEE Transactions on Robotics and Automation*, 5(3):345–358, 1989.

[Ull79] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, 1979.

[Vog01a] F. Vogt, C. Klimowicz, D. Paulus, W. Hohenberger, H. Niemann, and C. H. Schick. Bildverarbeitung in der Endoskopie des Bauchraums. In H. Handels, A. Horsch, T. Lehmann, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 320–324, Lübeck, Germany, 2001. Springer, Berlin, Heidelberg, New York.

[Vog01b] F. Vogt, D. Paulus, and C. H. Schick. Fast Implementations of Temporal Color Image Filtering. In D. Paulus and J. Denzler, editors, *Proceedings of the Workshop Farbbildverarbeitung*, pages 89–98, Erlangen, Germany, 2001. Universität Erlangen-Nürnberg, Institut für Informatik. Arbeitsberichte des Instituts für Informatik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Band 34, Nr. 15.

[Vog02a] F. Vogt, D. Paulus, B. Heigl, C. Vogelgsang, H. Niemann, G. Greiner, and C. H. Schick. Making the Invisible Visible: Highlight Substitution by Color Light Fields. In *Proceedings of the First European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*, pages 352–357, Poitiers, France, 2002. IS&T — The Society for Imaging Science and Technology, Springfield.

[Vog02b] F. Vogt, D. Paulus, and H. Niemann. Highlight Substitution in Light Fields. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 637–640, Rochester, USA, 2002. IEEE Computer Society Press, Los Alamitos.

[Vog02c] F. Vogt, D. Paulus, I. Scholz, H. Niemann, and C. H. Schick. Glanzlichtsubstitution durch Lichtfelder. In M. Meiler, H. Handels, F. Kruggel, T. Lehmann, and D. Saupe, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 103–106, Leipzig, Germany, 2002. Springer, Berlin, Heidelberg, New York.

[Vog03a] F. Vogt, S. Krüger, H. Niemann, and C. H. Schick. A System for Real-Time Endoscopic Image Enhancement. In R. E. Ellis and T. M. Peters, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 356–363, Montreal, Canada, 2003. Springer, Berlin, Heidelberg, New York.

[Vog03b] F. Vogt, S. Krüger, D. Paulus, H. Niemann, W. Hohenberger, and C. H. Schick. Endoskopische Lichtfelder mit einem kameraführenden Roboter. In T. Wittenberg, P. Hastreiter, U. Hoppe, H. Handels, A. Horsch, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 418–422, Erlangen, Germany, 2003. Springer, Berlin, Heidelberg, New York.

[Vog04a] F. Vogt, S. Krüger, J. Schmidt, D. Paulus, H. Niemann, W. Hohenberger, and C. H. Schick. Light Fields for Minimal Invasive Surgery Using an Endoscope Positioning Robot. *Methods of Information in Medcine*, 43(4):403–408, 2004.

[Vog04b] F. Vogt, S. Krüger, T. Zinßer, T. Maier, H. Niemann, W. Hohenberger, and C. H. Schick. Fusion von Lichtfeldern und CT-Daten für minimal-invasive Operationen. In T. Tolxdorff, J. Braun, H. Handels, A. Horsch, and H.-P. Meinzer, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 309–313, Berlin, Germany, 2004. Springer, Berlin, Heidelberg, New York.

[Vog04c] S. Vogt, A. Khamene, H. Niemann, and F. Sauer. An AR System with Intuitive User Interface for Manipulation and Visualization of 3D Medical Data. In J. D. Westwood, H. M. Hoffmann, G. T. Mogel, D. Stredney, and R. A. Robb, editors, *Proceedings of the 12th Annual Medicine Meets Virtual Reality Conference (MMVR)*, pages 397–403, Newport Beach, USA, 2004. IOS Press, Amsterdam.

[Vog04d] S. Vogt, F. Wacker, A. Khamene, D. R. Elgort, T. Sielhorst, H. Niemann, J. Duerk, J. Lewin, and F. Sauer. Augmented Reality System for MR-guided Interventions: Phantom Studies and First Animal Test. In R. L. Galloway, editor, *Proceedings of SPIE's Conference of Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display*, pages 100–109, San Diego, USA, 2004. SPIE.

[Vog05a] C. Vogelgsang. *The lgf3 Project: A Versatile Implementation Framework for Image-Based Modeling and Rendering*. PhD thesis, Institute of Computer Science, University Erlangen-Nuremberg, Germany, 2005.

[Vog05b] F. Vogt, S. Krüger, M. Winter, H. Niemann, W. Hohenberger, G. Greiner, and C. H. Schick. Erweiterte Realität und 3-D Visualisierung für minimal-invasive Operationen durch Einsatz eines optischen Trackingsystems. In H.-P. Meinzer, H. Handels, A. Horsch, and T. Tolxdorff, editors, *Proceedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, pages 217–221, Heidelberg, Germany, 2005. Springer, Berlin, Heidelberg, New York.

[Wag02] A. Wagner, K. Schicho, W. Birkfellner, M. Figl, R. Seemann, F. König, F. Kainberger, and R. Ewers. Quantitative analysis of factors affecting intraoperative precision and stability of optoelectronic and electromagnetic tracking systems. *Med. Phys.*, 29(5):905–912, 2002.

[Wan92] C. Wang. Extrinsic Calibration of a Vision Sensor Mounted on a Robot. *IEEE Transactions on Robotics and Automation*, 8(2):161–175, 1992.

[Wan01] P. Wang, S. M. Krishnan, C. Kugean, and M. P. Tjoa. Classification of Endoscopic Images Based on Texture and Neural Network. In *Proceedings of the 23rd Annual EMBS International Conference*, pages 3691–3695, Istanbul, Turkey, 2001. IEEE Computer Society Press, Los Alamitos.

[Wan02] Y. Wang, J. Ostermann, and Y.-Q. Zhang. *Video Processing and Communications*. Prentice Hall, New Jersey, 2002.

[Wen03] M. Wendt, F. Sauer, A. Khamene, B. Bascle, S. Vogt, and F. K. Wacker. A Head-Mounted Display System for Augmented Reality: Initial Evaluation for Interventional MRI. *Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 175(3):418–421, 2003.

[Wes99] G. Wessels, H. Feussner, H. Allescher, R. Graumann, and G. Herold. Interdisziplinärer gastroenterologisch-chirurgischer Arbeitsplatz. *electromedica*, 67(2):1–7, 1999.

[Wes04] J. B. West and C. R. Maurer. Designing Optically Tracked Instruments for Image-Guided Surgery. *IEEE Transactions on Medical Imaging*, 23(5):533–545, 2004.

[Wil45] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[Wil02] B. Wilburn, M. Smulski, H.-H. K. Lee, and M. Horowitz. The Light Field Video Camera. In S. Panchanathan, V. M. Bove, and S. I. Sudharsanan, editors, *Proceedings of the Conference Media Processors, SPIE Electronic Imaging*, San Jose, USA, 2002. SPIE.

[Win05] M. Winter, G. Greiner, F. Vogt, H. Niemann, and S. Krüger. Visualizing distances between light field and geometry using projective texture mapping. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, editors, *Proceedings of the Workshop Vision, Modeling, and Visualization (VMV)*, pages 257–264, Erlangen, Germany, 2005. Infix, St. Augustin.

[Wu03] R. Wu, K. V. Ling, W. Shao, and W. S. Ng. Registration of Organ Surface with Intra-operative 3D Ultrasound Image Using Genetic Algorithm. In R. E. Ellis and T. M. Peters, editors, *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 383–390, Montreal, Canada, 2003. Springer, Berlin, Heidelberg, New York.

[Yam02] S. M. Yamany and A. A. Farag. Surface Signatures: An Orientation Independent Free-Form Surface Representation Scheme for the Purpose of Objects Registration and Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1105–1120, 2002.

[Zha96] Z. Zhang. On the Epipolar Geometry Between Two Images With Lens Distortion. In *Proceedings of the International Conference Pattern Recognition (ICPR)*, pages 407–411, Vienna, Austria, 1996. IEEE Computer Society Press, Los Alamitos.

[Zha98] Z. Zhang. A Flexible New Technique for Camera Calibration. Technical Report MSR-TR-98-71, Microsoft Research, 1998.

[Zha99] Z. Zhang. Flexible Camera Calibration By Viewing a Plane From Unknown Orientations. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 666–673, Corfu, Greece, 1999. IEEE Computer Society Press, Los Alamitos.

[Zha00] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[Zha02] X. Zhang and S. Payandeh. Application of Visual Tracking for Robot-Assisted Laparoscopic Surgery. *Journal of Robotics Systems*, 19(7):315–328, 2002.

[Zim02] M. Zimmermann, R. Krishnan, A. Raabe, and V. Seifert. Robot-assisted navigated neuroendoscopy. *Neurosurgery*, 51(6):1451–1452, 2002.

[Zin03] T. Zinßer, J. Schmidt, and H. Niemann. A Refined ICP Algorithm for Robust 3-D Correspondence Estimation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, 2003. IEEE Computer Society Press, Los Alamitos.

[Zin04] T. Zinßer, C. Gräßl, and H. Niemann. Efficient Feature Tracking for Long Video Sequences. In C. E. Rasmussen, H. H. Bülthoff, M. A. Giese, and B. Schölkopf, editors, *Proceedings of the Conference on Pattern Recognition, 26th DAGM Symposium*, pages 326–333, Tübingen, Germany, 2004. Springer, Berlin, Heidelberg, New York.

Index

2-D Delaunay triangulation, 114
3-D endoscope, 26
3-D monitor, 5
3-D triangular mesh, 114
3-D visualization, 25

adapter, 103
Advanced Realtime Tracking GmbH, 101
AESOP, 18, 94
AESOP's kinematics, 95
Amira, 128
anatomical structures, 122
artificial patient, 137
ARTtrack2 camera, 101
augmented reality, 121
augmented reality (AR), 5, 24
AURORA, 27

back-projection, 38
  error, 51, 65

calibration kit, 102
calibration matrix, 37
calibration pattern, 61, 86
calibration points, 62
camera calibration, 61
camera parameter, 36
circularity criterion, 62
codebook vectors, 110
color cluster rotation, 68
color gradients, 79
color normalization, 68
computer assisted endoscopic surgery, 11
confidence map, 38
  computation, 111
conventional surgery, 12
cropping, 87

data selection, 108
DaVinci, 19
de-interlacing, 84
depth, 38
depth map, 38
  computation, 111
di-chromatic reflectance model, 78
distortion correction, 57, 86
distortion model, 59
double blind setup, 140

effective focal length, 36
EndoAssist, 19
endoscope
  flexible, 14
  rigid, 14
  side view, 75
endoscope adapter, 103
endoscopic surgery, 1
EvaMedIm, 140
exhaustive search, 109

factorization, 51
fiducial markers, 121
focal length, 36
forward kinematic transform, 96
freedom of movement, 3
fusion, 121, 133

ghosting artefacts, 42
ground truth, 142

hand-eye calibration, 97, **106**
    guidelines, 108
hand-eye coordination, 3
head-mounted display (HMD), 5
highlight detection, 78
homogeneous coordinates, 37, 207
homogeneous vector, 207
horizon, 75

illumination correction, 68
image degradations, **2**, 4
image distortion, 59
image enhancement, 20, 55
Image Geometry Transformations, 75
image modalities, 6
image rotation, 75
image skew, 36
image-based modeling, 5
imaging technologies
    advantages and disadvantages, 8
    computer tomography (CT), 6
    magnetic resonance imaging (MRI), 7
    positron emission tomography (PET), 7
interlacing, 55, 84
intrinsic camera parameters, 87
iterative-closest-point (ICP) algorithm, 131

keyhole, 1
keyhole surgery, 1
kinematics, 95

laparoscopic cholecystectomy, 15
LBG algorithm, 110
least median squares (LMedS), 112
lgf3, 117
Lichtleiter, 12
light field, 5, 35
    DC, 38
    definition and concept, 33
    dynamic, 53
    dynamic, DC, 53
    free form, 43
    reconstruction, 48
    two-plane (PP), 40
    visualization, 39, 117
light slab, 40
limited vision, **2**, 5
liver/gall bladder model, 137

MacBeth color checker, 68
Mann-Whitney U test, 141
mapping via planes, 43
marching cubes, 127
marker, 104
master-slave manipulator, 17
mechanical calibration, 48
MeVis, 125
MeVisLab, 125
MINIBIRD, 27
minimally invasive operation, 1
minimally invasive surgery, 1, **14**
mixed reality, 24

multi-modality visualization, 133

notch, 99
null-space of a matrix, 210

orthogonalization, 211
outlier detection, 50

PAL, 55
palpatation, 3
Picky ICP algorithm, 131
pinhole camera model, 36
plenoptic function, 33
  DC, 39
  discrete, 40
  reconstructed, 40
point tracking, 50
POLARIS, 28
pose determination, 27
preprocessing, 84
principal point, 36
projection matrix, 37
projective 2-D mapping, 63

range, 38
rank of a matrix, 210
region of interest, 75
registration, 121
  coarse, 130
  fine, 131
  intrinsic, 121
  transformation parameters, 129
registration and fusion, 29
robot, 17, 94
robot assisted interventions, 17

selection of 3-D points, 126
self-calibration, 52
sense of touch, 3
sensor noise, 56
Shi-Tomasi-Kanade tracker, 91
significance level, 141
singular value decomposition (SVD), 209
smARTtrack1, 28, 101
static scene assumption, 34
stereo endoscope, 26
stereoscopic depth perception, **3**, 5
structure matrix, 50
structure-from-motion, 50, 90
sum of squared differences (SSD), 91
synchronization, 102

target, 27, 102
  design, 103
temporal color median filter, 71
temporal filtering, 71
temporal noise, 71
trail, 165
transparent medium assumption, 35
trauma, 1
trocar, 14

unstructured lumigraph rendering, 39, 46

vector median filter, 72
vector quantization, 110
vertically interleaved, 117
video-endoscopic system, 55
virtual data, 28
virtual reality (VR), 24
VOXEL-MAN, 125

Wilcoxon test, 141

X-rays, 6

ZEUS, 19
zoom, 75

Curriculum Vitae

3rd May 1975	Date of birth (place of birth: Spaichingen)
Sep 1981 – Jul 1985	Grundschule Nendingen
Sep 1985 – Jul 1994	Otto-Hand Gymnasium Tuttlingen (Allgemeine Hochschulreife)
Sep 1994 – Sep 1995	Social Service at the University Clinic Tübingen
Oct 1995 – Sep 2000	Study of Computer Science at the University of Ulm (Diplom-Informatiker (Univ.))
Oct 2000 – Oct 2005	Researcher at the Computer Science Department 5 of the Friedrich-Alexander-University Erlangen-Nuremberg
since Nov 2005	Employed at Siemens Medical Solutions, Forchheim