# Stereo Vision-Guided Laser Microsurgery

Von der Fakultät für Maschinenbau
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades
Doktor-Ingenieur
genehmigte Dissertation

von

**Dipl.-Ing. Andreas Schoob**

**2018**

1. Referent: Prof. Dr.-Ing. Tobias Ortmaier
2. Referent: Prof. Dr.-Ing. Eduard Reithmeier

Tag der Promotion: 23. Oktober 2018

# Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Mechatronische Systeme der Gottfried Wilhelm Leibniz Universität Hannover.

Besonderer Dank gilt meinem Doktorvater und Institutsleiter, Herrn Professor Tobias Ortmaier, für die persönliche und fachliche Untersützung sowie die Möglichkeit an seinem Institut forschen und lehren zu dürfen. Rückblickend bin ich sehr dankbar für die Erfahrungen, die ich im Rahmen des internationalen Forschungsprojektes µRALP sammeln konnte. Weiterhin danke ich Herrn Professor Eduard Reithmeier für die Anfertigung des Zweitgutachtens und Frau Professorin Annika Raatz für die freundliche Übernahme des Prüfungsvorsitzes.

Meinen ehemaligen Kolleginnen und Kollegen möchte ich ganz herzlich für unsere Zusammenarbeit und die tolle Arbeitsatmosphäre am Institut danken. Besonderer Dank gebührt hierbei meinem Forschungsgruppenleiter Lüder Alexander Kahrs und meinem Bürokollegen Dennis Kundrat. Als Team konnten wir das µRALP–Projekt zu einem äußerst erfolgreichen Abschluss führen. Lüder danke ich sehr für die fachliche Unterstützung und das mir stets entgegengebrachte Vertrauen in meine Forschungsarbeit. Dennis möchte ich besonders für die tolle Zusammenarbeit, das freundschaftliche Verhältnis und die äußerst gründliche Korrektur dieser Arbeit danken. Seine Unterstützung bei den zahlreichen Laserversuchen ist hierbei besonders hervorzuheben. Er wusste immer die optimale Einstellung der Laserparameter – *Hauptsache ungerade!*

Großer Dank sei zudem an meine Studentinnen und Studenten gerichtet, von denen ich glücklicherweise Jan Bergmeier und Max-Heinrich Laves nach ihrer Abschlussarbeit als neue Kollegen am Institut begrüßen durfte. Weiterhin seien gebührend erwähnt: Florian Podszus, Lukas Kleingrothe, Stefan Lekon, Mariana Guerra M. Garcia, Daniel P. Coelho und Thomas Piskon – vielen Dank für eure äußerst wertvollen Beiträge im Rahmen dieser Arbeit.

Besonderer Dank gilt meiner Familie und meinen Freunden für die Unterstützung während dieser Arbeit. Mein größter Dank gebührt allerdings meiner Freundin, Ulrike, für die tollen Momente während meiner Promotionszeit, ihren Rückhalt und vor allem ihre Geduld bei der Fertigstellung dieser Arbeit. Vielen Dank!

Hannover, November 2018 *Andreas Schoob*

# Kurzfassung

Die transorale Lasermikrochirurgie (TLM) ist ein klinisch etabliertes Therapieverfahren zur kontaktlosen, atraumatischen Behandlung von Karzinomen im Mund- und Rachenraum sowie des Larynx. Das Ziel eines derartigen Eingriffs ist die radikale Entfernung dieser Läsion. Die Wahl eines hinreichend großen Resektionsabstandes steht hierbei jedoch im Konflikt mit der maximalen Funktionserhaltung des betreffenden Organs. Der Schutz des umliegenden Gewebes bedarf einer hochgenauen Laserpositionierung, die eine langjährige Erfahrung des Chirurgen voraussetzt.

Die Vorteile der TLM sind vielfältig, jedoch gehen diese mit wesentlichen Einschränkungen einher, die das postoperative Ergebnis und damit die Lebensqualität des Patienten beeinträchtigen. So liegen die größten Herausforderungen des Eingriff nicht nur in der eingeschränkten Sicht und dem erreichbaren Arbeitsraum, sondern auch in einer ungenauen Laserfokuspositionierung, einer unzureichenden Bildgebung der subepithelialen Tumorgrenzen, einer suboptimalen Schnittplanung und einem Genauigkeitsverlust der Ablation durch Weichgewebsbewegungen.

Die vorliegende Arbeit adressiert diese Einschränkungen durch einen neuartigen Ansatz der stereobildgestützten Lasermikrochirurgie mit Anwendung im Larynx. Obwohl zahlreiche Lasertechnologien und Algorithmen aus der Computer Vision existieren, gibt es bis heute keine ganzheitliche Lösung zur Integration der Stereobildverarbeitung in der Laserchirurgie, insbesondere unter Berücksichtigung der Lasereigenschaften sowie Informationen zur Struktur und Bewegung des Gewebes.

Zunächst wird ein Echtzeitverfahren zur stereobasierten Oberflächenrekonstruktion vorgestellt. Die Rekonstruktionsgenauigkeit wird anhand von *in vivo* Bilddaten und verschiedenen Stereokamerasystemen evaluiert. Anschließend wird ein Ansatz zur Laser-Kamera-Registrierung beschrieben, um so eine abstandsgeregelte Laserfokusnachführung für eine optimale Ablation zu realisieren. Weiterhin wird dem Livebild eine farbkodierte Distanzkarte überlagert, welche den Chirurgen bei der manuellen Fokuspositionierung unterstützt. Als weiterer Beitrag wird die Registrierung von Stereobilddaten und der optischen Kohärenztomographie vorgestellt, welche den Ausgangspunkt für die Detektion und Anzeige von Änderungen der laryngealen Submukosa darstellt.

Ferner werden die genannten Algorithmen in eine tabletbasierte Schnittstelle integriert. Geeignete Planungsstrategien werden abgeleitet und hinsichtlich Genauigkeit sowie Gebrauchstauglichkeit untersucht. Abschließend wird die Schnittstelle durch ein echtzeitfähiges Weichgewebsdeformationstracking ergänzt. Experimentelle Versuche weisen den positiven Beitrag einer Bildstabilisation während der Planung und einer Bewegungskompensation während der Ablation nach.

**Schlagwörter:** Lasermikrochirurgie, Stereo Vision, Multimodale Registrierung, Erweiterte Realität, Deformationstracking, Bewegungskompensation.

# Abstract

Transoral laser microsurgery (TLM) provides the most advanced microscopic technique for contactless, atraumatic treatment of oral, pharyngeal, and in especially laryngeal carcinomas. The primary goal of the laser ablation is radical removal of the lesion. This necessitates the selection of sufficiently large resection margins and thus conflicts with a further principle of surgery which is function preservation. Preserving as much healthy tissue as possible requires very accurate laser positioning that, as of today, demands a highly experienced surgeon.

Even though the advantages of TLM are manifold, there are substantial technical limitations compromising the post-operative outcome and hence the quality of life of the patient. Major challenges do not solely arise from a limited field of view and range of motion, but in particular from inaccurate laser focusing, inadequate imaging of the submucosal extent, suboptimal incision planning and loss of ablation accuracy due to soft tissue motion.

To overcome these limitations, this dissertation presents a novel approach for stereo vision-guided laser microsurgery with special emphasis on laryngeal interventions. A variety of laser technologies and computer vision algorithms exist; however, holistic integration of real-time stereoscopic image processing into soft tissue laser surgery, considering the laser characteristics on the one hand and surgical scene information, e.g., structure and non-rigid motion, on the other hand, has not been addressed in its entity thus far.

A computational method for stereo vision-based real-time surface estimation is developed as a prerequisite to multimodal registration. The reconstruction accuracy is assessed on an *in vivo* dataset and a variety of stereo imaging devices considered for microsurgery. A method for laser-to-camera registration is proposed facilitating distance-based laser focus adjustment and thus optimal ablation. To assist the manual focus repositioning process, color-encoded distance superimposed to the live view is implemented as part of the surgeon interface. A further contribution aims at the fusion of optical coherence tomography imaging with stereo vision. A registration and segmentation framework enables the detection and visualization of submucosal changes of laryngeal tissue.

Moreover, three-dimensional surface information and laser-to-camera registration are integrated into a stylus-tablet-based interface. Several path planning strategies are evaluated in terms of accuracy and usability. Finally, laser ablation in a dynamic soft tissue environment is addressed with an algorithm for non-rigid tracking. Experiments demonstrate the benefit of live view stabilization during incision planning and closed loop control for motion compensation during laser ablation.

**Keywords:** Laser microsurgery, stereo vision, multimodal registration, augmented reality, non-rigid tracking, motion compensation.

# Contents

**Vorwort** iii

**Kurzfassung** v

**Abstract** vii

**Nomenclature** xiii

**1 Introduction** 1
- 1.1 Transoral Laser Microsurgery of the Vocal Folds . . . . 4
- 1.2 Related Work . . . . 6
  - 1.2.1 Robot-Assisted and Endoscopic Laser Surgery . . . . 6
  - 1.2.2 Vision Guidance in Planning and Laser Control . . . . 8
  - 1.2.3 Vision-Based Scene Structure and Motion Estimation . . . . 9
  - 1.2.4 Optical Techniques for Laryngeal Tumor Imaging . . . . 14
- 1.3 Problems under Consideration . . . . 15
- 1.4 Components of the Research and Contributions to the µRALP Project . . . . 17
- 1.5 Outline of the Thesis . . . . 19

**2 Background** 21
- 2.1 Surgical Lasers . . . . 21
  - 2.1.1 Function Principle . . . . 21
  - 2.1.2 Laser-Tissue Interaction . . . . 22
- 2.2 Optical Coherence Tomography . . . . 23
- 2.3 Stereo Vision . . . . 24
  - 2.3.1 Coordinate Transformation . . . . 25
  - 2.3.2 Pinhole Camera Model . . . . 25
  - 2.3.3 Epipolar Geometry and Rectification . . . . 26
  - 2.3.4 Triangulation . . . . 27
- 2.4 Parallel Computing on the GPGPU . . . . 28
- 2.5 Surgical Laser Setup . . . . 29

**3 Online Estimation of Tissue Surface Structure** 31
- 3.1 Algorithmic Implementation and Evaluation . . . . 32
  - 3.1.1 Stereo Matching Algorithm . . . . 32

3.1.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.2 Reconstruction Accuracy of Stereo Imaging Devices for Microsurgery . . . . . . 42
3.2.1 Stereo Imaging Devices for Microsurgery . . . . . . . . . . . . . . . . 42
3.2.2 Accuracy Assessment . . . . . . . . . . . . . . . . . . . . . . . . . 43
3.2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
3.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

**4 Registration of Stereo Vision with Laser and OCT 53**
4.1 Tissue Surface Information for Distance-Based Laser Focus Adjustment . . . . . 54
4.1.1 Trifocal Model of Stereo Vision and Laser . . . . . . . . . . . . . . . . 54
4.1.2 Laser Ablation Experiments . . . . . . . . . . . . . . . . . . . . . . 57
4.1.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
4.2 Fusion of Stereo Vision and OCT . . . . . . . . . . . . . . . . . . . . . . 64
4.2.1 OCT Surface Segmentation and Registration . . . . . . . . . . . . . . 64
4.2.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
4.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68

**5 Development of a Visualization and Planning Interface 69**
5.1 Color-Encoded Distance for Interactive Laser Focus Positioning . . . . . . . . . 70
5.1.1 Visual Augmentation of the Live View . . . . . . . . . . . . . . . . . . 70
5.1.2 User Study Design . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
5.1.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
5.1.4 Transfer to an Endoscopic Scenario . . . . . . . . . . . . . . . . . . 75
5.2 OCT Imaging and Visualization of Subepithelial Tissue Structure . . . . . . . . . 79
5.2.1 Phantom Design and Segmentation in OCT Images . . . . . . . . . . . 79
5.2.2 Visualization of the Subepithelial Segmentation . . . . . . . . . . . . . 81
5.2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
5.3 Tablet-Based Strategies for Incision Planning in Laser Microsurgery . . . . . . . 85
5.3.1 Incision Planning Strategies . . . . . . . . . . . . . . . . . . . . . . 85
5.3.2 User Study Design . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
5.3.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
5.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90

**6 Motion Compensation in Planning and Laser Ablation 91**
6.1 Stereo Vision-Based Tracking of Soft Tissue Motion . . . . . . . . . . . . . . 92
6.1.1 Monoscopic Non-Rigid Motion Tracking . . . . . . . . . . . . . . . . . 92
6.1.2 Robust Stereoscopic Motion Tracking . . . . . . . . . . . . . . . . . . 95
6.1.3 Epipolar Constraint-Based Mesh Refinement . . . . . . . . . . . . . . 99
6.1.4 Considerations for Real-Time Implementation . . . . . . . . . . . . . . 103
6.1.5 Upsampling of the Motion Measurements . . . . . . . . . . . . . . . . 106

6.1.6 Performance Assessment on *In Vivo* Sequences . . . 108
6.1.7 Results . . . 109
6.2 Application to Image Stabilization during Incision Planning . . . 113
6.2.1 User Study Design . . . 114
6.2.2 Results . . . 115
6.3 Application to Motion Compensation during Laser Ablation . . . 116
6.3.1 Experimental Design . . . 116
6.3.2 Results . . . 120
6.4 Conclusion . . . 124

**7 Conclusion and Perspectives 127**

**Bibliography 129**

**A Reprint Permissions 151**

**B Curriculum Vitae 153**

# Nomenclature

## List of Latin Symbols

| | |
|---|---|
| $(\mathrm{CF})_\mathrm{L}$ | Coordinate frame of the left camera |
| ${}_{(\mathrm{L})}\boldsymbol{C}$ | Center point with respect to $(\mathrm{CF})_\mathrm{L}$ |
| $C(\boldsymbol{p},d)$ | Similarity-based stereo matching cost at image position $\boldsymbol{p}$ and disparity $d$ |
| $CT(\boldsymbol{p})$ | Census transform at image position $\boldsymbol{p}$ |
| $\{\boldsymbol{c},\boldsymbol{p}\}$ | Feature correspondence set of the non-rigid tracking |
| $d$ | Disparity defining the distance between the left and right image correspondence |
| $d_\mathrm{MHD}(\boldsymbol{p}_t)$ | Pairwise Mahalanobis distance of the non-rigid tracking evaluated at $\boldsymbol{p}_t$ |
| $\boldsymbol{d}_\mathrm{MHD}(t)$ | L1–norm of the Mahalanobis distances $d_\mathrm{MHD}(\boldsymbol{p}_t)$ of the mesh triangle $t$ |
| $D$ | Raw disparity map |
| $D_\mathrm{BF}$ | Smoothed disparity map obtained by bilateral filtering |
| $D_\mathrm{JBF}$ | Refined disparity map obtained by joint bilateral filtering |
| $\boldsymbol{D}_\mathrm{L}$ | Diagonal matrix of the weights in the left view during mesh refinement |
| $\boldsymbol{D}_\mathrm{R}$ | Diagonal matrix of the weights in the right view during mesh refinement |
| ${}^\mathrm{L}\boldsymbol{F}_\mathrm{R}$ | Fundamental matrix mapping image points from $(\mathrm{CF})_\mathrm{R}$ to $(\mathrm{CF})_\mathrm{L}$ |
| $\boldsymbol{F}_{\tau-1}$ | State transition matrix of the Kalman filter |
| $H(\boldsymbol{p},d)$ | Hamming distance of stereo matching at image position $\boldsymbol{p}$ and disparity $d$ |
| ${}^\mathrm{L}\boldsymbol{H}_\mathrm{G}$ | Homography defining the plane-to-plane projection from $(\mathrm{CF})_\mathrm{G}$ to $(\mathrm{CF})_\mathrm{L}$ |
| $\boldsymbol{H}$ | Hessian matrix used in the monoscopic tracking |
| $\boldsymbol{H}_\mathrm{L}$ | Hessian matrix of the left view used in the stereoscopic tracking |
| $\boldsymbol{H}_\mathrm{MR}$ | Hessian matrix of the mesh refinement |
| $\boldsymbol{H}_\mathrm{R}$ | Hessian matrix of the right view used in the stereoscopic tracking |
| $\boldsymbol{H}_\tau$ | Measurement matrix of the Kalman filter |
| $I(\boldsymbol{p})$ | Grayscale image evaluated at image position $\boldsymbol{p}$ |
| $\boldsymbol{I}(\boldsymbol{p})$ | Color image evaluated at image position $\boldsymbol{p}$ |
| $\boldsymbol{I}_\mathrm{R}(\boldsymbol{p})$ | Rank transform of color image $\boldsymbol{I}$ evaluated at image position $\boldsymbol{p}$ |
| $\Delta\boldsymbol{I}_\mathrm{S,L}$ | Stacked image residuals of the mesh refinement in the left view |
| $\Delta\boldsymbol{I}_\mathrm{S,R}$ | Stacked image residuals of the mesh refinement in the right view |
| $\boldsymbol{J}_\mathrm{L}$ | Left Jacobian of the mesh refinement |
| $\boldsymbol{J}_\mathrm{R}$ | Right Jacobian of the mesh refinement |
| $\boldsymbol{J}_\mathrm{S,L}$ | Stacked left Jacobian of the mesh refinement |
| $\boldsymbol{J}_\mathrm{S,R}$ | Stacked right Jacobian of the mesh refinement |
| $\boldsymbol{K}_\mathrm{L}$ | Left camera matrix |

| | |
|---|---|
| $\boldsymbol{K}_{\mathrm{R}}$ | Right camera matrix |
| $\boldsymbol{K}_{\tau}$ | Kalman gain |
| $\boldsymbol{M}_{\mathrm{L}}$ | Left camera projection matrix |
| $\boldsymbol{M}_{\mathrm{R}}$ | Right camera projection matrix |
| ${}_{(\mathrm{L})}\boldsymbol{n}$ | Normal vector with respect to $(\mathrm{CF})_{\mathrm{L}}$ |
| $\boldsymbol{p}_{\mathrm{L}}$, | Image position associated with $(\mathrm{CF})_{\mathrm{L}}$ |
| $\tilde{\boldsymbol{p}}_{\mathrm{L}}$ | Homogeneous image position associated with $(\mathrm{CF})_{\mathrm{L}}$ |
| ${}_{(\mathrm{L})}\boldsymbol{P}$ | 3D position with respect to $(\mathrm{CF})_{\mathrm{L}}$ |
| ${}_{(\mathrm{L})}\tilde{\boldsymbol{P}}$ | Homogeneous 3D position with respect to $(\mathrm{CF})_{\mathrm{L}}$ |
| $\boldsymbol{P}_{\tau}$ | State covariance of the Kalman filter |
| $\boldsymbol{q}$ | Parameter of the stereoscopic tracking |
| ${}_{(\mathrm{L})}\boldsymbol{Q}$ | Corner point with respect to $(\mathrm{CF})_{\mathrm{L}}$ |
| $\boldsymbol{Q}_{\tau-1}$ | Process noise covariance of the Kalman filter |
| ${}^{\mathrm{R}}\boldsymbol{R}_{\mathrm{L}}$ | Rotation matrix for transformation from $(\mathrm{CF})_{\mathrm{L}}$ to $(\mathrm{CF})_{\mathrm{R}}$ |
| $R_{\tau}$ | Measurement variance of the Kalman filter |
| $\{\boldsymbol{s}_j,\boldsymbol{s}_k,\boldsymbol{s}_l\}$ | Vertex positions defining a triangle of the deformable mesh $\boldsymbol{S}$ |
| $\boldsymbol{S}$ | Deformable triangular mesh in the mono view |
| $\boldsymbol{S}_{\mathrm{L}}$ | Deformable triangular mesh in the left view |
| $\boldsymbol{S}_{\mathrm{R}}$ | Deformable triangular mesh in the right view |
| ${}_{(\mathrm{R})}\boldsymbol{t}_{\mathrm{L}}$ | Translational vector of $(\mathrm{CF})_{\mathrm{L}}$ with respect to $(\mathrm{CF})_{\mathrm{R}}$ |
| ${}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{L}}$ | Homogeneous transformation from $(\mathrm{CF})_{\mathrm{L}}$ to $(\mathrm{CF})_{\mathrm{R}}$ |
| $(u,v)^{\mathrm{T}}$ | Image coordinate in horizontal and vertical direction |
| $\boldsymbol{W}(\boldsymbol{p},\boldsymbol{S})$ | Affine warp of image point $\boldsymbol{p}$ located in the deformable mesh $\boldsymbol{S}$ |

## List of Greek Symbols

| | |
|---|---|
| $\Delta\boldsymbol{q}$ | Parameter update increment of the stereoscopic tracking |
| $\Delta\boldsymbol{S}$ | Mesh update increment of the monoscopic tracking |
| $\varepsilon_{\mathrm{A}}$ | Appearance energy of the mesh refinement |
| $\varepsilon_{\mathrm{C}}$ | Correspondence energy of the stereoscopic tracking |
| $\varepsilon_{\mathrm{D}}$ | Deformation energy of the stereoscopic tracking and mesh refinement |
| $\varepsilon_{\mathrm{MR}}$ | Mesh refinement energy |
| $\zeta$ | Conditional function of the census and rank transform |
| $\boldsymbol{\xi}$ | Vector with the barycentric coordinates $(\xi_j,\xi_k,\xi_l)$ defined by vertices $(\boldsymbol{s}_j,\boldsymbol{s}_k,\boldsymbol{s}_l)$ |
| $\boldsymbol{\pi}$ | Plane definition used for the epipolar constraint and laser workspace border |
| $\rho(u)$ | Huber loss function |
| $\rho'(u)$ | Derivation of the Huber loss function |
| $\boldsymbol{\Sigma}_{\mathrm{C}}$ | Covariance matrix of center point $\boldsymbol{C}$ |
| $\boldsymbol{\Sigma}_{t}$ | Covariance matrix of triangle $t$ |

| | |
|---|---|
| $\Omega_{\mathrm{C}}$ | Spatial proximity considered for the stereo matching cost aggregation |
| $\Omega_{\mathrm{CT}}$ | Spatial proximity considered for the census transform |
| $\Omega_{\mathrm{d}}$ | Disparity search range |
| $\Omega_{\mathrm{R}}$ | Spatial proximity considered for the rank transform |
| $\Omega_{\mathrm{S}}$ | Spatial proximity considered for the appearance energy of the mesh refinement |

## List of Abbreviations

| | |
|---|---|
| ACC | Asymmetric Color Coding |
| AF | Autofocus |
| ASQ | After Scenario Questionnaire |
| CAD | Computer-Aided Design |
| CMM | Coordinate Measuring Machine |
| cMR(KF) | Concurrent Mesh Refinement (with Kalman Filtering) |
| CT | Computed Tomography |
| CUDA | Compute Unified Device Architecture |
| DLK | Deformable Lucas-Kanade |
| DOF | Depth of Field |
| DSI | Disparity Space Image |
| Er:YAG | Erbium-Doped Yttrium Aluminum Garnet Laser |
| EVT | *Ex Vivo* Tissue |
| FLE | Fiducial Localization Error |
| FPFH | Fast Point Feature Histograms |
| FRE | Fiducial Registration Error |
| GM | Gray Mask |
| GPGPU | General Purpose Computation on Graphics Processing Unit |
| GT | Ground Truth |
| HMA | Hierarchical Multi-Affine Toolbox |
| ICP | Iterative Closest Point |
| IVT | *In Vivo* Tissue |
| LV | Laser View |
| MC | Motion-Compensated |
| MDE | Maximum Distance Error |
| MDF | Medium-Density Fiberboard |
| MHD | Mahalanobis Distance |
| MIS | Minimally Invasive Surgery |
| MR | Mesh Refinement |
| MRI | Magnetic Resonance Imaging |
| MV | Mono View |

| | |
|---|---|
| NAF | No Autofocus |
| NBI | Narrow Band Imaging |
| NC | Non-Compensated |
| OCT | Optical Coherence Tomography |
| OpenGL | Open Graphics Library |
| PB | Point-Based |
| PCA | Principal Component Analysis |
| PD | Pen Display |
| PFN | Progressive Finite Newton |
| RANSAC | Random Sample Consensus |
| RMSE | Root Mean Square Error |
| RT | Real-Time |
| SCC | Symmetric Color Coding |
| SD | Standard Deviation |
| SEQ | Sequence |
| sMR | Sequential Mesh Refinement |
| SUS | System Usability Scale |
| SV | Stereo View |
| TLM | Transoral Lasermicrosurgery |
| TPS | Thin Plate Spline |
| WTA | Winner-Take-All |
| µRALP | Micro-Technologies & Systems for Robot-Assisted Laser Phonomicrosurgery |

# 1 Introduction

Cancer therapy is considered as one of the major challenges of society today. A significant percentage of cancer is associated to head and neck, especially in the larynx where surgical treatments are highly challenging and demanding for the clinician. Small imperfections significantly affect the post-operative quality of the patient's life. In particular, traditional open surgery of the upper aerodigestive tract involving the oral cavity, the pharynx, and the larynx can lead to postoperative functional disorders of voice formation, swallowing, or respiration. Thus, transoral laser microsurgery (TLM) has been established in recent years as a widely used technique for contact-less treatment of a variety of pathologies. Precise and atraumatic removal of the lesion is achieved at superior preservation of the anatomical integrity of the affected organ [SA00].

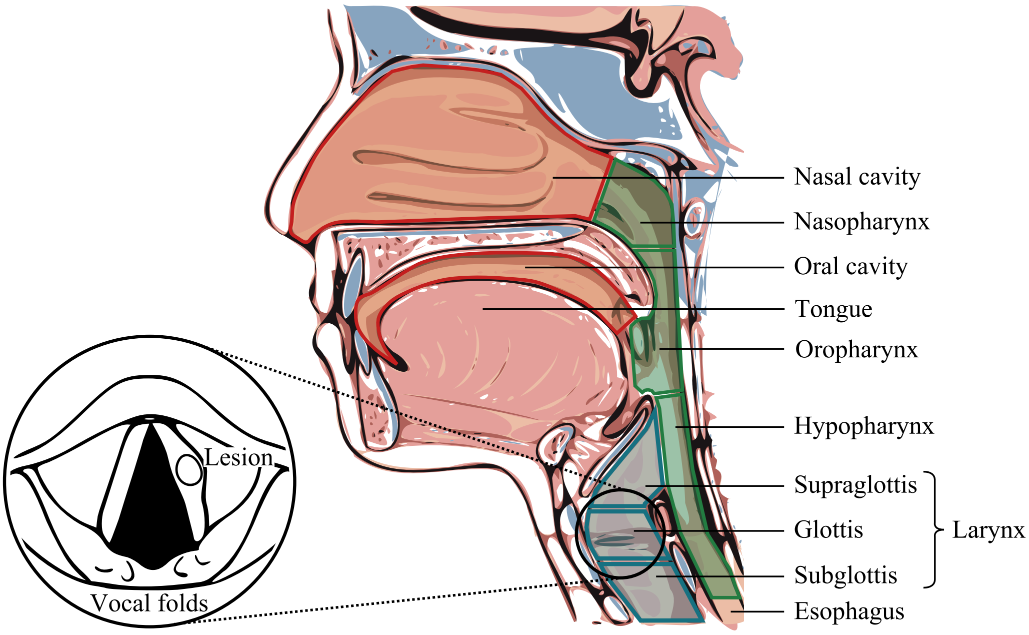


Figure 1.1: Anatomy of the upper aerodigestive tract.

As illustrated in Figure 1.1, the main anatomical structures of the upper aerodigestive tract include the nasal and oral cavity (red), the pharynx including nasopharynx, oropharynx, and hypopharinx (green), and the larynx (blue). The latter connects to the pharynx and is located within the anterior portion of the neck. It consists of three unpaired and three pairs of smaller cartilages, and many minuscule muscles [RS08]. Its spatial extent can be divided into supraglottic, glottic, and subglottic space (see Figure 1.1). The glottis houses the vocal folds that are complex microanatomical structures constituted by the squamous epithelium, the lamina propria with the vocal ligaments, and subjacent vocalis muscle fibers [RS08]. In the most primitive function, the larynx controls respiration and protects the lower airway against foreign objects by adduction reflex or coughing. More complex functions such as voice formation require coordinated breath support and simultaneous contraction of the laryngeal musculature, i.e., the vocalis muscle, in order to induce oscillations on the tensioned vocal ligaments and thus to generate sound.

The characteristics of pathological disorders in the upper aerodigestive tract differ between benign, premalignant, and malignant stage. In case of benign lesions (e.g. nodules, polyps, cyst), epithelium cells are not infiltrated. Initial changes are observed in case of premalignancies (e.g. dysplasia, atypia). In case of malignant lesions, i.e., the squamous cell carcinoma, that is the most prevalent entity among all tumors in the oral and nasal cavity, pharynx and larynx, the epithelium layer is significantly altered. Malignancies arising in other areas such as the sinonasal tract or the salivary glands are relatively uncommon [SW14].

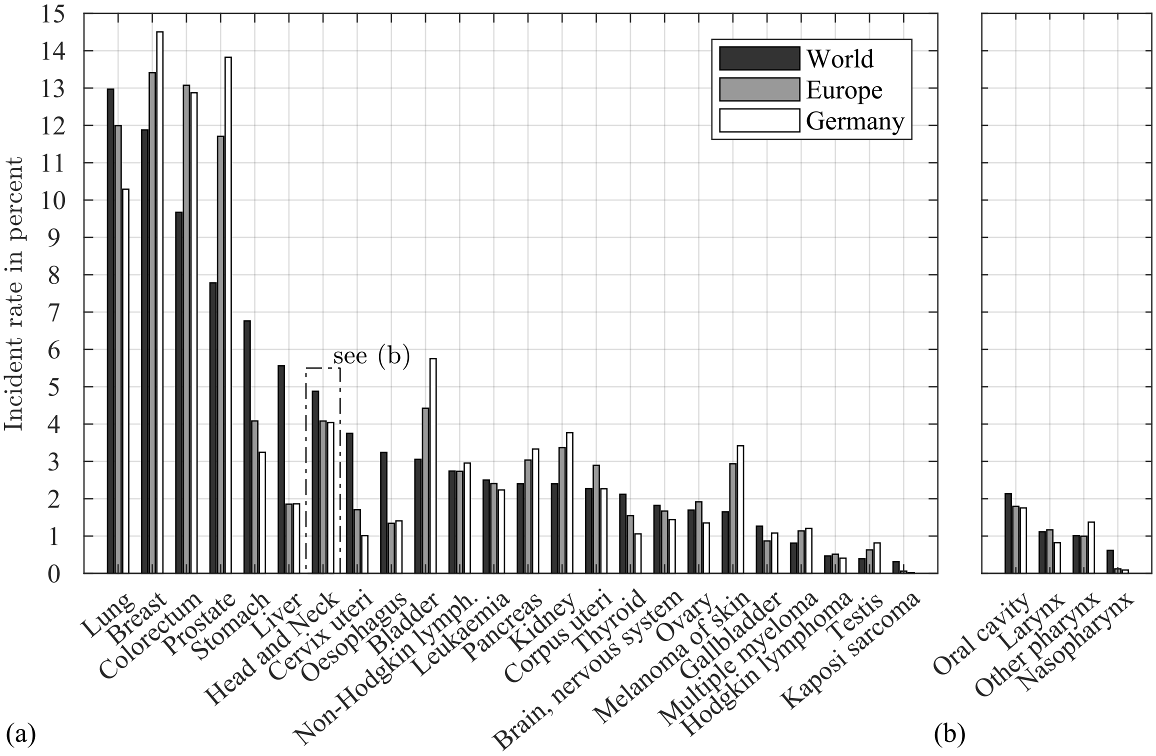


Figure 1.2: Estimated new cancer cases in 2012 with (a) the incident rate related to the total number of recorded cancer diseases and (b) the occurrences in head and neck only [FSD+15].

As the program Global Cancer Observatory of the International Agency for Research on Cancer reported for 184 countries worldwide in 2012 [FSE+13], head and neck squamous cell carcinoma is the seventh leading cancer in incidence worldwide with 686 000 cases next to lung, breast, colorectum, prostate, stomach, and liver cancer [FSD+15]. Figure 1.2a shows the estimated incident rate in relation to all new cancer cases in 2012. Head and neck malignancies can be further assigned to their local anatomy as depicted in Figure 1.2b. In total, 376 000 patients died due to cancer in the upper aerodigestive tract indicating a high mortality.

With regard to laryngeal cancer where glottic carcinoma is the most frequent entity compared with those in the supraglottic and subglottic space, the worldwide number of incidents in 2012 is estimated with 156 877 thereof 39 921 and 4 064 in Europe and Germany, respectively [FSFLT+13, FSD+15, PKW+16]. Surprisingly, the probability of developing cancer is more than seven times higher for male than for female [FSD+15].

The most significant causes of developing head and neck squamous cell carcinoma are mainly related to smoking and excessive alcohol use [LPK+09]. Regarding tobacco consumption, the probability of developing cancer drastically increases with the number of packs per day and years of smoking and has been found to be higher in high-income countries [SW14]. If combined with heavy alcohol consumption of several years, there is a synergistic effect leading to an increased multiplicative risk of malignancies in the upper airway [HBC+09]. Even though there is a falling trend at all ages and both sexes in most of the developed countries not least because of ongoing anti-smoking campaigns, an increase in head and neck squamous cell carcinoma related to infection with the human papillomavirus (HPV) has been found. This is supposed to be caused by changes in sexual habits, i.e., oral sex and sexual activity beginning in earlier ages as well as by the higher number of varying sexual partners nowadays [MDWF10, OEM+12].

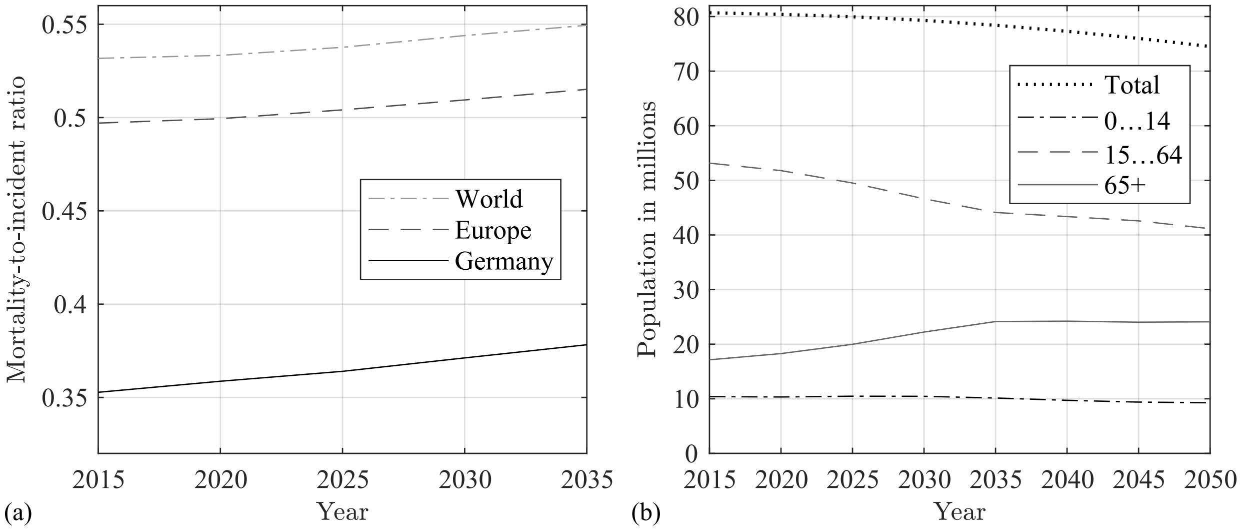


Figure 1.3: Prognosis of (a) mortality-to-incident ratio for larynx cancer of the years [FSE+13], and (b) expected demographic development in Germany by age in years [Nat13].

Considering the population forecast of the United Nations [Nat13], Figure 1.3a depicts the predicted mortality-to-incident ratio between the years 2015 and 2035 revealing the future burden of laryngeal cancer not only worldwide, but also in highly developed European countries and in particular with rising incidents in Germany. The increase in Germany is mainly associated with the demographic change of the society, as a result of higher life expectancy on the one hand and declined birth rate for the last decades on the other hand (see Figure 1.3b). Even though laryngeal cancer nowadays represents only 1.1 percent of all reported incidents, efforts have to be made for both laryngeal cancer awareness campaigns as well as for further advanced diagnostic and therapeutic systems.

Today, TLM represents the most recent surgical technique for organ preserving treatment of oral, pharyngeal, and in especially laryngeal pathologies. These are not solely limited to cancerous stages as underlined by the statistics shown in Figure 1.2. Premalignant entities as antecedents to squamous cell carcinoma and benign pathologies also demand for surgical treatment provided that physiology of swallowing, breathing, or voice formation is affected. To determine the therapeutic approach, the patient has to undergo a diagnostic work flow based on indirect or direct laryngoscopy as well as radiology in case of tumor spread is suspected. If preoperative imaging with computed tomography (CT) or magnetic resonance imaging (MRI) reveals deep infiltration of the tumor or nodal metastases, open surgery or radiotherapy in combination with cancer-targeted drug therapies are inevitable [SA00].

## 1.1 Transoral Laser Microsurgery of the Vocal Folds

If indication for transoral laser microsurgery (TLM) is given and the laryngeal space can be exposed adequately without any restrictions due to the patient's anatomy, the common work flow considers the lesion to be resected with a focused carbon dioxide ($CO_2$) laser (wavelength $\lambda_{CO_2} = 10.6\,\mu m$) coupled to a surgical microscope. Compared to cold instrument surgery, early studies demonstrated atraumatic and bloodless cutting with a $CO_2$ laser due to the optimal absorption property in soft tissue and the haemostatic effect of the coagulation [Jak72, Str75, VSJ78]. The patient does not require any blood transfusion or tracheotomy and perioperative complication rate is reduced to a minimum [SA00]. Even though instrumentation is expensive, TLM attains comparable cancer survival rates with improved outcomes in terms of voice quality, aesthetics (no scars) and shortened patient recovery as well as hospitalization time; thus, it leads to a better cost effectiveness ratio [HSG$^+$07, SRS$^+$12]. Numerous studies have shown that TLM of even advanced-stage diseases such as invasive squamous cell carcinoma constitutes a valuable alternative to open partial laryngectomy [RA11, CIM$^+$13, VBB$^+$16, DSN$^+$16, PPP$^+$16].

Referring to the surgical work flow, state-of-the-art TLM is performed using a microscope, a high-power $CO_2$ laser and a manually operated micromanipulator as shown in Figure 1.4. While being positioned supine with a strongly extended neck, the patient is anesthetized and ventilated by endotracheal intubation. The vocal folds are accessed by establishing a direct line of sight through a laryngoscope that is fastened on the patient's chest with a suspension arm (see Figure 1.4). The microscope is placed at a distance of usually 400 mm providing a magnified view onto the lesion. The experienced surgeon is able to locate the tumor and to determine its extent in order to derive an adequate surgical plan. If changes of the cellular structure cannot be identified clearly, a diagnostically more conclusive inspection is conducted with an endoscope fed through the laryngoscope to obtain a closer perspective and magnified view. A further expansion stage is provided by narrow band imaging (NBI) that enhances the visibility of mucosal vessel structures indicating characteristic changes in case of malignancies [PCDB+10].

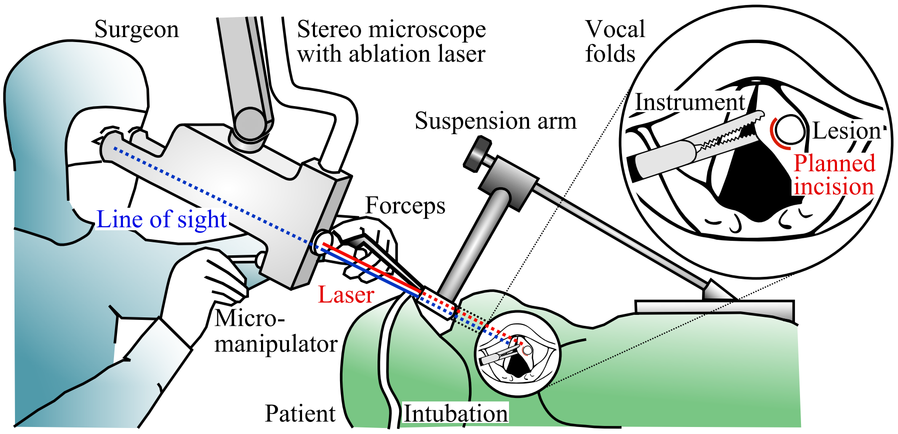


Figure 1.4: Surgical setup for transoral laser microsurgery (TLM).

Once the resection strategy is elaborated, the tumor is excised by manually deflecting the laser beam while executing laser ablation through a foot-switch. The tissue is laterally tensioned with grasping forceps to expose subjacent structures while simultaneously occluding the subglottic space and thus reducing the risk of hitting the tracheal mucosa [SA00]. However, associated hand-eye-coordination for both instrument handling and micromanipulator control is highly demanding and thus requires extensive training [ORJ+14]. To accurately trace the desired path while minimizing heat exposure to the tissue, semi-automated scanning micromanipulators are increasingly utilized for automatic, multi-pass execution of pre-programmed line and circular patterns. Noteworthy commercial systems are the Digital Acublade Scanning Micromanipulator (Lumenis Ltd., Yokneam, Israel), the SoftScan plus R (KLS Martin Group, Tuttlingen, Germany), or the HiScan Surgical (DEKA M.E.L.A. SRL, Calenzano, Italy) providing a scanning range up to several millimeters. Compared to manual beam guidance, operating time is not only shortened up to 30 percent, but also an evenly distributed coagulation thickness of only 10 µm can be achieved [RHC+05].

The primary concern of the laser-assisted ablation is radical dissection of the lesion. If the tumor extent is clearly visible under microscopic observation, it is resected until the transition to healthy tissue appears. Small tumors are commonly removed en bloc while larger entities are excised in pieces with oncologic safety [SA00]. Subsequently, histopathological analysis is carried out in order to assess if the margins are tumor-free. Radicality of the dissection has to be confirmed to preclude potential recurrence of the cancer disease; otherwise, the physician has to repeat the aforementioned steps of the surgical treatment.

Complete removal of laryngeal tumors necessitates selection of sufficiently large resection margins and thus conflicts with a further principle of surgery aiming at function preservation [SA00]. To minimize the amount of dissected non-malignant tissue, the laser has to be guided as close as possible to the tumor boundary. While resection margins of 5 mm are common in the oral cavity and oropharynx [HFBG+13, ACBP+13], less than a millimeter is targeted when resecting small glottic tumors [RA11]. Exceeding this distance can compromise the anatomical integrity of the vocal ligaments that are located only a few hundred micrometers below the topmost epithelium layer. As a consequence, precise and narrow cutting with minimal thermal damage is inevitable. Minimizing the irradiation time is not only achieved by aforementioned scanning technique, but also by applying a pulsed laser. Multiple high-power, short duration bursts are emitted per second whereas the interpulse time period ensures thermal relaxation of the tissue [RLND08].

If the scanning micromanipulator is combined with a microspot laser focusing unit such as the Micro Point 2 (KLS Martin Group, Tuttlingen, Germany), that provides a minimum focus diameter of 0.11 mm at high-power, the energy density threshold required for ablation is rapidly reached which significantly reduces thermal spread to the vocal fold ligaments [SCJ+11]. At the same time, incisions remain almost char-free with clearly recognizable tumor margins; thus, this technique facilitates an unobstructed view during the intervention and a distinct histological examination after surgery [SA00].

Due to the diversity of cutting and ablation tasks in TLM, the postoperative outcome is furthermore strongly affected by selected laser parameters including wavelength, power, spot size, exposure time and energy delivery mode. While the maximum output power can be very high, e.g., up to 20 W for cutting through large tumors [RA11], a laser power 3 W is usually sufficient to manage a variety of clinical cases [CMI+14].

## 1.2 Related Work

In the following, the state-of-the-art for laser-assisted soft tissue surgery as well as stereo vision guidance in minimally invasive surgery is discussed. This section reveals that a variety of laser technologies and computer vision methods exists in the field of medical applications and research.

### 1.2.1 Robot-Assisted and Endoscopic Laser Surgery

Since commercial, microscope-attached laser scanner can operate only within a few millimeters, recent research has been focused on the development of novel concepts for motorized beam deflection providing a scanning range of several centimeters and frequencies up to 200 Hz while maintaining micrometer precision [MDC11, DMC15b]. Results demonstrated not only superior performance in terms of usability, accuracy, and controllability, but also that novice surgeons can perform almost as good as experienced clinicians since required level of training is drastically reduced [BWD15].

Initial work in the field of robot-assisted laser surgery discusses a $CO_2$ laser laparoscope actuated with four degrees of freedom in order to maintain the trocar remote center of motion while precisely ablating tissue in the abdominal cavity [TVBR+03]. In regard to transoral robotic surgery (TORS), the da Vinci Surgical System (Intuitive Surgical, Inc, Sunnyvale, CA, USA) has become a valuable alternative to TLM, i.e., for oropharyngeal surgery, providing the clinician with wristed instrumentation, high-definition stereo vision, motion scaling, and tremor compensation [WOM+12]. A flexible, hollow core $CO_2$ laser fiber (OmniGuide, Inc., Lexington, MA, U.S.) attached to the instrument tip is deployed for contact-based tissue cutting by moving the robotic arm [SS07, DSJG08]. Successful tumor treatment in the upper aerodigestive tract is demonstrated in human cadaver and canine experiments as well as patient trials. However, setup time and costs as well as inadequate exposure of the narrow supraglottic space limit the widespread use compared with conventional TLM equipment.

The Flex® Robotic System (Medrobotics Corp., Raynham, MA, U.S.) shown in Figure 1.5a combines the advantages of flexible endoscopy and robotic assistance. The problem of rigid and straight instrumentation as in TLM is circumvented by providing a versatile tool that passively adapts to the anatomy of the larynx and thus mitigates the need of suspension laryngoscopy [RSJZ+12]. The system is equipped with two articulating instruments whereas one of them is deployed to guide a flexible $CO_2$ laser fiber for contact-based ablation. However, direct access to the vocal folds with the Flex® Robotic System system is still a challenging task due to the endotracheal tube [LMH+16].

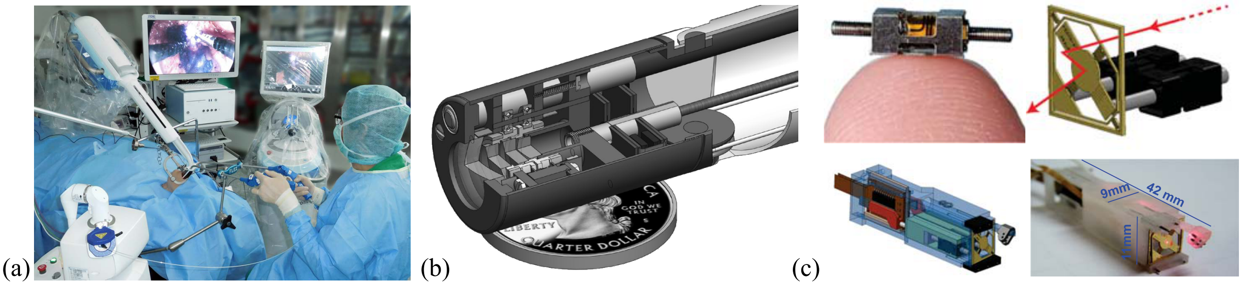


Figure 1.5: TLM with (a) the Flex® system with fiber laser and instruments (reprinted from [LMH+16], © 2016, with kind permission from John Wiley and Sons), (b) an endoscopic laser scalpel with an actuated Risley prism pair (reproduced from [PRK+12], © 2012, with kind permission from SPIE International Society for Optics and Photonics), and (c) an endoscopic laser scanner based on two linear piezoelectric motors (reprinted from [RTR+16], © 2016 IEEE).

As a consequence, endoscopic flexible access to the glottic space necessitates a high level of miniaturization of laser optics, vision system, and instruments. Since contact-based laser ablation dispense with optical focusing and distal scanning, the integration of an optical fiber into the endoscopic tip is straightforward [HHW14, RDM15, STL15]. If minimal thermal damage is targeted, contactless ablation with automatic and high-frequent scanning is inevitable. Various techniques are discussed in literature. For instance, the laser beam can initially be deflected by a proximal galvanometric scanner and then coupled into the optical path of a rigid, relay lens system-based endoscope yielding a distal scanning range of $(10.7 \times 9.7)\,\mathrm{mm}^2$ [YYM$^+$10].

Flexible laser endoscopy with distal beam scanning in a range of a few centimeters can either be facilitated with miniaturized piezoelectric drives that actuate a Risley prism pair (see Figure 1.5b) [PRK$^+$12] or with a deformable, two-axis silicon mirror operated at $40\,\mathrm{Hz}$ (see Figure 1.5c) [RTR$^+$16]. In contrast, ultrafast scanning at $300\,\mathrm{kHz}$ repetition rate in a $(150 \times 150)\,\mu\mathrm{m}^2$ workspace can be achieved with a piezoelectric tube actuator deflecting an embedded laser fiber [FYSBY14]. However, the most promising solution in terms of integrability to an endoscopic device is offered by micro-electromechanical systems (MEMS) combining a compact design with frequencies up to several kilohertz and enlarged scanning ranges [SKG$^+$14, PRRA16].

### 1.2.2 Vision Guidance in Planning and Laser Control

Once the laser system offers computerized scanning, vision-based planning and control can be established as prerequisites for enhanced surgeon-machine interfaces. In a first expansion stage, homography-based mapping between image and laser task space, assuming planarity of the tissue surface, facilitates incision planning in the surgical live video. An early development of intuitive stylus-tablet-based planning is dedicated to laser-assisted laparoscopy demonstrating that laser aiming accuracy is improved by making use of familiar hand writing skills [TBV$^+$06]. In the context of TLM, a virtual scalpel system based on a pen display provides superior precision, control, and usability compared with a state-of-the-art micromanipulator [MDB$^+$14]. In addition, safe zones can be defined in the live view yielding virtual fixtures in task space that protect risk structures from unintended laser movements [DOCM14].

To overcome inaccuracies of the homography-based open-loop control, i.e. at non-planar tissue surface, recent research in laser surgery has been focused on visual servoing. Quasi closed-loop control is achieved by a two-step strategy adding a vision-based correction phase to fast open-loop laser positioning [DMC15a]. By observing the red aiming laser within a trial run, the misalignment between the actual and the desired path is measured and corrected prior to high-power scanning ablation. In contrast to this approach, direct position control can be modeled taking trifocal geometry constraints between stereo camera and laser deflection unit into account (see Figure 1.6a) [AT15]. Based on tracking the visible spot, accurate vision-based laser control, as demonstrated on three-dimensional targets and on vocal fold tissue of a human cadaver, is feasible. Further extension is discussed for decoupling longitudinal and lateral laser positioning in order to counteract thermal damage and carbonization by adapting scanning speed without compromising path following accuracy [STA15].

However, aforementioned concepts for vision-guided TLM do not account for explicit laser focus adaptation as a prerequisite for atraumatic and char-free incisions. Furthermore, fast laser scanning with closed-loop control demands for robust laser spot detection at increased image acquisition

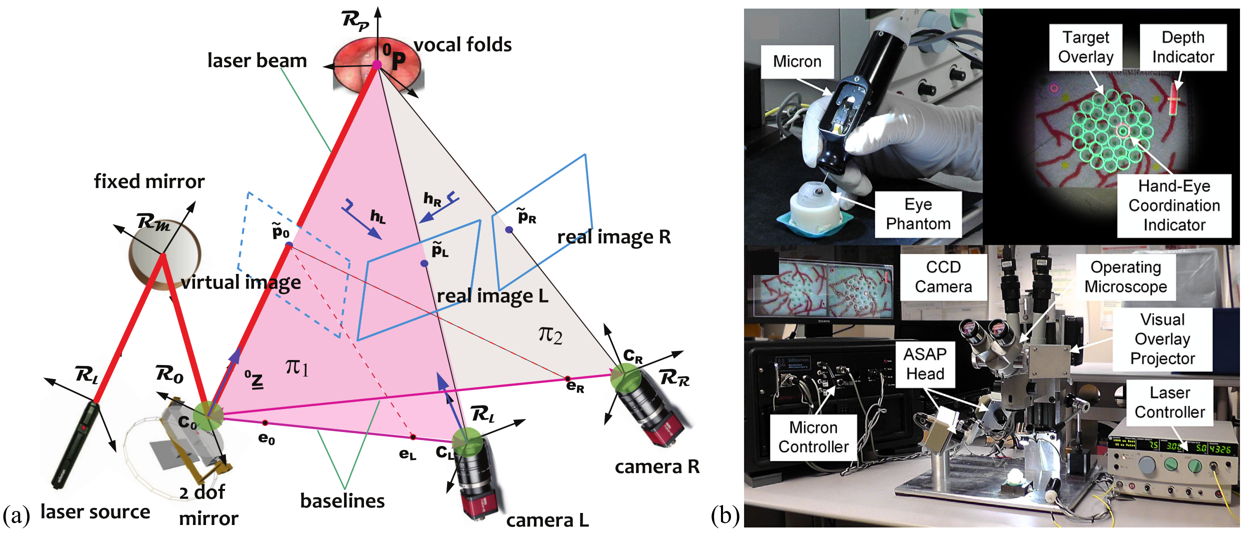


Figure 1.6: Vision-guidance for (a) path following in endoscopic laser phonomicrosurgery based on trifocal geometry (reprinted from [AT15], © 2015, with kind permission from SAGE Publications), and (b) motion-compensation in intraocular laser surgery (reprinted from [YLMR15], © 2015, with kind permission from John Wiley and Sons).

rate. This requirement usually conflicts with the camera exposure time that is necessary to adequately visualize the surgical scene when using conventional white-light imaging. In view of motion compensation during ablation, initial work is discussed for intraocular laser surgery. Here, homography-based tracking is employed for photocoagulation with a handheld laser probe while compensating for eye movements (see Figure 1.6b) [YLMR15]. While a planar retina scene can adequately be represented by rigid, affine, or similarity transforms [PB16], laser ablation executed on deforming tissue has not been addressed so far. In especially in TLM, tissue deformation is prone to by respiratory motion artifacts and tissue exposure with surgical forceps; thus, the risk of unintended injury of adjunct cell compounds is increased.

### 1.2.3 Vision-Based Scene Structure and Motion Estimation

Visual guidance in robot-assisted MIS is extensively discussed in literature. In particular, stereo vision-based methods have grown in popularity since the associated hardware for both imaging and three-dimensional visualization has already found its way into the operating room.

As highlighted in the previous section, current vision-guidance in laser microsurgery considers scene depth only in an implicit manner and within in a very small margin of the laser spot. However, the surgical work flow demands for enhanced techniques enabling intuitive laser focus positioning as well as temporally consistent planning and ablation in a soft tissue environment. A promising solution is offered by dense stereo matching and tracking that provide structural as well as motion information of the tissue in a global context. As a consequence, online acquired depth allows for instant feedback and input on the surgical planning, e.g., if the planned incision is located within the laser focal range. Beyond that, multimodal image registration and augmented reality visualization in the stereo view without compromising depth perception are feasible.

In the following, recent methods for intraoperative tissue surface estimation and non-rigid motion tracking, not solely focusing on stereo vision, are briefly reviewed.

### Three-Dimensional Reconstruction of Soft Tissue Surface Structure

One of the classical problems in computer-assisted interventions is dedicated to accurate, online acquisition of tissue surface structure as a prerequisite to intraoperative navigation. A variety of optical techniques for surface reconstruction, generally classified into active and passive imaging, exists [MIH11]. While passive methods process raw camera images, active systems additionally use controlled light that the surgical scene is exposed to.

Common active techniques are based either on structured light or time-of-flight imaging. In the first case, a light pattern of known geometry is projected onto the surface while being detected by at least one camera. Object depth is computed with the trigonometry of optical triangulation. Recent research in MIS demonstrates significant miniaturization of the projector enabling the integration into laparoscopes (see Figure 1.7a) [CSMH$^+$11, MADdM12, SFSA12, FMS$^+$16], and also to imaging systems for application in the upper aerodigestive tract [CLZQ03, MNSU13, NHH$^+$16]. In contrast, time-of-flight imaging devices measure the travel time of a light signal between camera and object surface. Since neither correspondence search nor a baseline are required, fast depth estimation and compact design enable endoscopic application [PHS$^+$09, MMS$^+$11]. Disadvantages of time-of-flight cameras are related to interference if multiple sensors are used concurrently or to multiple reflections at the scene surface distorting the distance information.

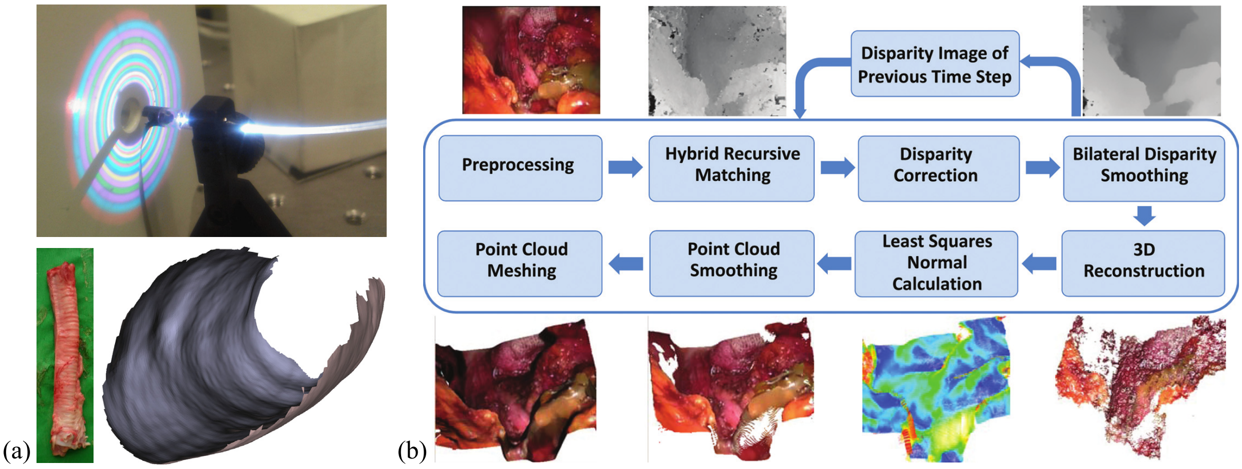

(a)
(b)
Disparity Image of Previous Time Step
Preprocessing
Hybrid Recursive Matching
Disparity Correction
Bilateral Disparity Smoothing
3D Reconstruction
Least Squares Normal Calculation
Point Cloud Smoothing
Point Cloud Meshing

Figure 1.7: Intraoperative surface reconstruction with (a) structured light endoscope prototype demonstrated on an *ex vivo* lamb trachea (reprinted from [SFSA12], © 2012, with kind permission from Elsevier), and (b) hybrid recursive matching based on laparoscopic stereo images (reprinted from [RBS$^+$12], © 2012, with kind permission from John Wiley and Sons).

Passive methods process images acquired from single, stereo, or multiple camera views. For image sequences acquired with a moving camera, structure-from-motion facilitates simultaneous computation of scene depth and camera motion. In MIS, recent work demonstrates structure-from-motion with extension to deformable surfaces and enhanced robustness to outliers resulting from incorrect endoscope tracking [MBC11, HPF+12]. If structure-from-motion is not feasible due to a limited range of motion, endoscopic shape-from-shading provides a valuable alternative [WNJ10, CB12, VSSY12]. This technique recovers shape from a gradual variation of the pixel brightness based on modeling the relationship between light source direction and surface normal.

Stereo vision, as further passive imaging technique, offers the clinician a three-dimensional view when deploying stereoscopic displays. Computational methods for metric surface reconstruction have been discussed in literature, i.e., for non-medical applications. Determining stereo correspondence, i.e., the disparity that describes the pixel displacement of an object point projected to the left and to the right view, generally considers the following four steps: matching cost computation, cost aggregation, disparity computation, and disparity refinement [SS02].

In general, stereo matching can be grouped into local and global methods. Local approaches use a pixelwise or a patch-based photometric similarity measure to determine the disparity with the minimum cost by a winner-take-all (WTA) strategy [SS02]. To regularize disparity matching, i.e., in homogeneous image regions, global algorithms consider a smoothness prior within the cost optimization but are computationally demanding [OK85, BVZ01, SZS03]. A lot of effort has been put into local methods due to their real-time capability. The most significant improvement has been achieved by edge-aware filtering of the disparity cost volume [YK06, HRB+13, YJL+14]. To improve the reconstruction of slanted surfaces either pixelwise cost aggregation along multiple directions [Hir05, GRU10] or piecewise affine functions can be considered [BRR11, HKJK13].

An early work of correlation-based matching on *in vivo* data is dedicated to robot-assisted cardiac interventions [DMCM01]. In this field of application, geometric models such as piecewise bi-linear maps or B-Splines have been used to regularize disparity estimation, assuming that the heart surface is smooth [LRC+04, SDY05]. In addition, a variety of near real-time methods has been discussed in the context of robot-assisted abdominal surgery. For instance, semi-dense, but locally consistent depth can be computed at $15$ frames per second (at an image resolution of $360 \times 288$ pixels) by propagating disparity around salient feature correspondences [SSPY10]. As shown in Figure 1.7b, nearly dense reconstruction at $60$ frames per second (image resolution of $320 \times 240$ pixels) is achieved by hybrid matching combining local correlation with recursive propagation of disparity information from the spatial and temporal neighborhood [RBS+12]. In contrast, global optimization can be addressed by a variational matching framework that considers the Huber loss function to robustly constrain the gradient of disparity [CSDE13]. Compared to the method from [RBS+12], reconstruction density is increased at the expense of the computation rate degrading to $20$ frames per second despite implementation and parallelization on a general purpose graphics processing unit (GPGPU) [CSDE13]. Reconstruction density can also be increased by superpixel-based refinement of invalid disparities [POM$^+$16]; however, this approach is not real-time capable yet.

Further application is dedicated to neurosurgical interventions utilizing operating microscopes. Computational stereo methods enable intraoperative reconstruction of the cortical surface as a prerequisite for registration in image-guided neurosurgery [PFJ05]. To address the problem of frequent intraoperative changes in the microscopic magnification, scale changes can be estimated by establishing feature correspondences between different views [KMP$^+$15].

Due to the variety of methods and applications, different open source validation frameworks have been provided to the research community. These are not solely restricted to images of a real-world scenario, as the Middlebury Stereo Vision toolbox [SS02, GLU12, SHK$^+$14], but also allow to benchmark endoscopic stereo vision methods. In this particular case, laparoscopic *in vivo* and *in vitro* images are provided through the Hamlyn Centre London [HCL17] and through the Open-CAS platform [CAS17], respectively. A first evaluation study on surface reconstruction methods demonstrates superior performance of stereo-based methods, i.e., of aforementioned algorithms [RBS$^+$12, CSDE13], over structured light and time-of-flight endoscopy [MHGB$^+$14]. However, those toolboxes are limited to comparing computational methods for stereo vision rather than different stereo imaging devices, e.g., with respect to their applicability in microsurgery.

### Non-rigid Tracking of Soft Tissue Motion

Recently, vision-based tracking has been focused on MIS due to advances in medical imaging, augmented reality, and robotics. Early studies discussed motion tracking, particularly in the field of beating heart surgery, considering similarity-based template matching [SMD$^+$05, SPT$^+$06, SY07, NTP07]. Tracking purely based on measurement information is computationally efficient, but motion estimation may fail if the measurement is lost. Thus, a priori information has to be taken into account, i.e., in case of measurement noise or scene occlusions. For instance in beating heart surgery, characteristic motion frequencies can be determined by Fourier analysis. Then, the heart motion can be predicted with the help of the Takens' theorem and measurements from the past [OGB$^+$05]. In case of non-predictable motion, tracking robustness, i.e., with respect to occlusions and deformation, can be improved by including geometrical constraints for spatial consistency [YLS$^+$12], multi-affine clustering of the target region [PSM13], affine-invariant feature descriptors (see Figure 1.8a) [GVSY13], or online learning and classification [MY08, YGMY16].

By contrast, non-rigid tracking based on physical models incorporates prior knowledge of the organ's biomechanical characteristics. If physical parameters can be properly modeled and identified during tracking, accurate and occlusion-invariant motion reconstruction is feasible [BWRH07, BPH11]. Furthermore, shape priors acquired preoperatively can be included to ease the initialization with respect to the intraoperative scene and to achieve near real-time capability of tracking on consecutive frames [SRB$^+$14, HCP$^+$15, CBBC16].

To overcome the computational complexity of physical representation, geometric models are often

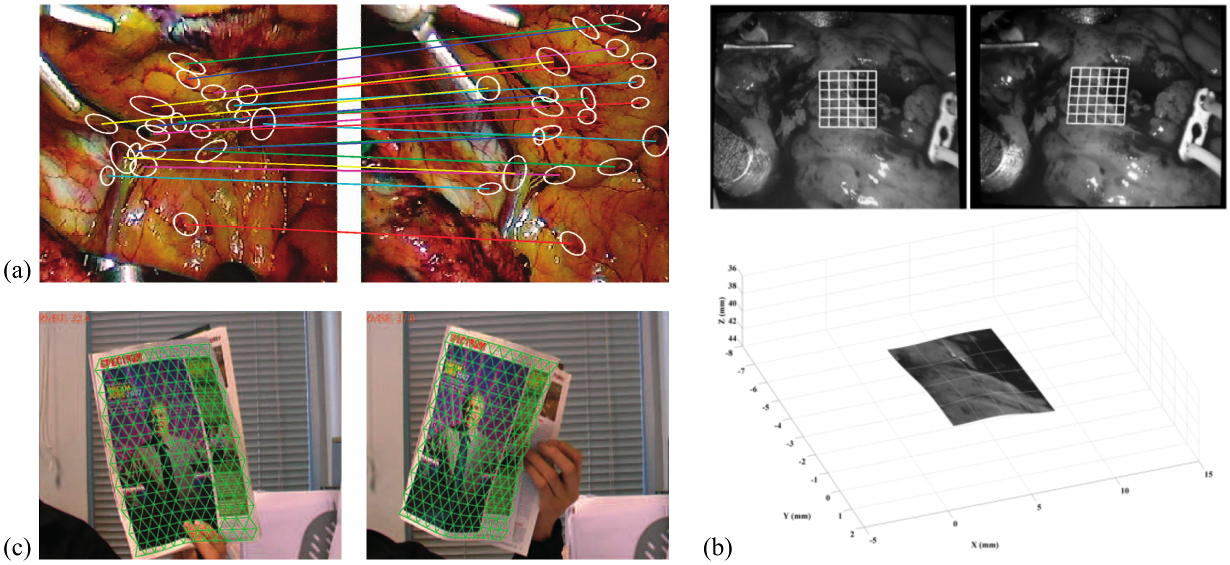


Figure 1.8: Non-rigid motion estimation utilizing (a) probabilistic tracking of affine-invariant anisotropic regions (reprinted from [GVSY13], © 2013 IEEE), (b) TPS-based stereoscopic deformation modeling (reprinted from [RPL10], © 2010, with kind permission from SAGE Publications), and (c) piecewise affine warps fusing features and appearance (reprinted from [ZLH09], © 2009 IEEE).

used taking spatial dependencies between object points into account. Associated optimization aims at minimizing the shape bending energy and the matching error between the current frame and its template model. Object surface approximation with Free-Form Deformation (FFD) (e.g. piecewise bi-linear maps or B-splines) or Radial Basis Functions (RBF) (e.g. Thin Plate Splines (TPS) shown in Figure 1.8b) has been shown to perform well for beating heart motion estimation [LRC$^+$04, RCL$^+$04, SMD$^+$05, RPL10, CBBC16]. In order to reduce the computational load when TPS are used, tracking can be split into intra-frame shape registration and inter-frame motion estimation [YWLP14]. If a deformation is small, primitive models, such as quasi-spherical triangles, can perform as accurately as TPS-based methods [WYLP13]. Accelerated tracking is achieved by inverse compositional optimization [BGBB$^+$11], or learning of non-linear template transformation [THNI14].

In contrast to RBF-based models, which are mainly limited to smooth and continuous deformations, alignment to local geometric changes can be efficiently achieved with piecewise warps providing local support and invertibility [SDP13]. A noteworthy method in the field of vision-based, non-rigid tracking [PLF08] estimates deformations with a triangular mesh of hexagonal elements. In this case, a quadratic energy term is formulated penalizing local surface curvature, whereas outliers are determined with a coarse-to-fine robust estimator function. To accelerate the piecewise affine tracking, as exemplarily shown in Figure 1.8c, the progressive finite Newton (PFN) scheme allows solving the optimization problem within a fixed number of steps [ZLH09]. Application to soft tissue motion estimation for white light and multispectral imaging has been recently discussed [SY09, SRH12, DCA$^+$15]. Piecewise affine warps have been considered not exclusively for endoscopic vision but also for online ultrasound image registration, due to their reduced computational complexity [PDLA$^+$14, RKD$^+$17].

### 1.2.4 Optical Techniques for Laryngeal Tumor Imaging

White light imaging in the form of flexible endoscopy or direct laryngoscopy does not reveal histological information of the laryngeal mucosa; thus, taking biopsies for an *ex vivo* analysis is currently the conventional procedure. To identify tumor margins intraoperatively as a prerequisite for vision-guided resection, a variety of optical imaging technologies has been discussed recently [HSK$^+$10]. Those include autofluorescence imaging based on monochromatic excitation of fluorophores (e.g. porphyrins, collagen, nicotinamide adenine dinucleotide - NADH) occurring in neoplastic mucosa [MDGA02, CSG$^+$13], or aminolevulinic acid induced fluorescence of the cancerous cells [CKI$^+$04, KBLA11]. Alternatively, narrow band imaging (NBI), as shown in Figure 1.9a, enhances the contrast and thus reveals changes of the mucosal vessel structure in abnormal tissue [NHX$^+$11, KFG$^+$16]. In regard to improving detection of laryngeal cancer in early stages and to finding adequate resection margins, initial work on image-based classification of the blood vessel shape and density demonstrates successful detection of laryngeal lesions with an accuracy of $84.3$ percent [BM16].

Fluorescence and narrow band imaging allow for determination of the lateral lesion extent whereas its subepithelial spread is not assessable due to the limited penetration depth of the excitation light [MGLvE$^+$13]. To overcome this limitation, optical coherence tomography (OCT) as a high-resolution, non-invasive, cross-sectional tomographic imaging technology has been extensively discussed in terms of *in situ* optical biopsy. OCT is based on the measurement of near-infrared light that is backscattered at tissue boundaries. Its resolution can vary from $1$ to $15$ µm while a penetration depth of $2$ to $3$ mm is feasible in most tissues [FPBB00]. Originally applied to oph-

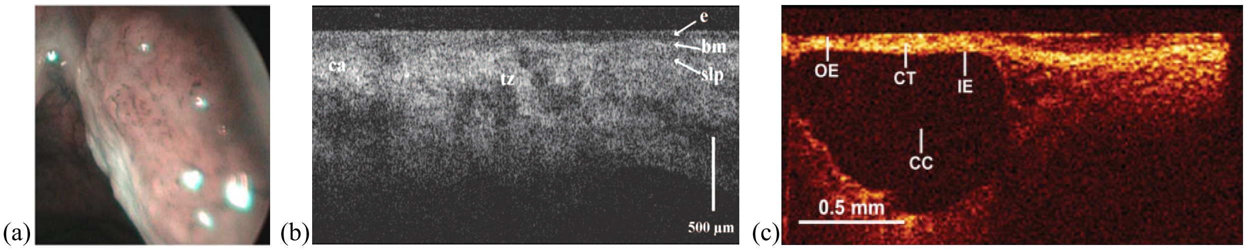


Figure 1.9: Intraoperative imaging of the vocal folds utilizing (a) NBI to enhance vessel contrast (reprinted from [BM16], © 2016 IEEE), (b) OCT imaging of squamous cell carcinoma (e–epithelium, bm–basement membrane, slp–superficial lamina propria, tz–transition zone, ca–cancer) (reprinted from [ARV$^+$06], © 2006, with kind permission from John Wiley and Sons), and (c) OCT imaging of a retention cyst (reprinted from [KGvG$^+$08], © 2008, with kind permission from John Wiley and Sons).

thalmology, several clinical studies highlight the potential of OCT for diagnosis in laryngology [WJG+05, ARV+06, KGvG+08]. Combining OCT with microlaryngoscopy for identifying benign and malignant lesions has shown a detection rate of 89 percent which is higher than for microlaryngoscopy alone yielding only 80 percent [KGvG+08]. In particular, penetration of the basement membrane, that is located between the thin translucent epithelium and the lamina propria, is the most important criterion for evaluating the invasiveness of the malignancy (see Figure 1.9b and 1.9c). If this intermediate layer cannot be clearly detected in the OCT scan, squamous cell carcinoma cannot be excluded. Due to the high clinical value of OCT, recent research additionally focuses on high resolution imaging for microsurgery guidance [ZK11, WDK+14] and on the integration to endoscopic vision [TPT+13, DBR+15] as well as to surgical microscopes [LKO+13]. Utilizing OCT during laser surgery is further motivated by its capability to monitor and control ablation depth with micrometer precision, as recently demonstrated for bone tissue [LWFY12, DKG+13, ZPW+14, FPB+15].

## 1.3 Problems under Consideration

Even though the advantages of tumor resection with high-power lasers are manifold, there are contraindications for TLM. Firstly, anatomical restrictions of the transoral access are mainly related to neck mobility as well as teeth, jaw, and tongue dimensions. Secondly, achieving clear margins while preserving as much healthy tissue as possible depends on accurate laser parametrization and control that, as of today, require a highly experienced surgeon. Otherwise, the risk of incomplete tumor resection and damage of healthy tissue is high. In addition, there are substantial technical challenges. As the literature review has shown, surgical vision in laser therapies has only been addressed to some extent so far. In particular, consistent integration of real-time image computing into soft tissue laser surgery, i.e., considering the laser model, structural and motion information of the scene, will facilitate novel assistive functions overcoming the following limitations.

**Narrow field of view and limited range of motion**
Even though a microscope provides a magnified view, a significant part of the surgical scene is hidden by the laryngoscope inserted to the throat. This leads to inadequate visualization and frequent repositioning in case of larger tumors. Furthermore, nearby malignant structures, i.e., in the supraglottic space, might be covered by the laryngoscope and thus potentially be overlooked. In contrast, endoscopes facilitate a wide or even angular perspective; however, they cannot be utilized simultaneously with the ablation laser. In addition to the narrow field of view, the tube-like access drastically limits the lateral range of motion for instruments. As a consequence, precise instrument movements are difficult to master and cause inevitable occlusions in the direct line of sight.

**Inaccurate laser focus positioning**
The radiation energy of a $CO_2$ lasers is strongly absorbed by tissue with a high content of water. A millisecond, high-power laser pulse leads to very local water evaporation whereby non-liquid tissue

particles are suddenly released; thus, the extent of collateral damage is reduced to a minimum [Nie13]. However, this scenario only applies for a precisely focused laser. As a consequence, microspot surgery on the vocal folds demands for focus positioning within submillimeter accuracy; otherwise, no ablation effect is achieved. Due to the quadratic relation between the focal depth and the spot size, decreasing the spot diameter of a $CO_2$ laser from $0.5$ to $0.1$ mm drastically reduces the available depth of focus from $17$ to $0.68$ mm, respectively [SA00]. However, current surgical systems do not provide online focus adaptation. Thus, surgeons prevalently tolerate spot sizes greater than $0.25$ mm to benefit from a larger depth of focus, i.e., for bulky tumors that would require frequent focus repositioning due to a adjustments of the laryngoscope [ORJ$^+$14].

**Limited imaging of tumor submucosal extent**

Even though experienced clinicians are able to identify the transition between benign and malignant cell compounds under microscopic vision, estimating the depth of infiltration to the subepithelial vocal fold layers is a challenging task. The assessment of tumor growth by preoperative diagnosis with CT or MRI has to be treated with caution, since the tumor size is often overestimated if inflammatory changes and edema surround the lesion; or is underestimated if the submucosal spread is not recognizable [SA00]. As a result, this might lead either to overtreatment with unnecessary radicality of the resection, or to undertreatment with an increased risk of recurrence. Another limitation of preoperative imaging comes along with soft tissue motion artifacts and deformation; thus, correct oncologic interpretation in the conventional modalities is difficult [SA00].

**Inadequate incision planning interface**

The micromanipulator generally enables accurate laser aiming and control but is vulnerable to erroneous user inputs arising from inexperience, tremor, or poor ergonomics. Operating the laser over a large distance ($\sim 400$ mm) while performing relatively small movements exacerbates the accuracy and thus requires intensive training. Due to the adverse position of the micromanipulator, that is attached to the microscopic head, manual laser control suffers from the sub-optimal forearm support to steady hand tremor [Gia08]. Although scanning-assistance is provided by commercial systems, the clinician is still encouraged to manually pre-position the scan pattern with respect to the tumor which is likewise compromised by aforementioned restrictions.

**Reduction of ablation accuracy due to soft tissue motion**

Laser aiming inaccuracies do not solely arise from manual control as outlined above. Misalignment of the laser path and focus is also evoked by scene motion. Firstly, there is no rigid, mechanical fastening between patient and laser system; thus, forces externally applied to the microscope body most likely results in positional deviation of the laser spot. Secondly, respiratory motion artifacts or tissue tensioning with grasping forceps lead to non-negligible deformation that is difficult to handle during ablation, i.e., when function preservation with resection margins of less than a millimeter is the aim. Despite already existing assistive functions such as high-speed scanning, the current surgical planning and ablation process does not take any motion compensation into account.

## 1.4 Components of the Research and Contributions to the µRALP Project

The work described in this dissertation is part of the achievements of the multidisciplinary EU-funded project µRALP [uRA15]. It proposes a redesign of the conventional setup for laser microsurgery through research on micro-robotic and endoscopic laser tools, stereo vision, augmented reality, human-machine interfaces, and cognitive supervision of the ablation process. The developed robotic endoscope, as shown in Figure 1.10a, provides easier access to the surgical site and a widened field of view. In combination with an information-rich augmented reality interface, a new level of perception, accuracy, and safety for laser microsurgery is achieved; thus, the µRALP concept constitutes a valuable extension to the conventional clinical setting.

Overcoming the limitations outlined in Section 1.3, this dissertation presents validated methods and algorithms for stereo vision-guided laser microsurgery with special emphasis on laryngeal interventions. This work is technically not limited to the endoscopic µRALP approach [KSMO13, KSKO15]. The developed methods can be like-wisely applied to microscopic TLM.

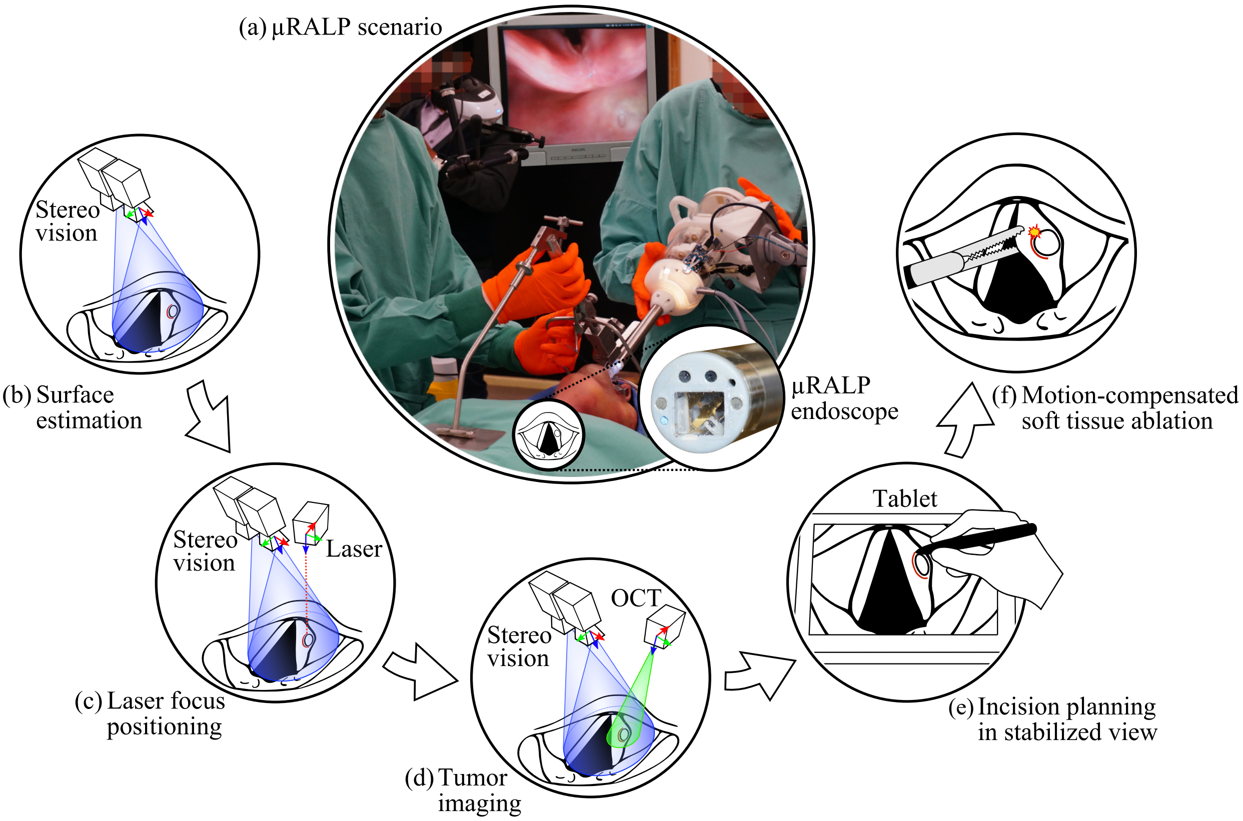


Figure 1.10: Contributions of this thesis are based on (a) the µRALP concept and include (b) three-dimensional tissue surface estimation, (c) distance-based laser focus positioning, (d) fusion of stereo vision with OCT, (e) tablet-based planning in the stabilized view, and (c) motion-compensated laser ablation on soft tissue.

In view of the intraoperative work flow outlined in Figure 1.10b–f, tissue surface information is acquired as a prerequisite to multimodal registration and vision-guided laser control. Since stereo vision is present anyway for visualization, it seems attractive to introduce computational stereo methods (see Figure 1.10a). An algorithm for real-time and robust surface estimation is developed and assessed on a variety of optical imaging devices including endoscopic and microscopic vision. In regard to the requirements of laser microsurgery, the impact of camera baseline, distance to the tissue surface and microscopic magnification level is discussed with respect to reconstruction accuracy.

To enhance intraoperative incision planning and laser focusing by means of computed tissue surface information, the second contribution bases on establishing a trifocal model of stereo camera and ablation laser (see Figure 1.10c). A methodology for laser-to-camera registration is proposed to facilitate automatic, distance-based laser focus adjustment and thus to always ensure optimal energy exposure to the tissue. To guide the surgeon during optimal focus positioning, i.e., at sub-millimeter accuracy, an augmented reality framework with color-encoded distance between target surface and laser focal range is implemented.

Since OCT imaging is considered being a promising diagnostic tool for detecting submucosal changes, the third contribution aims at fusing OCT imaging with stereo vision (see Figure 1.10d). A registration and segmentation framework is developed and practically demonstrated on a phantom replicating the optical properties of laryngeal tissue, i.e., for the epithelium layers. Compared to conventional tomographic imaging, the integration of OCT enables tumor visualization directly in the live view. A color gradient ranging from maximum function preservation (red color) to maximal radicality (green color) is proposed for a potential resection margin.

As the literature review has shown, stylus-tablet-based planning outperforms the state-of-the-art micromanipulator control. Therefore, the fourth contribution of this work, as depicted in Figure 1.10e, concentrates on the integration of three-dimensional surface information and laser-to-camera registration into a tablet interface. This allows not only to fully exploit stereoscopic visualization by means of overlaying the incision to both the left and the right view, but also to compute a virtual laser view for planning from the laser perspective. Different visualization and path definition strategies are developed and evaluated in a user study.

The fifth contribution extends the developed planning interface and ablation system by a novel non-rigid tracking algorithm to estimate tissue motion at runtime and to achieve live view stabilization and closed loop control for motion compensation during laser ablation. This will relieve the surgeon with respect to steadying the surgical scene while accurately tracing the desired path. Beyond that, the tracking enables to perform cuttings with repeated scanning very easily, regardless of concurrent tissue deformation. As a benefit, unintended ablation of nearby structures is avoided and thus function preservation is maximized. The performance of the developed tracking framework is assessed on *in vivo* data and practically demonstrated by a user study as well as by ablation trials on *ex vivo* tissue.

## 1.5 Outline of the Thesis

The following paragraphs provide the structure of this dissertation with associated methods and results that have been published in scientific journals, international and national conference proceedings. Preliminary work and contributions to the referenced publications were accomplished within five supervised student theses [Pod13, Kle13, Lek14, Ber14, Lav15].

**Chapter 2** introduces the theoretical background of this thesis. The fundamentals of laser-tissue interaction, OCT imaging, and stereo camera geometry are described. Furthermore, parallel computation on graphics processing units and the experimental setup used throughout this thesis are briefly presented.

**Chapter 3** presents a real-time method for surface reconstruction. The algorithmic performance is analyzed with respect to runtime and accuracy on *in vivo* image data. Afterwards, the reconstruction algorithm is applied to different stereo imaging devices in order to evaluate applicability in laser microsurgery. The methods and results of this chapter are in great part published in [SPK$^+$13, SKKO16].

**Chapter 4** describes a methodology for aligning laser workspace information and volumetric OCT scans to the stereo video as a prerequisite to surgical guidance. Based on a trifocal model and registration of laser and camera, distance-based focus adjustment is demonstrated by ablation on *ex vivo* tissue. Subsequently, results of a two-step, surface feature-based registration of stereo vision and OCT are presented. The methods and results of this chapter are published in [SKK$^+$15, BKS$^+$15].

**Chapter 5** presents surgical planning concepts considering stereo vision, laser, and OCT. Based on the laser-to-camera registration, a color-encoded laser workspace mapping is described providing visual feedback during focus positioning. Next, OCT-based segmentation of subepithelial lesions its visualization in the live view are demonstrated with a laryngeal tissue phantom. Finally, tablet-based incision planning is discussed in terms of ablation accuracy, time, and usability. The methods and results of this chapter are published in [SLK$^+$14, SLK$^+$15, CGS$^+$15, SKL$^+$16].

**Chapter 6** presents a novel non-rigid tracking framework for soft tissue motion estimation in the context of laser microsurgery. Special emphasis is given to real-time optimization. The algorithm is validated on *in vivo* data and compared with state-of-the-art. For practical demonstration, the method is embedded into the planning interface providing image stabilization during path definition and compensating for tissue motion during laser ablation. User performance and ablation accuracy are discussed. The methods and results of this chapter are published in [SLKO16, SKKO17].

**Chapter 7** concludes this thesis and provides directions of further research related to this work.

# 2 Background

This chapter provides the background for the research conducted in this dissertation. Since this work aims at stereo vision-guided microsurgery considering a multimodal setup, the fundamentals of surgical lasers, OCT imaging, and stereo vision are presented. Subsequently, parallel computing on the GPGPU for real-time implementation of the image processing is described briefly. The chapter closes with an introduction of the laser setup used throughout this thesis.

## 2.1 Surgical Lasers

Surgical lasers are an emerging field enabling the surgeon further advanced ablation techniques. Almost bloodless and char-free incisions are feasible allowing for an unobstructed view during surgery and a clear histopathological confirmation after tumor removal [ORJ$^{+}$14]. A further advantage is that the application of visible or infrared lasers is categorized as non-ionizing whereas the radiation energy is transformed into heat in a very locally defined proximity inside the tissue.

### 2.1.1 Function Principle

Laser is an acronym for light amplification by stimulated emission of radiation. This term describes the principle of generating optical radiation that is characterized by an extraordinarily high level of coherence, monochromaticity, and directionality [Trä12].

A laser mainly consists of a gain medium, two mirrors functioning as a resonator, and an external pump source. The latter component creates electromagnetic waves that excite the atoms in a solid, liquid, or gaseous gain medium to a higher unstable energy state. These atoms tend to fall back into the stable low-energy state while emitting a photon of random direction and phase (spontaneous emission). Laser light originates from a second emission type, denoted by stimulated emission. Radiation of the same frequency and phase as of the incoming light is emitted. Sustaining this effect requires perpetual excitation of the gain medium. In other words, the number of atoms occupying a higher energy state has to exceed the amount of those remaining in the ground state (population inversion). This is achieved by optical (e.g., incoherent light source) or electrical pumping (e.g., current flow through the gain medium) [Trä12]. This emission is further amplified by two mirrors encapsulating the gain medium and forming a resonant optical cavity. The laser beam is constituted by a fraction of the light passing through one of the mirrors being partially permeable.

In terms of clinical use, the most relevant parameters of a surgical ablation laser are the wavelength,

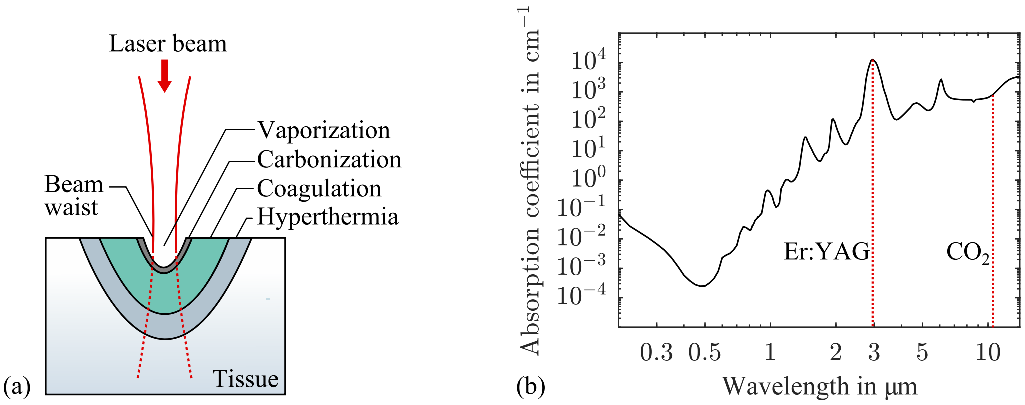


Figure 2.1: Laser ablation is predominated by wavelength-dependent energy absorption in water resulting in (a) different thermal effects, while (b) maximum absorption is enabled by Er:YAG lasers [HQ73].

the focal spot size, the beam profile, and the energy density [SA00]. The wavelength typically varies from the visible to the infrared spectrum. The focal spot defines the smallest diameter of the beam waist (see Figure 2.1a) and it is minimal for a Gaussian beam profile. The depth of focus is determined by the Rayleigh length that describes the distance on the optical axis where the cross section area is doubled with respect to the beam waist [Trä12]. In practice, small spot sizes result in a short depth of focus; hence, the target has to be aligned accurately [SA00]. Furthermore, ablation quality depends on whether the laser is operated in continuous or pulsed mode. The latter ensures thermal relaxation of the tissue between the high-energy pulses.

### 2.1.2 Laser-Tissue Interaction

In principle, there are three effects of laser light interacting with matter: (1) reflection and refraction, (2) absorption, and (3) scattering [Nie13]. If the electromagnetic wave impinges on a physical boundary between different materials, a portion of the light returns from the surface (reflection) while the remaining fraction penetrates into the material. For the latter, a part passes through whereas its direction and velocity can change with the optical density of the propagation medium (refraction). Another part of the radiation is attenuated due to conversion into thermal energy (absorption), or because of dispersion as a result of collision with particles (scattering). These effects usually appear simultaneously while their proportions are determined by the laser wavelength and optical properties of the material.

Regarding biological tissue ablation, the most relevant mechanism is absorption. The energy conversion is classified into photochemical interaction, thermal interaction, photoablation, plasma-induced ablation, and photodisruption [Nie13]. Even though the occurrence of these interaction principles requires roughly the same energy density, it is the exposure time that selects one of the transformation effects [Bou86]. In laser surgery, ablation is predominated by thermal interaction as the exposure time commonly exceeds a microsecond. This mechanism is briefly presented in the following. For details on the other mechanisms, the reader is kindly referred to [Nie13].

Thermal interaction is based on heating-up tissue in the local proximity of the focused laser striking the target surface. Depending on the heat conduction and associated increase of tissue temperature, several effects appear. As shown in Figure 2.1a, these are primarily hyperthermia, coagulation, vaporization, and carbonization [Nie13]. Hyperthermia (42– 50 °C) usually is reversible, provided that the irradiation does not last longer than a few seconds. For temperatures above 60 °C, permanent changes are induced by coagulation of the tissue, i.e., by protein denaturation. Tissue ablation is primarily achieved by vaporization at 100 °C. This effect benefits from wavelength-dependent energy absorption in water. The vaporization causes a rapid increase of pressure due to the expansion of the water molecules in biological tissue. Microexplosions finally cause the desired ablation that is denoted by thermal decomposition [Nie13].

The wide clinical use of $CO_2$ lasers (wavelength $\lambda_{CO_2} = 10.6\,\mu m$) is motivated by the high absorption coefficient as shown in Figure 2.1b. However, the maximum absorption in water is obtained with the erbium-doped yttrium aluminum garnet (Er:YAG) laser (wavelength $\lambda_{Er:YAG} = 2.94\,\mu m$) that is utilized throughout this dissertation. Clinical studies have demonstrated reduced thermal damage zones when compared to the $CO_2$ laser [WFD89, LRVdM92]. For laser phonomicrosurgery, the Er:YAG laser provides a valuable alternative since carbonization and associated adverse effects on voice production are noticeably reduced [HCHF93, LLP07, BJK$^+$13].

## 2.2 Optical Coherence Tomography

Optical coherence tomography (OCT) is a recent high-resolution, cross-sectional imaging technique. Its resolution commonly varies from 1 to 15 µm while a penetration depth of 2 to 3 mm is feasible in most tissues [FPBB00]. The clinical use is motivated by diagnostic capabilities such as *in situ* optical biopsy. Originally applied to ophthalmology, several clinical studies highlight the potential of OCT for diagnosis in laryngology [WJG$^+$05, ARV$^+$06, KGvG$^+$08].

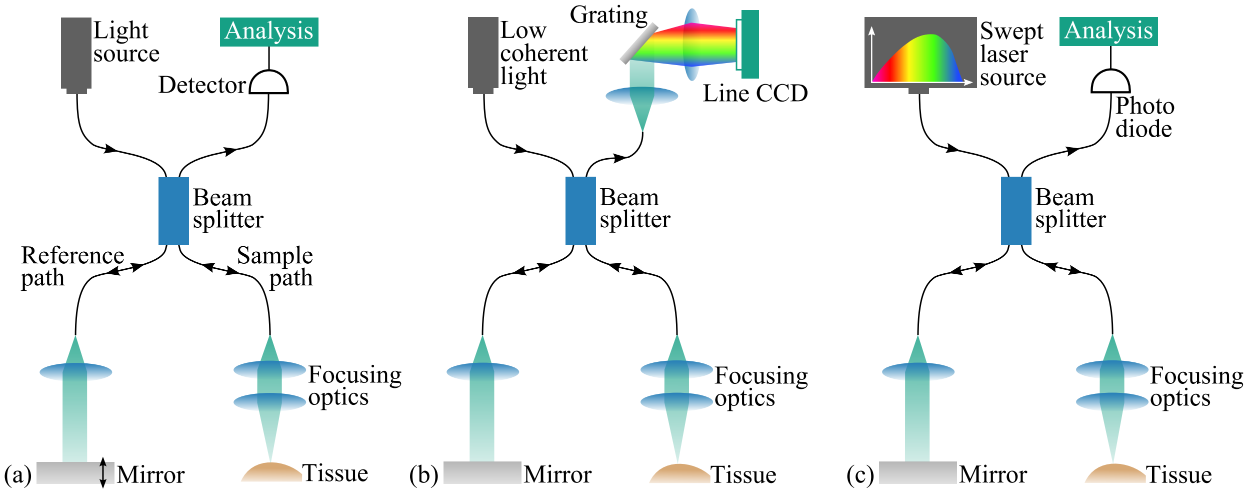


Figure 2.2: Optical biopsy based on (a) TD-OCT, (b) SD-OCT, and (c) SS-OCT.

Near-infrared, low-coherent light is backscattered at tissue boundaries and measured with the interferometry principle. A beam splitter directs the emitted light into a tissue sample path and a reference arm, as shown in Figure 2.2a. Regarding the first-generation devices based on Time-Domain OCT (TD-OCT), the interference amplitude of the reflected light from each arm is measured with a photodetector. Since interference only occurs when the distance traveled by the two light beams is within the coherence length, the measurable depth range is limited to a few micrometers. Millimeter-range depth scanning of the tissue layers is achieved by axially moving the mirror in the reference path. This mode is called A-scan. Cross-sectional slices or entire volumetric images are acquired by lateral scanning, categorized as B-scans and C-scans, respectively. However, since axial scanning is mechanically limited to a few kHz, TD-OCT does not provide fast imaging.

This limitation is addressed by Frequency-Domain OCT (FD-OCT). Instead of scanning the reference arm, depth information is encoded by the frequency of the light. Analyzing the spectral interference with the Fourier transform allows locating the reflective tissue layers. Two common techniques exist. The first one, the Spectral-Domain OCT (SD-OCT), requires a spectrometer splitting the reflected signal into different wavelength (see Figure 2.2b). The second principle, that is Swept-Source OCT (SS-OCT), bases on sweeping the wavelength of the light source over time while detecting the signal with a photo diode (see Figure 2.2c). In contrast to TD-OCT, both techniques operate at high acquisition speed enabling video-rate B-scans of the tissue morphology.

## 2.3 Stereo Vision

In the following, the fundamentals of mono and stereo vision are presented. A dual camera configuration imitates natural binocular vision and is mainly characterized by the baseline and the camera convergence. Regarding human stereo vision, objects are focused by altering the angle between the eyes, as shown in Figure 2.3a. For stereo cameras, as shown in Figure 2.3b, the field of view overlap is limited due to the parallel arrangement. In practice, the cameras are angled towards each other to maximize the view overlap at the working distance, as depicted in Figure 2.3c.

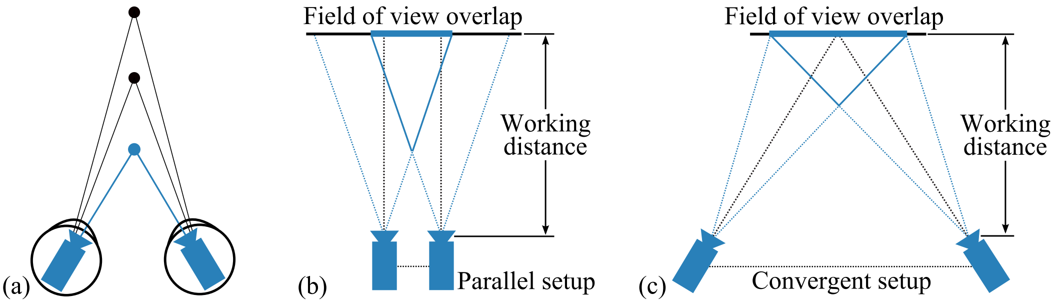


Figure 2.3: Human binocular vision, as shown in (a), enables depth and structure perception of objects at different distances. Based on this principle, computer stereo vision is commonly uses (b) a parallel or (c) a convergent stereo setup in order to maximize the field of view overlap.

### 2.3.1 Coordinate Transformation

Computational stereo methods often demand for transferring an object point between the two camera views. An arbitrary point $\boldsymbol{P} = (x,y,z)^{\mathrm{T}}$ in object space, as shown in Figure 2.4a, can be described with respect to a certain coordinate frame, e.g., the left camera frame $(\mathrm{CF})_{\mathrm{L}}$ of a stereo configuration yielding the notation ${}_{(\mathrm{L})}\boldsymbol{P} = ({}_{(\mathrm{L})}x,{}_{(\mathrm{L})}y,{}_{(\mathrm{L})}z)^{\mathrm{T}}$. A simple matrix multiplication defined by

$$ {}_{(\mathrm{R})}\tilde{\boldsymbol{P}} = {}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{L}}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = \left(\begin{array}{ccc|c} \multicolumn{3}{c|}{{}^{\mathrm{R}}\boldsymbol{R}_{\mathrm{L}}} & {}_{(\mathrm{R})}\boldsymbol{t}_{\mathrm{L}} \\ \hline 0 & 0 & 0 & 1 \end{array}\right) {}_{(\mathrm{L})}\tilde{\boldsymbol{P}} \tag{2.1} $$

allows for transforming the point ${}_{(\mathrm{L})}\boldsymbol{P}$ to another coordinate frame, e.g., the right camera frame $(\mathrm{CF})_{\mathrm{R}}$. The homogeneous transformation ${}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{L}} \in \mathrm{SE}(3)$ is composed of a rotation matrix ${}^{\mathrm{R}}\boldsymbol{R}_{\mathrm{L}} \in \mathrm{SO}(3)$ and a translation vector ${}_{(\mathrm{R})}\boldsymbol{t}_{\mathrm{L}} \in \mathbb{R}^{3\times1}$. The point ${}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = ({}_{(\mathrm{L})}x,{}_{(\mathrm{L})}y,{}_{(\mathrm{L})}z,1)^{\mathrm{T}}$ is given in homogeneous coordinates. Same applies for the point ${}_{(\mathrm{R})}\tilde{\boldsymbol{P}}$. Within a single multiplication, the transform considers both the relative rotation and the translation between two coordinate systems.

If the point of interest is provided with reference to a third coordinate frame, e.g. the OCT imaging frame $(\mathrm{CF})_{\mathrm{O}}$, the transform to the frame $(\mathrm{CF})_{\mathrm{R}}$ can be written as

$$ {}_{(\mathrm{R})}\tilde{\boldsymbol{P}} = {}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{O}}\,{}_{(\mathrm{O})}\tilde{\boldsymbol{P}} = {}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{L}}\,{}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{O}}\,{}_{(\mathrm{O})}\tilde{\boldsymbol{P}}, \tag{2.2} $$

where the homogeneous transformation matrix ${}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{O}}$ is concatenated by ${}^{\mathrm{R}}\boldsymbol{T}_{\mathrm{L}}$ and ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{O}}$.

### 2.3.2 Pinhole Camera Model

The most common representation of a camera in computer vision is the pinhole model depicted in Figure 2.4b. The perspective transform that projects a point ${}_{(\mathrm{L})}\boldsymbol{P} = ({}_{(\mathrm{L})}x,{}_{(\mathrm{L})}y,{}_{(\mathrm{L})}z)^{\mathrm{T}}$ from object space to pixel position $\boldsymbol{p}_{\mathrm{L}} = (u_{\mathrm{L}},v_{\mathrm{L}},)^{\mathrm{T}}$ in the image plane is defined by

$$ s\,\tilde{\boldsymbol{p}}_{\mathrm{L}} = \begin{pmatrix} f_u & \alpha & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix} (\boldsymbol{I}\,|\,\boldsymbol{0})\;{}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = \boldsymbol{K}_{\mathrm{L}}\,(\boldsymbol{I}\,|\,\boldsymbol{0})\;{}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = {}^{\mathrm{L}}\boldsymbol{M}_{\mathrm{L}}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}}, \tag{2.3} $$

where $\boldsymbol{K}_L \in \mathbb{R}^{3\times3}$ is the camera matrix. The left and right hand sides are equal up to a non-zero scaling $s$. The intrinsic parameters are defined by the optical center $\boldsymbol{c}_{\mathrm{L}} = (c_u,c_v)^{\mathrm{T}}$, known as the principal point, and the focal lengths $f_u$ as well as $f_v$. The shearing $\alpha$ is considered to be zero in today's camera image sensors. The pixel $\tilde{\boldsymbol{p}}_{\mathrm{L}} = (u_{\mathrm{L}},v_{\mathrm{L}},1)^{\mathrm{T}}$ is given in homogeneous coordinates. Even though this section refers to single view geometry, the frame $(\mathrm{CF})_{\mathrm{L}}$ is already introduced here referring to the left camera of the stereo model described in Section 2.3.3. The left camera's optical center likewisely defines the world coordinate system. Thus, the projection matrix ${}^{\mathrm{L}}\boldsymbol{M}_{\mathrm{L}} \in \mathbb{R}^{3\times4}$ is obtained by multiplying the camera matrix $\boldsymbol{K}_L$ with an external transformation that is set to $(\boldsymbol{I}\,|\,\boldsymbol{0}) \in \mathbb{R}^{3\times4}$ as the point ${}_{(\mathrm{L})}\boldsymbol{P}$ is already referred to $(\mathrm{CF})_{\mathrm{L}}$.

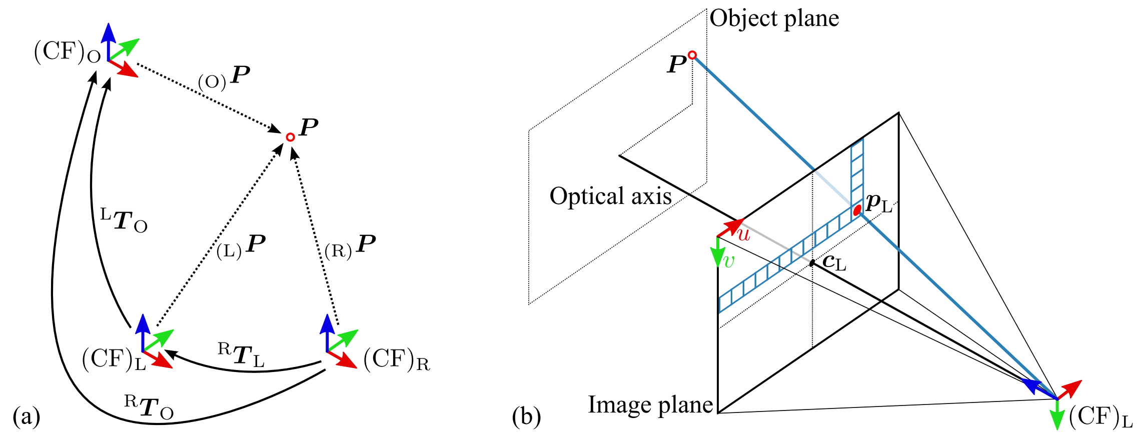

$(\mathrm{CF})_\mathrm{O}$
$_{(\mathrm{O})}\boldsymbol{P}$
$\boldsymbol{P}$
$^\mathrm{L}\boldsymbol{T}_\mathrm{O}$
$_{(\mathrm{L})}\boldsymbol{P}$
$_{(\mathrm{R})}\boldsymbol{P}$
$(\mathrm{CF})_\mathrm{L}$
$^\mathrm{R}\boldsymbol{T}_\mathrm{L}$
$(\mathrm{CF})_\mathrm{R}$
$^\mathrm{R}\boldsymbol{T}_\mathrm{O}$
(a)
Object plane
$\boldsymbol{P}$
Optical axis
$\boldsymbol{p}_\mathrm{L}$
$u$
$v$
$\boldsymbol{c}_\mathrm{L}$
Image plane
$(\mathrm{CF})_\mathrm{L}$
(b)

Figure 2.4: Transformation of a point $\boldsymbol{P}$ is depicted in (a). Its projection to the image plane is shown in (b).

The camera model commonly takes radial and tangential distortion into account. Correction of the radial distortion is achieved by

$$\begin{aligned} u_{\mathrm{cor,r}} &= u(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ v_{\mathrm{cor,r}} &= v(1 + k_1 r^2 + k_2 r^4 + k_3 r^6), \end{aligned} \tag{2.4}$$

whereas the tangential distortion correction is accomplished by

$$\begin{aligned} u_{\mathrm{cor,t}} &= u + 2p_1 vu + p_2(r^2 + 2u^2) \\ v_{\mathrm{cor,t}} &= v + p_1(r^2 + 2v^2) + 2p_2 uv \end{aligned} \tag{2.5}$$

with $r^2 = (u - c_u)^2 + (v - c_v)^2$. In short, five distortion coefficients $\boldsymbol{d}_L = (k_1,k_2,k_3,p_1,p_2)$ have to be found that are, along with the intrinsic parameters, determined by planar pattern-based calibration [Zha00, Bou04]. The camera parameters are derived by capturing $n$ images of a planar calibration pattern with $m$ grid points $_{(\mathrm{P})}\boldsymbol{P}_k$ under varying rotation and translation as given by the transform $^\mathrm{L}\boldsymbol{T}_{\mathrm{P},j}$. Camera calibration aims at minimizing the re-projection error

$$\sum_{j=1}^{n}\sum_{k=1}^{m}\left\|\boldsymbol{p}_{\mathrm{L},j,k} - \hat{\boldsymbol{p}}_\mathrm{L}\left(\boldsymbol{K}_\mathrm{L}, \boldsymbol{d}_L, {}^\mathrm{L}\boldsymbol{T}_{\mathrm{P},j}, {}_{(\mathrm{P})}\boldsymbol{P}_k\right)\right\|_2^2, \tag{2.6}$$

where $\boldsymbol{p}_{\mathrm{L},j,k}$ are the measured and $\hat{\boldsymbol{p}}_\mathrm{L}$ the estimated grid points. The non-linear optimization problem is solved using the Levenberg–Marquardt algorithm [Zha00].

### 2.3.3 Epipolar Geometry and Rectification

Considering the stereo view shown in Figure 2.5a, the geometrical relation between the projection of an object point $\boldsymbol{P}$ into the left and the right image plane, yielding $\boldsymbol{p}_\mathrm{L}$ and $\boldsymbol{p}_\mathrm{R}$, respectively, can be described through epipolar geometry. Any point $\boldsymbol{P}$ defines an epipolar plane $\boldsymbol{\pi}$ with the two optical

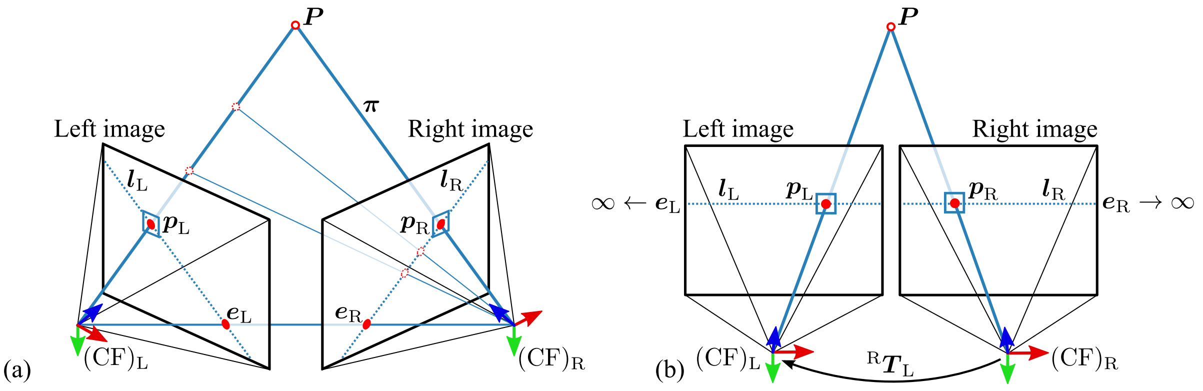

$\boldsymbol{P}$
$\boldsymbol{\pi}$
Left image
Right image
$\boldsymbol{l}_\mathrm{L}$
$\boldsymbol{p}_\mathrm{L}$
$\boldsymbol{l}_\mathrm{R}$
$\boldsymbol{p}_\mathrm{R}$
$\boldsymbol{e}_\mathrm{L}$
$\boldsymbol{e}_\mathrm{R}$
$(\mathrm{CF})_\mathrm{L}$
$(\mathrm{CF})_\mathrm{R}$
(a)
$\boldsymbol{P}$
Left image
Right image
$\infty \leftarrow \boldsymbol{e}_\mathrm{L}$
$\boldsymbol{l}_\mathrm{L}$
$\boldsymbol{p}_\mathrm{L}$
$\boldsymbol{p}_\mathrm{R}$
$\boldsymbol{l}_\mathrm{R}$
$\boldsymbol{e}_\mathrm{R} \rightarrow \infty$
(b)
$(\mathrm{CF})_\mathrm{L}$
${}^\mathrm{R}\boldsymbol{T}_\mathrm{L}$
$(\mathrm{CF})_\mathrm{R}$

Figure 2.5: Epipolar geometry of (a) a convergent stereo configuration and (b) a rectified stereo view.

centers at $(\mathrm{CF})_\mathrm{L}$ and $(\mathrm{CF})_\mathrm{R}$. The intersection of $\boldsymbol{\pi}$ with the image planes yields the epipolar lines $\boldsymbol{l}_\mathrm{L}$ and $\boldsymbol{l}_\mathrm{R}$. The fundamental matrix ${}^\mathrm{L}\boldsymbol{F}_\mathrm{R} \in \mathbb{R}^{3\times 3}$ relates the two points by the epipolar constraint

$$\tilde{\boldsymbol{p}}_\mathrm{L}^{\mathrm{T}}\,{}^\mathrm{L}\boldsymbol{F}_\mathrm{R}\,\tilde{\boldsymbol{p}}_\mathrm{R} = 0 \tag{2.7}$$

such that $\boldsymbol{p}_\mathrm{R}$ must lie on the epipolar line $\boldsymbol{l}_\mathrm{R} = \tilde{\boldsymbol{p}}_\mathrm{L}^{\mathrm{T}}\,{}^\mathrm{L}\boldsymbol{F}_\mathrm{R}$ [HZ04]. As a result, Equation 2.7 reduces the two-dimensional correspondence search space to one dimension. For computational efficiency, the epipolar lines can be horizontally aligned by image rectification [FTV00, Bou04]. In a rectified view, as depicted in Figure 2.5b, the image planes are enforced to be parallel while the intrinsic parameters remain unchanged. Considering the rotation ${}^\mathrm{R}\boldsymbol{R}_\mathrm{L}$ and the translation ${}_{(\mathrm{R})}\boldsymbol{t}$ between the two cameras, the projection transform of the right view is defined by

$$s\,\tilde{\boldsymbol{p}}_\mathrm{R} = \boldsymbol{K}_\mathrm{R}\left({}^\mathrm{R}\boldsymbol{R}_\mathrm{L}\,|\,{}_{(\mathrm{R})}\boldsymbol{t}_\mathrm{L}\right)\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = {}^\mathrm{R}\boldsymbol{M}_\mathrm{L}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}}\,. \tag{2.8}$$

After rectification, the rotation simplifies to ${}^\mathrm{R}\boldsymbol{R}_\mathrm{L} = \boldsymbol{I}$ while the translation is only defined by the horizontal shift ${}_{(\mathrm{R})}\boldsymbol{t}_\mathrm{L} = (-b{,}0{,}0)^\mathrm{T}$. Stereo calibration aims at minimizing the re-projection error

$$\begin{aligned}\sum_{j=1}^{n}\sum_{k=1}^{m}&\left\|\boldsymbol{p}_{\mathrm{L},j,k} - \hat{\boldsymbol{p}}_\mathrm{L}\left(\boldsymbol{K}_\mathrm{L},\boldsymbol{d}_\mathrm{L},{}^\mathrm{L}\boldsymbol{T}_{\mathrm{P},j},{}_{(\mathrm{P})}\boldsymbol{P}_k\right)\right\|_2^2\\&+\left\|\boldsymbol{p}_{\mathrm{R},j,k} - \hat{\boldsymbol{p}}_\mathrm{R}\left(\boldsymbol{K}_\mathrm{R},\boldsymbol{d}_\mathrm{R},{}_{(\mathrm{P})}\boldsymbol{P}_k,{}^\mathrm{R}\boldsymbol{R}_\mathrm{L},{}_{(\mathrm{R})}\boldsymbol{t}_\mathrm{L}\right)\right\|_2^2\end{aligned} \tag{2.9}$$

in order to determine the intrinsic parameters and the external transformation [SDY05].

### 2.3.4 Triangulation

Once the stereo view is rectified, triangulation of a found pixel correspondence $\boldsymbol{p}_\mathrm{L} = (u_\mathrm{L}{,}v_\mathrm{L})^\mathrm{T}$ and $\boldsymbol{p}_\mathrm{R} = (u_\mathrm{R}{,}v_\mathrm{R})^\mathrm{T}$, whereas $v_\mathrm{L} = v_\mathrm{R}$, is given by

$$u_\mathrm{L} = \frac{xf}{z} \quad \text{and} \quad u_\mathrm{R} = \frac{(x-b)f}{z} \quad \Leftrightarrow \quad z = \frac{bf}{u_\mathrm{L} - u_\mathrm{R}} = \frac{bf}{d}\,. \tag{2.10}$$

The pixel shift $d = u_{\mathrm{L}} - u_{\mathrm{R}}$ represents the so-called disparity. If the images are not rectified or if the point correspondence most likely does not satisfy the epipolar constraint in Equation 2.7, an optimal solution can be found by linear least squares triangulation [HS97]. In particular if the two triangulated rays do not intersect, one has to find the shortest distance between the two lines to get the position $\boldsymbol{P}$. Its projection $\boldsymbol{p}_{\mathrm{L}}$, as defined by Equation 2.3, considers equality up to the scale factor $s$. Same applies for the right coordinate $\boldsymbol{p}_{\mathrm{R}}$ according to Equation 2.8. Elimination of $s$ in the perspective transforms yields the following system of linear equations

$$\begin{pmatrix} u_{\mathrm{L}}\,\boldsymbol{m}_{3,\mathrm{L}}^{\mathrm{T}} - \boldsymbol{m}_{1,\mathrm{L}}^{\mathrm{T}} \\ v_{\mathrm{L}}\,\boldsymbol{m}_{3,\mathrm{L}}^{\mathrm{T}} - \boldsymbol{m}_{2,\mathrm{L}}^{\mathrm{T}} \\ u_{\mathrm{R}}\,\boldsymbol{m}_{3,\mathrm{R}}^{\mathrm{T}} - \boldsymbol{m}_{1,\mathrm{R}}^{\mathrm{T}} \\ v_{\mathrm{R}}\,\boldsymbol{m}_{3,\mathrm{R}}^{\mathrm{T}} - \boldsymbol{m}_{2,\mathrm{R}}^{\mathrm{T}} \end{pmatrix} {}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = \boldsymbol{A}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = \boldsymbol{0}, \tag{2.11}$$

where $\boldsymbol{m}_{\mathrm{i,L}}^{\mathrm{T}}$ is the $i$-th row of the projection matrix $\boldsymbol{M}_{\mathrm{L}}$. Same applies for the rows $\boldsymbol{m}_{\mathrm{i,R}}^{\mathrm{T}}$ of $\boldsymbol{M}_{\mathrm{R}}$. Based on Equation 2.11, the three-dimensional object position is calculated by

$${}_{(\mathrm{L})}\boldsymbol{P} = \boldsymbol{A}_{\mathrm{red}}^{+}\boldsymbol{b} \tag{2.12}$$

in a least square solution considering the Moore-Penrose pseudoinverse $\boldsymbol{A}_{\mathrm{red}}^{+}$.

## 2.4 Parallel Computing on the GPGPU

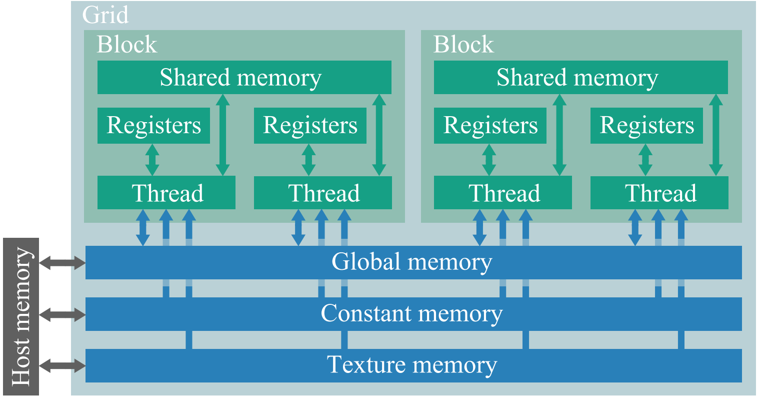


Figure 2.6: GPU architecture and available memory types.

In this dissertation, general-purpose computation on graphics processing units (GPGPU) based on the CUDA programming interface (Nvidia Corporation, Santa Clara, CA, US) had been deployed for implementing the image processing algorithms. Even though a GPU operates at lower frequencies as a CPU, a high level of parallelization can be achieved due to the immense number of processing cores, e.g., 2,688 for the Nvidia GeForce GTX Titan that had been utilized throughout this thesis.

A CUDA core provides a specific memory architecture, as shown in Figure 2.6. Data from the host, mostly a CPU, is copied to the read-only constant or texture memory, or to the global memory with read-write access. Those memory banks are accessible from all threads. Several threads associated to a block process data in parallel whereas the blocks are organized in a grid. If data are processed multiple times, registers or the shared memory provide an efficient solution, but their access is limited to the threads in the related block. Due to bandwidth limitations, multiple data transfer between the host and the GPU has to be avoided to fully exploit its computational power.

## 2.5 Surgical Laser Setup

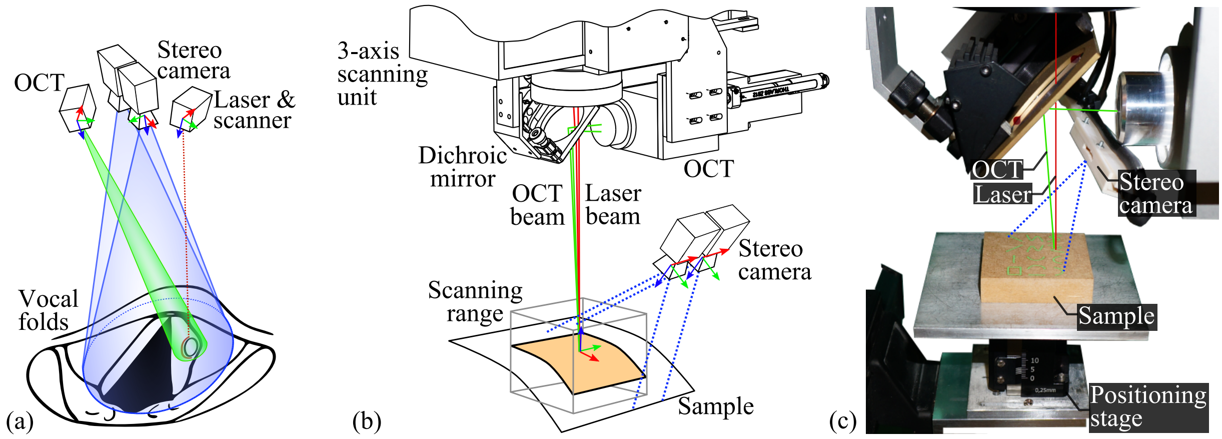


Figure 2.7: Configuration of stereo vision, laser and OCT (a). Related experimental setup is shown in (b,c).

Practical demonstration of the research presented in this thesis is provided with a multimodal setup. Referring to the scenario shown in Figure 2.7a, an ablation laser, an OCT, and a stereo vision system are aligned such that the field of view coincides with the laser and OCT scanning range.

The surgical setup shown in Figure 2.7b and 2.7c consists of an Er:YAG laser ($\lambda_{\mathrm{Er:YAG}} = 2.94\,\mu\mathrm{m}$, DPM-15, Pantec Engineering AG, Ruggell, Liechtenstein). The laser spot can be positioned in a cubic workspace of $10\,\mathrm{mm}$ in each direction by passing a galvanometer-based three-axis scanning unit (modules varioSCAN and hurrySCAN, SCANLAB, Puchheim, Germany) and a f-theta flat-field focusing lens [FPB$^+$15]. As discussed in Section 2.1, the Er:YAG laser represents a valuable alternative to conventional $CO_2$ lasers due to the superior ablation characteristics.

Moreover, the setup comprises a SD-OCT (Ganymede, Thorlabs Inc, Newton, New Jersey, USA) with a maximum scan range of $(15 \times 15 \times 2.76)\,\mathrm{mm}^3$ at a voxel definition of $512 \times 512 \times 1024$, resulting in a depth resolution of $2.69\,\mu\mathrm{m}$. A dichroic mirror couples the OCT beam into the laser optical path whereas a positioning stage aligns the OCT scan range with the laser focus.

Images are acquired with a stereo camera attached to this setup. Different imaging devices are discussed in Section 3.2. The laser and the camera are arranged non-coaxially while the field of view and the scanning range overlap at a distance of $30\,\mathrm{mm}$. Scene illumination is accomplished by a cold-light source (LB24 Solarc Light, Ushio America Inc., CA, USA).

# 3 Online Estimation of Tissue Surface Structure

Nowadays, laser microsurgery is performed under a magnified stereo view providing the surgeon with depth perception. On the one hand, the risk of unintended injury of the delicate vocal fold structure is lowered during tissue manipulation with instruments. On the other hand, stereo vision enables computational methods for three-dimensional tissue structure and motion estimation that are of interest when targeting vision-guided and robot-assisted laser surgery.

As the literature review in Section 1.2 has demonstrated, major investigations of the last two decades have been conducted with respect to surface reconstruction and tissue motion tracking in endoscopic surgery. In this context, a variety of toolboxes for benchmarking stereo algorithms, mainly with a focus on laparoscopic interventions, are available online. By contrast, evaluation of different stereo optical settings, i.e., for comparing endoscopic and microscopic vision, has not been addressed so far in literature. In particular in laser surgery with microspot manipulators, highly accurate positioning of the laser focus is required, otherwise, optimal ablation is not obtainable.

As a prerequisite to distance-based laser focus adjustment and to intraoperative fusion of stereo vision with OCT imaging, as described in Chapter 4, Section 3.1 presents a real-time method for surface reconstruction as shown in Figure 3.1. The algorithmic performance is assessed with respect to runtime and accuracy on *in vivo* image data. In Section 3.2, the reconstruction algorithm is considered for measurements with different stereo imaging devices in order to evaluate their applicability in laser microsurgery.

The methods and results of this chapter are in great part published in [SPK$^{+}$13, SKKO16].

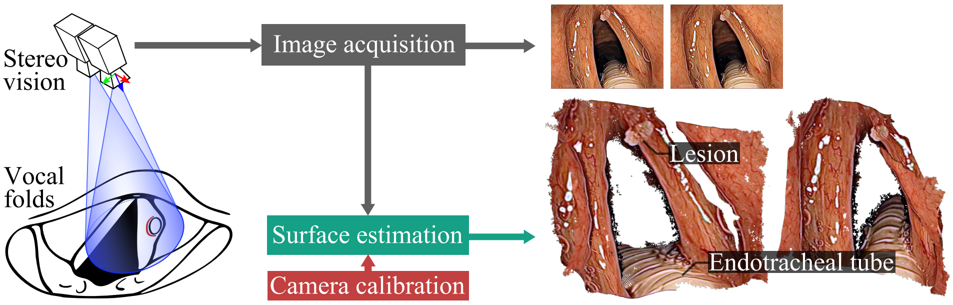


Figure 3.1: Stereoscopic surface estimation as exemplarily shown for a vocal fold situs. Image courtesy of Prof. Giorgio Peretti, Department of Otorhinolaryngology, University of Genoa, Italy.

## 3.1 Algorithmic Implementation and Evaluation

Even though stereo matching has been extensively discussed in literature, it is still an active research topic, i.e., in MIS with associated challenges such as specular highlights arising on glossy tissue or changing illumination of the surgical scene. This section presents the implementation and validation of a real-time method for robust surface reconstruction.

### 3.1.1 Stereo Matching Algorithm

Given a rectified stereo view, a point $\boldsymbol{p} = \boldsymbol{p}_{\mathrm{L}} = (u,v)^{\mathrm{T}}$ in the left image, that is regarded as the reference frame, is related to its corresponding point $\boldsymbol{p}_{\mathrm{R}} = \boldsymbol{p} - (d,0)^{\mathrm{T}}$ in the right view considering the disparity $d$. In other words, matches are found along the horizontal epipolar lines as depicted in Figure 3.2a. The stereo matching aims at estimating the correct disparity within the solution space $d \in \Omega_{\mathrm{d}}$, also known as disparity range. A three-dimensional matrix is constructed containing the similarity-based matching costs $C(\boldsymbol{p},d)$ to determine the optimal disparity

$$D(\boldsymbol{p}) \overset{!}{=} d_{\mathrm{opt}} = \arg \min_{d \in \Omega_{\mathrm{d}}} C(\boldsymbol{p},d) \tag{3.1}$$

for pixel $\boldsymbol{p}$. The cost volume is termed disparity space image (DSI). Establishing correspondence for entire image points yields a dense disparity map $D$ that has the same dimensions as the image.

According to the processing sequence depicted in Figure 3.2b, the stereo image pair is initially rectified based on stereo camera calibration. Subsequently, the images are census-transformed to compensate for different illumination or exposure settings between the two views. After pixel-wise matching and cost aggregation, the optimal disparity is found taking sub-pixel refinement, left-right consistency, and edge-preserving bilateral smoothing into account. Speckle detection and joint bilateral filtering are applied to detect and to refine mismatches, respectively.

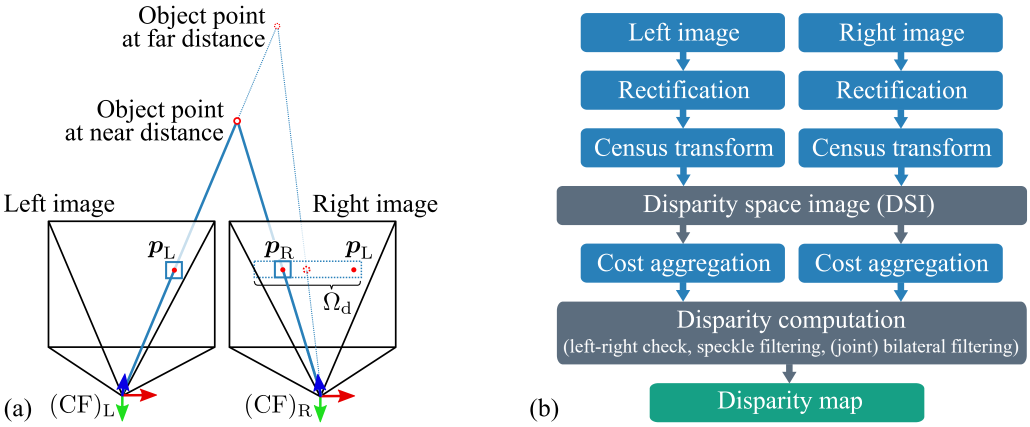


Figure 3.2: Stereo-based depth computation considering the disparity range $\Omega_{\mathrm{d}}$ is shown in (a). Related computation sequence is depicted in (b).

### Census Transform

State-of-the-art methods mostly rely on common similarity measures, i.e., sum of absolute differences (SAD) or normalized cross correlation (NCC). By contrast, this thesis adopts the non-parametric census transform and the Hamming distance that have been shown to robustly handle radiometric differences such as nonlinear illumination changes [ZW94, HS09, PN12]. The census transform $CT(\boldsymbol{p})$ encodes the relative ordering of the grayscale intensities $I(\boldsymbol{p})$ in a local neighborhood $\Omega_{\mathrm{CT}}$ of the center pixel $\boldsymbol{p}$, also denoted by support region, and is defined by

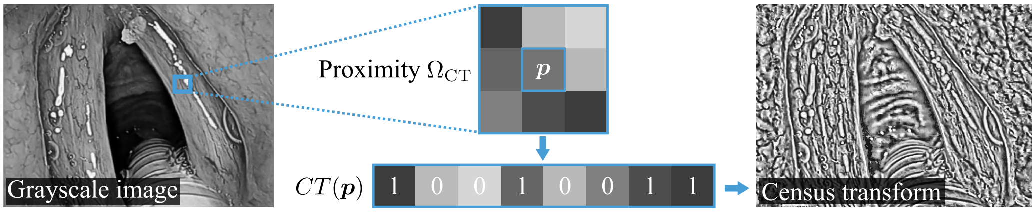


Figure 3.3: Computation and normalized representation of the census transform of a grayscale image.

$$CT(\boldsymbol{p}) = \bigotimes_{\boldsymbol{q}\in\Omega_{\mathrm{CT}}} \zeta\left(I(\boldsymbol{p}),I(\boldsymbol{q})\right) \tag{3.2}$$

with $\bigotimes$ denoting the concatenation to a bit string of the length $|\Omega_{\mathrm{CT}}| - 1$. The function $\zeta$ for compares the two intensities such that

$$\zeta(I_1,I_2) = \begin{cases} 0, & \text{if } I_1 \leq I_2 \\ 1, & \text{else.} \end{cases} \tag{3.3}$$

The census transform is computed for the left and the right image. For computational efficiency, the sparse census transform that considers every second pixel per row and column is applied. This reduction has been shown to be almost as efficient as the full transform [ZHAK08].

### Pixelwise Cost Computation

Subsequently, a pixelwise correspondence search is performed along the horizontal epipolar lines. For this purpose, the Hamming distance similarity measure is computed between the pixel $\boldsymbol{p}$ of the left transform $CT_{\mathrm{L}}$ and the point $\boldsymbol{p} - (d,0)^{\mathrm{T}}$ of the right transform $CT_{\mathrm{R}}$. The Hamming distance $H(\boldsymbol{p},d)$ at a certain disparity $d$ is given by

$$H(\boldsymbol{p},d) = \sum_{j=1}^{|\Omega_{\mathrm{CT}}|-1} CT_{\mathrm{L},j}(\boldsymbol{p}) \oplus CT_{\mathrm{R},j}(\boldsymbol{p} - (d,0)^{\mathrm{T}}), \tag{3.4}$$

where index $j$ defines the appropriate bit in the census string and the symbol $\oplus$ denotes the XOR operation. The Hamming distance can be computed in three different ways. A straightforward implementation to count the nonzero elements is bitwise shifting. Another efficient solution, denoted by the Wegner method, is to recursively perform a bitwise AND operation of the distance string with itself, but decremented by one [Weg60]. As a result, a string containing five nonzero elements will be iterated exactly five times, regardless of its length. A third option is provided by the CUDA __popc intrinsic function that counts the number of ones in a string [Nvi17]. It is directly called from the GPGPU device code.

#### Cost Aggregation

Smooth and consistent matching, i.e., in image areas with less texture information, is achieved by aggregating the pixelwise costs from a spatial neighborhood $\Omega_\mathrm{C}$ according to the function

$$C(\boldsymbol{p},d) = \sum_{\boldsymbol{q} \in \Omega_\mathrm{C}} H(\boldsymbol{q},d) \; . \tag{3.5}$$

However, aggregation over large areas is computationally expensive; thus, it usually conflicts with the real-time capability required for intraoperative application of the reconstruction algorithm. To overcome this limitation, this thesis addresses the implementation and evaluation of five aggregation strategies that are schematically illustrated in Figure 3.4.

Regarding the straightforward window-based, star-like access pattern, as shown in Figure 3.4a, the computation complexity grows with an increasing aggregation area $\Omega_\mathrm{C}$. Furthermore, every cost value is accessed multiple time due to the overlap of different windows. This results in an inefficient memory usage, i.e., if the global memory architecture of a GPGPU is utilized.

When applying the sliding window approach (see Figure 3.4b), the aggregation area is shifted through the cost matrix [FHM$^+$93]. In order to cover the costs in two dimensions, this method is split into an initial horizontal scan followed by a vertical scan. For each pass, only two memory read accesses are required per element; when it slips into (added to the costs) and when it leaves the aggregation area (subtracted from the costs). Even though the runtime is independent of the region size, the sequential processing pattern impedes parallelization on a GPGPU.

An alternative solution is provided by integral images that allow to rapidly aggregate the matching costs over a subregion whereas every pixel represents the sum of the elements above and to the left of it [Vek03]. As shown in Figure 3.4c, only four memory accesses are needed to compute the summation at constant time (two additions and two subtractions), regardless of the window size. Even though the cost aggregation is fast and highly parallelizable, the computational load is shifted to the computation of the integral image itself. Common computer vision libraries, such as OpenCV, often provide the functionality to compute integral images on the GPGPU.

Run time strongly depends on the memory access pattern. A work-efficient algorithm that is adopted in this thesis for computing integral images is called parallel prefix sum scan [HSO07]. As exemplarily illustrated in Figure 3.4d for the Hamming distances $H_0 ... H_3$, the aggregation

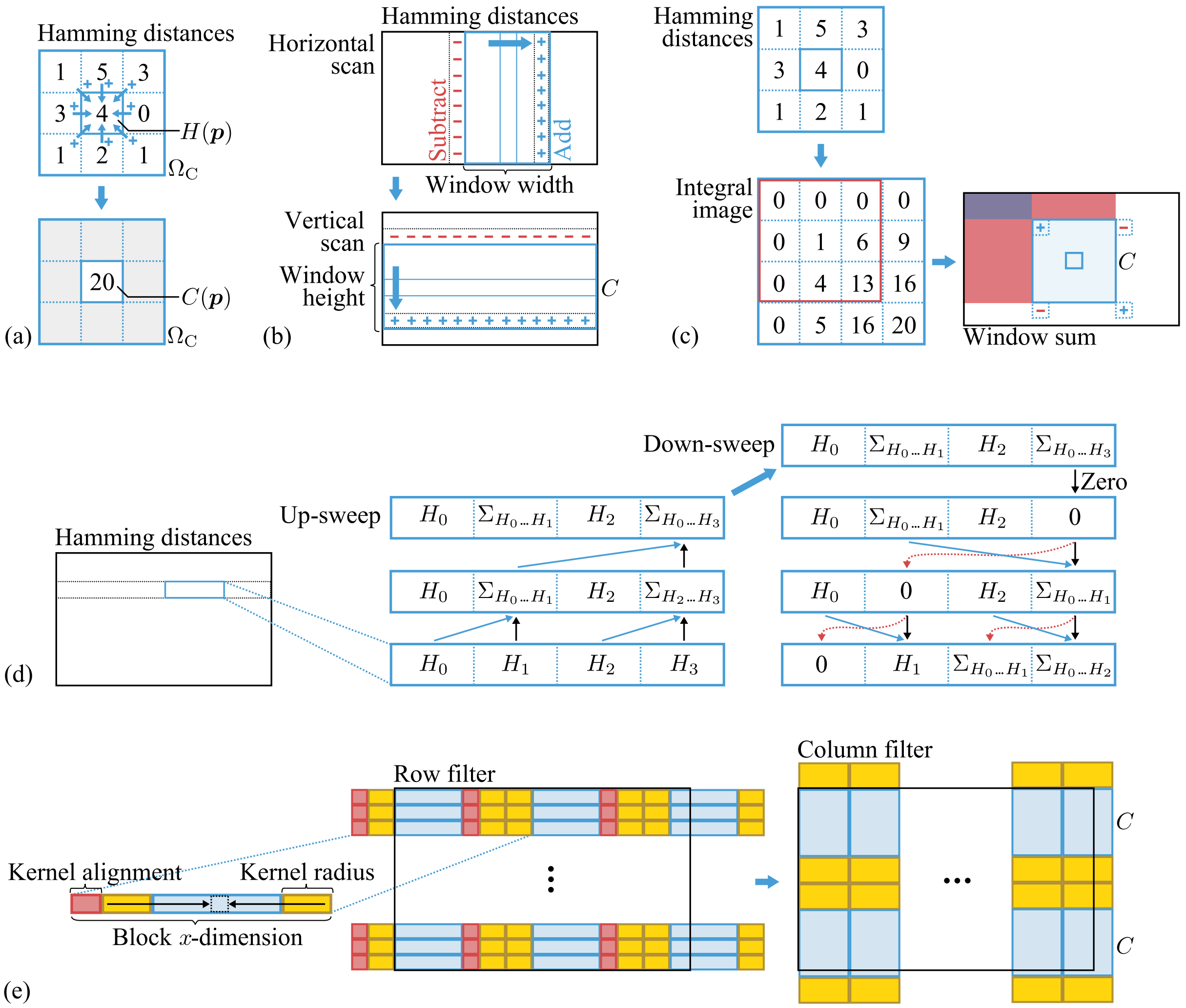


Figure 3.4: Cost aggregation strategies. Computing the sum of the Hamming distances in a rectangular region can be achieved by (a) the straightforward approach, (b) the sliding window method, (c) using integral images, (d) the parallel prefix sum scan, or (e) convolution with separable filters.

scheme consists of two phases: the up-sweep phase (commonly known as reduce phase) and the down-sweep phase. In the first phase a parallel reduction is performed, such that the last element holds the sum of the Hamming distances (see Figure 3.4d). Subsequently, the down-sweep takes the partial sums from the reduce phase and performs an exclusive scan (i.e., the total sum of all four Hamming distances is not included) where the zero propagates back to the first element; thus, the integral of the Hamming distance row is obtained. The two phases are performed row-by-row and then column-by-column, both with optimized access to the shared memory of the GPGPU [HSO07].

The cost aggregation can furthermore be addressed by separable convolution, as depicted in Figure 3.4e. The two-dimensional discrete convolution is defined by

$$C(\boldsymbol{p},d) = (F * H)(\boldsymbol{p},d), \tag{3.6}$$

where $H$ denotes the Hamming distance image and $F$ describes the box filter kernel containing ones. Considering a $3 \times 3$ pixel neighborhood $\Omega_\mathrm{C}$ as exemplarily illustrated in Figure 3.4a, the filter kernel $F$ can be separated into two consecutive one-dimensional convolution operations with kernel functions $F_1$ and $F_2$ given by

$$F_1 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \qquad \text{and} \qquad F_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} . \tag{3.7}$$

In especially for high performance GPGPU computation, separable filters drastically reduce the arithmetic complexity and bandwidth usage for the computation of each point [Pod07]. In this thesis, the idea of separable convolution is adopted to the cost aggregation process. As shown in Figure 3.4e, the Hamming distances are initially added in a horizontal pass (row filter) followed by a vertical pass (column filter) while being processed in the shared memory allowing for minimal access time. For each of the two scanning directions, there is an apron of pixels (yellow) that has the size of the kernel radius and that is required to process the entire image block (blue). As a consequence, the apron overlaps with adjacent regions that are loaded by another thread block of the GPGPU. Pixels outside the image borders are set to zero. However, there are additional (inactive) threads (red) required on the leading edge to meet global memory alignment constraints and thus to achieve the maximum bandwidth, i.e., for the row filtering step. During the vertical scan, additional threads are not necessary. For further details, the reader is kindly referred to [Pod07].

### Disparity Computation and Refinement

According to Equation 3.1, the optimal disparity is determined by the minimum value of the aggregated costs. To provide a continuous depth map, sub-pixel interpolation is achieved by fitting a parabola to the minimum and its neighboring costs. Moreover, a consistency check between the left and the right disparity map is performed to invalidate outliers, i.e., occlusions. Further mismatches are localized by adopting a parallelized connected component labeling, in the following denoted by speckle filtering, implemented on the GPGPU shared memory architecture [KRKS11]. Based on that, a speckle filter is derived that successfully removes disparity outliers.

For denoising of the disparity map $D$, bilateral filtering, as defined by

$$D_\mathrm{BF}(\boldsymbol{p}) = \frac{1}{W_\mathrm{BF}} \sum_{\boldsymbol{q} \in \Omega_\mathrm{BF}} G_\mathrm{s}(\|\boldsymbol{p} - \boldsymbol{q}\|)\, G_\mathrm{r}(|D(\boldsymbol{p}) - D(\boldsymbol{q})|)\, D(\boldsymbol{q}) \,, \tag{3.8}$$

provides an effective solution. In contrast to the conventional Gaussian weighting $G_\mathrm{s}$ (spatial kernel) that only considers the Euclidean distance, the bilateral filter also takes a weighting $G_\mathrm{r}$ (range kernel) based on the disparity difference into account. As a result, edge-aware smoothing, i.e., at object boundaries, is guaranteed as illustrated in Figure 3.5. Normalization is achieved by the term $W_{\mathrm{BF}}$ considering the sum of the weighting product in Equation 3.8.

After invalidating disparities by the left-right consistency check and the speckle filtering, they are iteratively refined with a joint bilateral filter that, in contrast to Equation 3.8, is guided by texture information of the image $I$ [ED04, PSA+04]. Valid depth information is recursively propagated from the spatial neighborhood considering the function

$$D_{\mathrm{JBF}}(\boldsymbol{p}) = \frac{1}{W_{\mathrm{JBF}}} \sum_{\boldsymbol{q}\in\Omega_{\mathrm{JBF}}} G_{\mathrm{s}}(\|\boldsymbol{p}-\boldsymbol{q}\|)\, G_{\mathrm{r}}(|I(\boldsymbol{p})-I(\boldsymbol{q})|)\, D_{\mathrm{BF}}(\boldsymbol{q}), \tag{3.9}$$

where the smoothed disparity map $D_{\mathrm{BF}}$ is consistently filled within a few iterations yielding the final disparity $D_{\mathrm{JBF}}$. Joint bilateral filtering is computationally complex not least due to the star-like memory access pattern, but also because of the nonlinear mathematical functions, i.e., the base exponential operation, that are applied to each point. To overcome this limitation, CUDA single precision (SP) floating-point intrinsics, e.g. for multiplication, division, or exponential operations, are utilized. Those are optimized low-level GPGPU hardware functions that are much faster than the conventional, mathematical library functions with a negligible loss of accuracy [Nvi17].

Even though applicable in one step, bilateral and joint bilateral filtering are split into two sequential operations for computational efficiency. In practice, the bilateral filter requires only a single iteration to sufficiently smooth the disparity estimate while the joint bilateral filter has to be repeated several times to fill larger gaps.

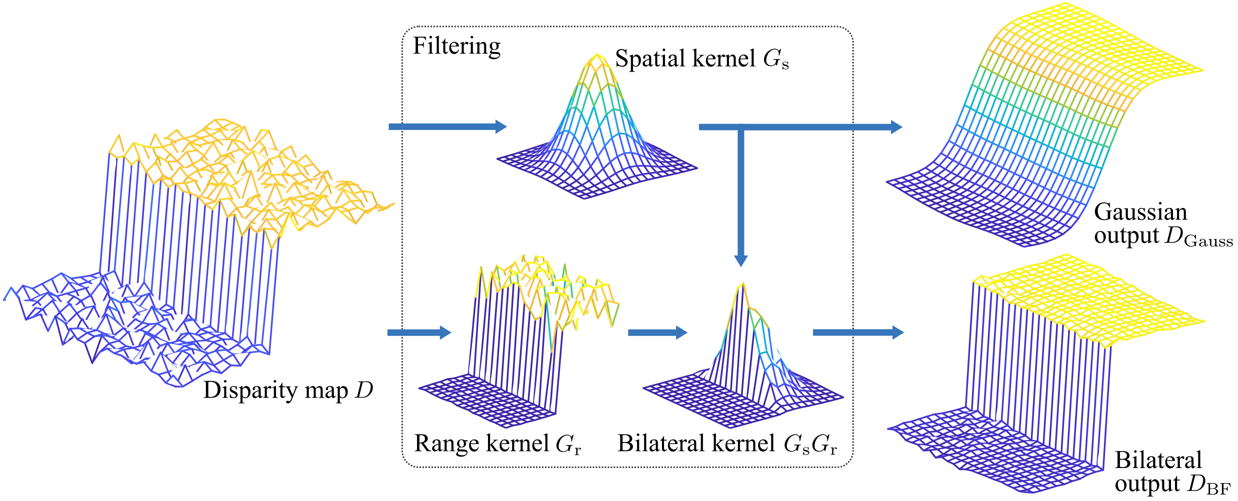

Filtering
Spatial kernel $G_{\mathrm{s}}$
Gaussian output $D_{\mathrm{Gauss}}$
Disparity map $D$
Range kernel $G_{\mathrm{r}}$
Bilateral kernel $G_{\mathrm{s}}G_{\mathrm{r}}$
Bilateral output $D_{\mathrm{BF}}$

Figure 3.5: Edge-aware disparity smoothness is achieved by bilateral filtering that, in contrast to the illustrated Gaussian smoothing, can successfully preserve depth discontinuities of the scene.

### 3.1.2 Results

In the following, the experimental design for runtime and accuracy assessment of the developed algorithm as well as the results of the algorithmic evaluation are presented.

### Accuracy Assessment on *In Vitro* Data

Since ground truth data for vocal fold surgery, such as depicted in Figure 3.1, is hard to acquire, reconstruction accuracy is evaluated on the *in vitro* laparoscopic dataset from the Open-CAS platform [MHGB$^+$14]. Ground truth data is provided by artificial landmark-based registration of surface information acquired from CT imaging.

The given calibration data is processed to rectify the 35 stereo images prior to disparity computation. Subsequently, the corresponding image points are triangulated to object space according to Section 2.3.4 and assessed with respect to the ground truth. The algorithm is parametrized with $9 \times 9$ pixels for the census window size (due to the sparse approach effectively $5 \times 5$ pixels), $15 \times 15$ pixels for the cost aggregation area, and a disparity range of 60 pixels. The disparity threshold for the left-right consistency check and speckle filtering is set to 0.3 pixels. Spatial smoothness and outlier compensation is achieved by bilateral filtering with a $7 \times 7$ kernel and joint bilateral filtering with a $3 \times 3$ kernel (in 2 iterations), respectively.

The reconstruction algorithm is assessed with the Open-CAS framework. This allows the comparison with two methods described in Section 1.2, denoted by UCL [SSPY10] and KIT [RBS$^+$12]. As exemplarily shown for dataset 14 (see Figure 3.6a and 3.6b), the developed method is able to provide a depth estimate of high density even in the presence of specular highlights and texture-less surface.

Figure 3.6c summarizes the root mean square error (RMSE) for each of the 35 images in the form of box plots [MTL78]. The interquartile range (IQR) comprises data points between the 25th and the 75th percentile and is defined by the bottom and top of the box, respectively, while the error median is represented by the notch. The upper whisker includes data within 1.5 times the IQR of the upper quartile, whereas the lower whisker contains data within 1.5 times the IQR of the lower quartile. Values are marked as outliers by a cross if not included between the whiskers.

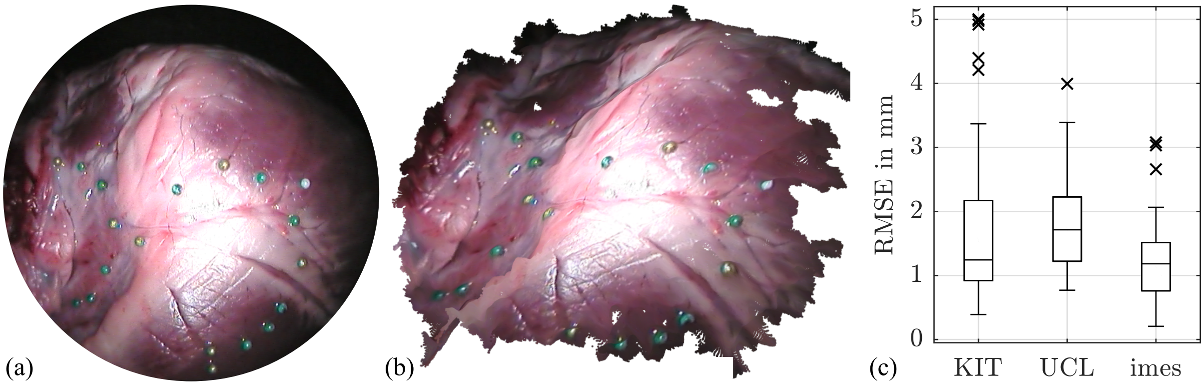


Figure 3.6: Results of the *in vitro* validation. Exemplarily, the left image of stereo pair 14 of the Open-CAS framework (from [MHGB$^+$14]) is shown in (a). Related surface points are depicted in (b). The boxplot in (c) illustrates the reconstruction error measured for entire image dataset. Results are compared to methods UCL [SSPY10] and KIT [RBS$^+$12].

The quantitative assessment reveals a high reconstruction accuracy. The developed algorithm provides an RMSE of $(1.27 \pm 0.68)$ mm (mean ± SD) and thus outperforms the two methods KIT and UCL yielding $(1.77 \pm 1.36)$ mm and $(1.83 \pm 0.79)$ mm, respectively. The reconstruction density of $(93.5 \pm 12.4)$ % is nearly as high as for the KIT method with $(95.8 \pm 10\,)\%$ and greater than for the UCL approach with $(92.6 \pm 16.8\,)\%$), while a reduced error variance is observed due to efficient removal of disparity outliers. In terms of runtime, the presented approach operates at 40 frames per second (at $720 \times 576$ pixels) whereas KIT and UCL only provide 14 frames per second (at $640 \times 480$ pixels) and one frame per second (at $720 \times 576$ pixels), respectively [MHGB$^+$14].

### Algorithmic Runtime Assessment

Accurate positioning of the laser focus on soft tissue requires the reconstruction algorithm to run at real time. Considering the different implementation strategies for computing the census transform and the Hamming distance, cost aggregation, and final disparity refinement, runtime performance is assessed for various parameter settings that are commonly used in practice, such as for the *in vitro* data described in the previous paragraph. A GeForce GTX Titan (Nvidia Corporation, Santa Clara, CA, USA) is deployed for GPGPU-based processing of the stereo image data.

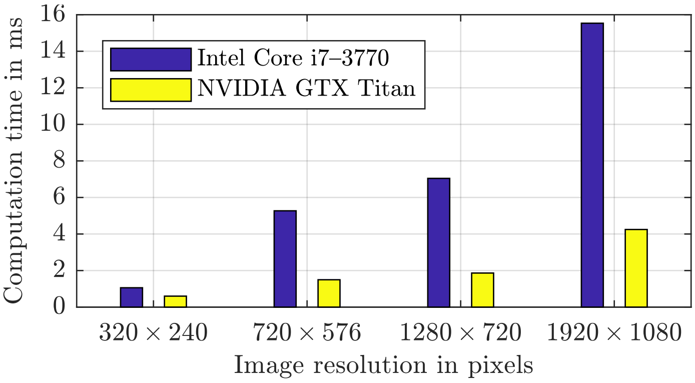


Figure 3.7: Computation time required for rectifying an image of the given resolution.

The potential of high-performance computing is illustrated in Figure 3.7 that compares the processing time required for image rectification measured for a CPU and a GPGPU implementation. Even though every pixel is accessed only once per acquired image frame, the GPGPU method outperforms the CPU-based computation by factor 3.6, i.e., at higher image resolutions. The speedup drastically increases if the data is processed multiple times; in that case, the runtime overhead due to memory allocation and data transfer between CPU and GPGPU can be neglected.

Regarding the census transform, computation requires less than 0.3 ms for an image pair of $720 \times 576$ pixels (see Figure 3.8a). Depending on the census support region size, the Hamming distance calculation requires up to 3.3 ms if the common bit shifting or even the Wegner method is utilized. Since bit shifting on the GPGPU is fast, there is no performance gain when using the

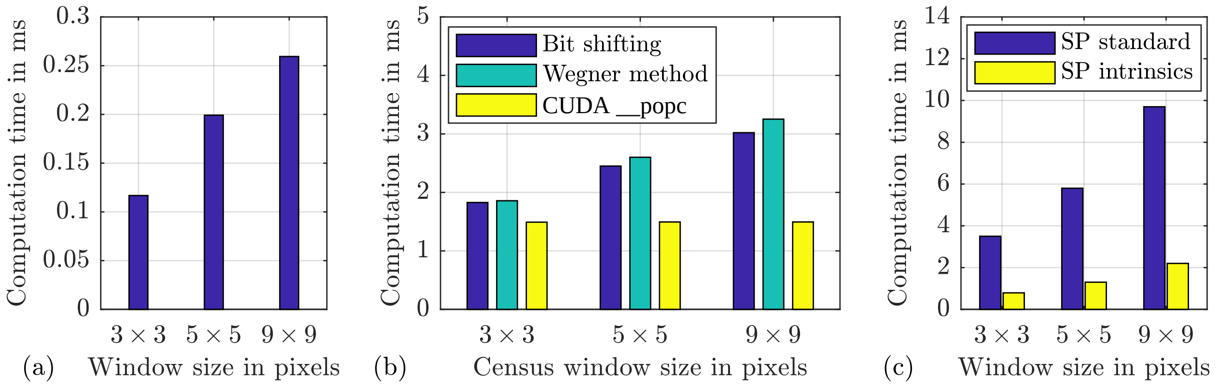


Figure 3.8: Computation time of (a) the census transform, (b) the Hamming distances, and (c) the joint bilateral filtering depending on the window size. The timing results apply for a stereo image pair from the Open-CAS dataset (720 × 576 pixels at 60 disparities).

Wegner method. By contrast, the runtime is kept constant at only $1.5$ ms in case of the CUDA intrinsic operation __popc, regardless of the window size (see Figure 3.8b).

The most time-critical routine is the cost aggregation. When applying the straightforward star-like access pattern (denoted by Straight-F) the computation time grows inadequately for large window sizes, as illustrated in Figure 3.9a. In comparison, constant aggregation time is achieved with the sliding window (Slid-Wnd), the OpenCV integral image (OCV-Int), and the parallel prefix sum scan (Par-Prefix). Related timing results are given by $(45.4 \pm 0.2)$ ms, $(61.2 \pm 6.8)$ ms, and

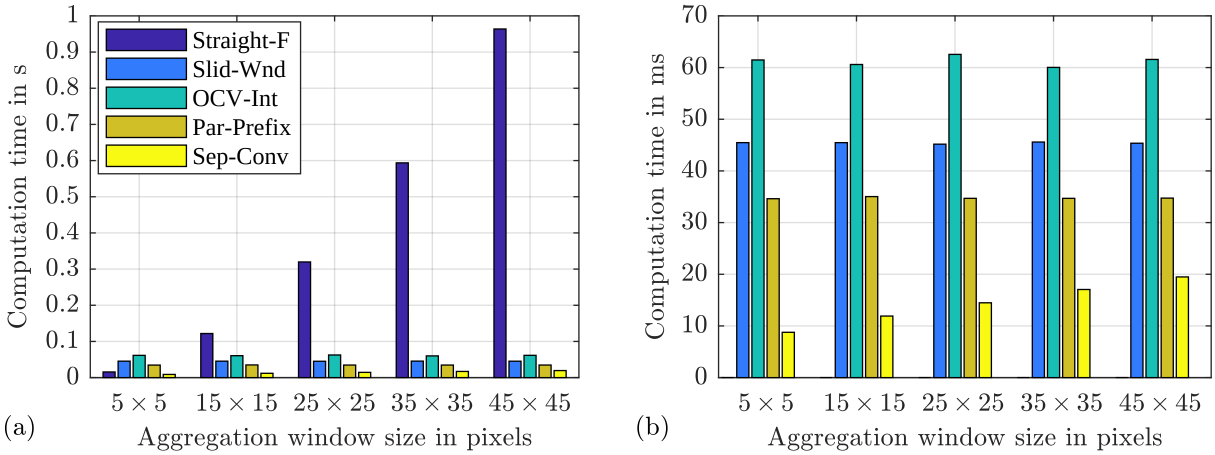


Figure 3.9: Computation time of the cost aggregation is shown in (a) for a stereo image pair from the Open-CAS dataset (720 × 576 pixels at 60 disparities). The results are presented for the straightforward method (Straight-F), the sliding window approach (Slid-Wnd), the OpenCV-based integral image computation (OCV-Int), the parallel prefix sum scan (Par-Prefix), and the separable convolution (Sep-Conv). A close-up view of (a) excluding the Straight-F method is given in (b).

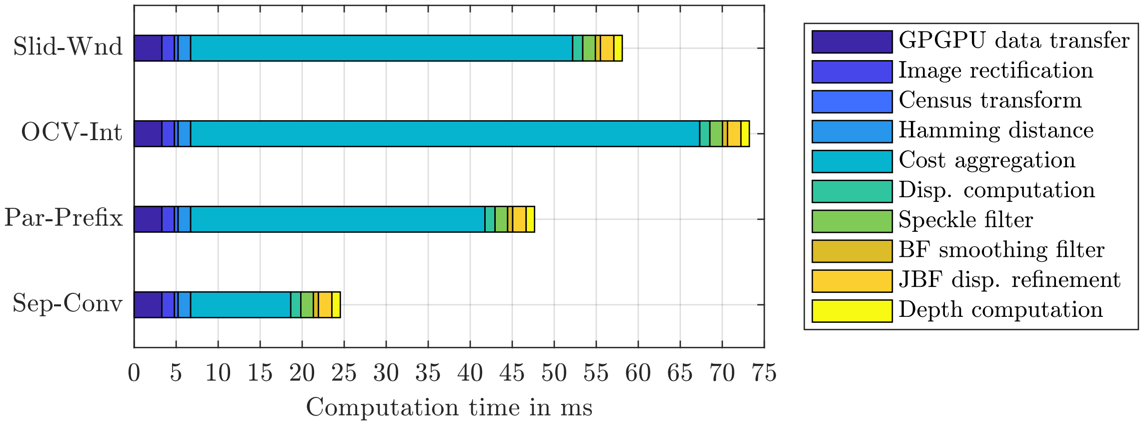


Figure 3.10: Computation time of the surface reconstruction algorithm considering cost aggregation for a $15 \times 15$ pixel window ($720 \times 576$ pixels at $60$ disparities).

$(34.8 \pm 0.9)$ ms, respectively. As shown in Figure 3.9b, the developed separable convolution scheme (Sep-Conv) drastically outperforms the other methods even though runtime depends on the window size. Regarding a support region of $15 \times 15$ pixels, as chosen for the *in vitro* validation, the costs are aggregated within $11.9$ ms yielding a speedup of factor $2.9$ compared to the second best method.

The post processing performance is exemplarily discussed for the joint bilateral filtering (see Figure 3.8c). For a $3 \times 3$ kernel, $3.5$ ms are required. In practice, the filter has to be applied iteratively to fill larger disparity gaps. Using the CUDA SP intrinsics reduces the computation time to $0.79$ ms and thus allows multiple repetitions of the joint bilateral filtering without significantly affecting the online capability of the matching.

A profiling of the overall computation times including the GPGPU data transfer, the image rectification, the optimized implementation for preprocessing and refinement as well as the final depth computation are summarized in Figure 3.10. The results clearly demonstrate superior performance of the separable convolution (Sep-Conv) approach that is capable to operate at $40$ frames per second. These findings motivate the online application of the matching scheme in laser microsurgery.

## 3.2 Reconstruction Accuracy of Stereo Imaging Devices for Microsurgery

This section evaluates different stereo-optical devices with respect to their applicability in laser microsurgery. In particular, the following questions are addressed:

1. Are stereo endoscopes and microscopes comparable in terms of reconstruction accuracy?
2. Can stereo imaging devices with small baselines provide accurate depth information?
3. How does the microscopic magnification contribute to reconstruction accuracy?

Answering those question has a high practical relevance and will help to select or design adequate stereo imaging devices that meet the accuracy requirements of the respective clinical application.

In practice, laser surgery systems with microspot manipulators do not provide real-time measurement and adjunct adaptation of the distance between tissue and laser focus. Thus, introducing computational stereo methods to vocal fold surgery is of high interest. As presented in Section 4.1, image-based assistance will allow to accurately adjust the laser focus onto the tissue surface in real-time in order to drastically reduce collateral tissue damage.

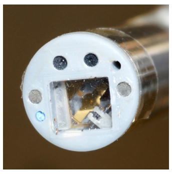
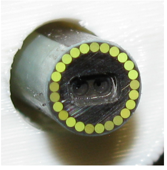
A direct comparison of the surface reconstruction accuracy obtained with two custom-made camera solutions (low cost, chip-on-the-tip) and a commercial endoscopic as well as microscopic system is presented. In detail, the analyzed camera systems are described including relevant parameters such as image resolution and stereo baseline. Subsequently, the study design of the stereo-based measurements by means of a reference object is described. In this context, the design of a custom-made sample and its measurement to acquire ground truth data is outlined. For quantitative evidence, the reconstruction errors based on image data from aforementioned imaging devices are discussed.

### 3.2.1 Stereo Imaging Devices for Microsurgery

As listed in Table 3.1, the following imaging devices for microsurgery are analyzed. The first custom-made stereo camera, denoted by µRALP, is part of a flexible endoscope laser phonomicrosurgery [KSKO15]. It embeds a low-cost chip-on-the-tip camera module that has a diameter of 4.8 mm. Two of those cameras can be positioned at a baseline of 4.3 mm due to a flat section on one side. By contrast, a commercial, passively bendable 3D endoscope VSii (Visionsense, Petach-Tikva, Israel) designed for neurosurgery is analyzed. This system is considered as alternative solution being integrated to flexible laser endoscopes such as described in [KSKO15]. It provides a camera baseline of only 1 mm (single chip, two apertures) at an outer diameter of 4.5 mm.

Those two endoscopic camera solutions are compared to microscopic imaging with three different magnifications, denoted by MS4×, MS8×, and MS16×, respectively. The stereo light microscope (Allegra 50, Möller-Wedel GmbH & Co. KG, Wedel, Germany) is dedicated to examinations and operations in the field of otorhinolaryngology. Two cameras are attached to the microscope (UI-3370CP-C-HQ, IDS Imaging, Obersulm, Germany), one for each optical path.

Table 3.1: Analyzed stereo imaging devices

| | <br>μRALP [KSKO15] | <br>SOM [BKS+15] | <br>VSii | <br>MS4× | MS8× | MS16× |
|---|---|---|---|---|---|---|
| Manufacturer | MISUMI<br>MO-BS0804P | SOMIKON<br>USB2.0 | Visionsense VSii<br>Black Tail Cobra | Haag-Streit Möller-Wedel Allegra 50<br>IDS Imaging UI-3370CP-C-HQ USB3.0 | | |
| Image definition in px* | 720 × 576 | 640 × 480 | 640 × 480 | 640 × 480 | 800 × 600 | 1200 × 1200 |
| Resolution in mm/px | 0.024 | 0.051 | 0.031 | 0.050 | 0.025 | 0.012 |
| Analyzed DOF* in mm | 21 . . . 31 | 33 . . . 41 | 26 . . . 34 | 250 ± 4 | 250 ± 2.5 | 250 ± 1 |
| Baseline in mm | 4.3 | 6.3 | 1.0 | | 25 | |
| Camera Ø in mm | 2×Ø4.8 | 2×Ø6.0 | Ø4.5 | | – | |

*Abbreviations: px – pixels, DOF – depth of field

Additionally, a second custom-made camera, denoted by SOM, is analyzed that is considered for fusion of stereo vision and OCT imaging, as addressed in Section 4.2. The camera module is characterized by increased lateral but reduced axial dimensions (flat design). The scene is illuminated with a cold-light source (LB24 Solarc Light,Ushio America Inc.).

#### Calibration of the Analyzed Cameras

In order to provide metric surface measurements, the intrinsic and extrinsic camera parameters have to be determined. In this context, one has to consider the optical design of modern stereo light microscopes having two optical paths that share one common main objective (CMO) aligning both views to the same focus. Due to eccentric arrangement of the two paths with respect to the CMO, a combined distortion with strong local gradients might be introduced that cannot be modeled by simple radial distortion functions [SGS04]. Literature provides solutions either by mathematically modeling CMO distortion [Dan99] or by distortion removal based on an interpolation model [SGS04]. However, reported accuracy improvement (max. 4.3 µm [Dan99]) is below the spatial resolution of the microscope listed in Table 3.1. For that reason and for simplicity of the calibration, the effect of CMO distortion is neglected. As a consequence, planar pattern-based calibration is applied to all analyzed cameras, as described on Section 2.3.3. A circular calibration grid with $9 \times 8$ dots and a spacing of 1 mm is used and approx. 10 to 15 images are acquired at different poses.

### 3.2.2 Accuracy Assessment

In the following, the experimental design and the measurement of the reference sample and of the reconstruction accuracy are described.

### Sample Design and Measurement

Providing ground truth from *in vitro* data, such as used for the algorithmic validation described in Section 3.1, requires a methodology of either tomographic imaging or surface scanning with a significantly higher resolution than offered by microscopic imaging. Instead, a sample of known shape and dimensions is used in order to objectively analyze reconstruction accuracy.

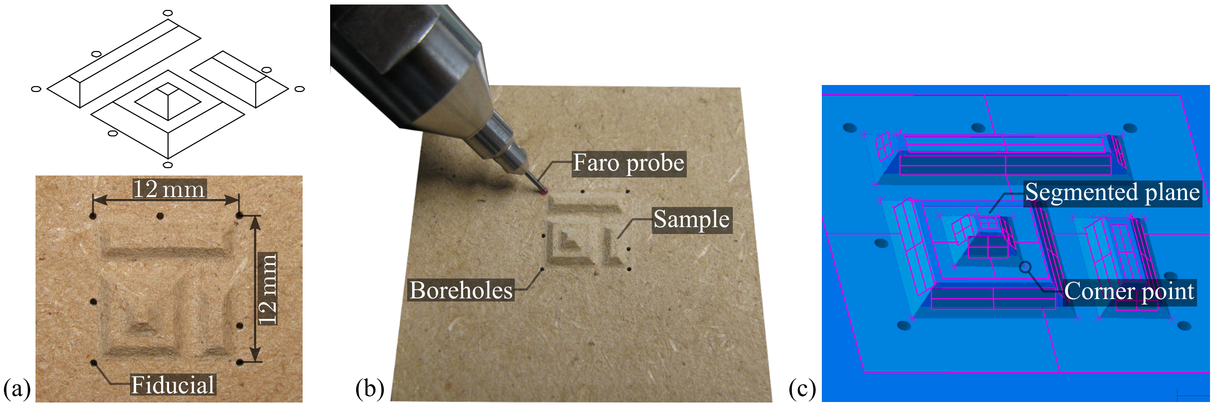


Figure 3.11: Design and photography of the sample are shown in (a). The shape is measured with a CMM as depicted in (b) for plane and corner segmentation required for the mesh update (c).

As depicted in Figure 3.11a, the specimen consists of several truncated pyramids with a step height of 1 mm in order to provide varying depth but also to be completely focused in a single shot image, i.e., at higher microscopic magnification. The outer contour is surrounded by seven boreholes located on a square of $(12 \times 12)$ mm$^2$ providing fiducials for image-based registration.

The specimen is made of medium-density fiberboards (MDF) and provides sufficient texture for dense surface reconstruction. It is fabricated with a CNC milling machine (FP3NC Dialog11, Deckel AG, Munich, Germany). Since the manufacturing tolerance is expected to be more than 20 µm (accuracy in as-new condition of the milling machine), a coordinate measuring machine (Faro Gage, Faro Technologies Inc., Lake Mary, FL, USA), abbreviated by CMM, is used in combination with a spherical probe (diameter of 1 mm) to determine ground truth (see Figure 3.11b).

For evaluating the CMM measurement uncertainty, the *Fiducial Localization Error* (FLE), that describes the Euclidean distance between true and measured marker positions, is estimated. Since the true locations are unknown and the CMM measurement contributes to the localization error, the FLE is estimated with a technique based on intramodal (IM) registration. This is achieved by fixating the sample on the granite surface plate of the CMM and by measuring the whole fiducial configuration that consists of $M = 7$ boreholes (see Figure 3.11b). This measurement is done $N = 10$ times with slightly varying arm configurations. Subsequently, the FLE is approximated by

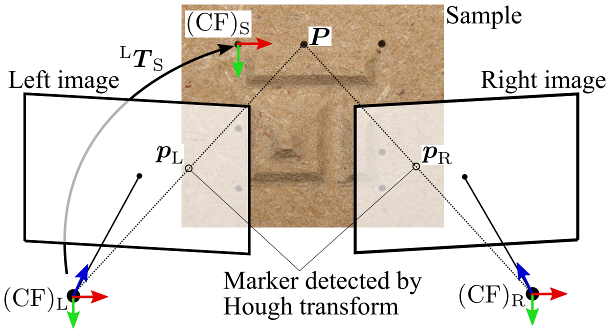
$$\mathrm{FLE} = \sqrt{\frac{M}{2K(M-2)} \sum_{i=1}^{K} \mathrm{FRE}_{\mathrm{IM}}^{2}(i)} \tag{3.10}$$

considering the *Fiducial Registration Errors* $\mathrm{FRE}_{\mathrm{IM}}(i)$ with $i = \{1, \ldots, K\}$ of the intramodal registration for all pairwise combinations ($K = 45$) of the fiducial sets [KDDF<sup>+</sup>14]. This metric quantifies the root mean square distance between the corresponding boreholes after registration. Although the $\mathrm{FLE}$ cannot be generalized, it correlates with the CMM measurement uncertainty.

Registration between the computer-aided design (CAD) and the CMM frame is performed with the mean fiducial position computed from its $N$ measurements. Then, the surface is sampled with the CMM probe ($n = 200$) and each pyramidal plane is segmented with the CMM toolbox. Following the assumption of planarity, the corners are obtained from the intersection of adjunct planes (see Figure 3.11c). The graphics software Blender (Blender Foundation, Amsterdam, Netherlands) is used to correct the triangular mesh, denoted by model update, and to export an STL-file.

### Marker Localization and Registration

In each stereo image pair, the fiducial positions are detected by the Hough transform for circles [IK88]. Prior to that, the images are up-sampled by factor $16$ applying cubic interpolation to increase the localization accuracy [KDDF$^+$14]. The two independently detected marker position $\boldsymbol{p}_{\mathrm{L}} = (u_{\mathrm{L}}, v_{\mathrm{L}})^{\mathrm{T}}$ and $\boldsymbol{p}_{\mathrm{R}} = (u_{\mathrm{R}}, v_{\mathrm{R}})^{\mathrm{T}}$ in the left and right view, denoted by frame $(\mathrm{CF})_{\mathrm{L}}$ and $(\mathrm{CF})_{\mathrm{R}}$, respectively, most likely do not satisfy the epipolar geometry constraint [HZ04]. To find an optimal solution for the object point $\boldsymbol{P} = (x, y, z)^{\mathrm{T}}$ (see Figure 3.12), linear least squares triangulation, as described in Section 2.3.4, is applied. Subsequent to this, rigid point-based registration with respect to the ground truth fiducial configuration is performed to compute homogeneous transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{S}}$ between the sample frame $(\mathrm{CF})_{\mathrm{S}}$ and the camera frame $(\mathrm{CF})_{\mathrm{L}}$ (see Figure 3.12). The registration error is quantified by $\mathrm{FRE}_{\mathrm{SL}}$, similar to $\mathrm{FRE}_{\mathrm{IM}}$, as described in the previous section.



$(\mathrm{CF})_{\mathrm{S}}$
$\boldsymbol{P}$
Sample
${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{S}}$
Left image
Right image
$\boldsymbol{p}_{\mathrm{L}}$
$\boldsymbol{p}_{\mathrm{R}}$
$(\mathrm{CF})_{\mathrm{L}}$
Marker detected by Hough transform
$(\mathrm{CF})_{\mathrm{R}}$

Figure 3.12: Triangulation of marker image position to object space.

### Experimental Work Flow

The measurement setup is exemplarily illustrated in Figure 3.13 for the VSii endoscope, but it applies for all cameras listed in Table 3.1. To entirely cover the focal range of the camera, the

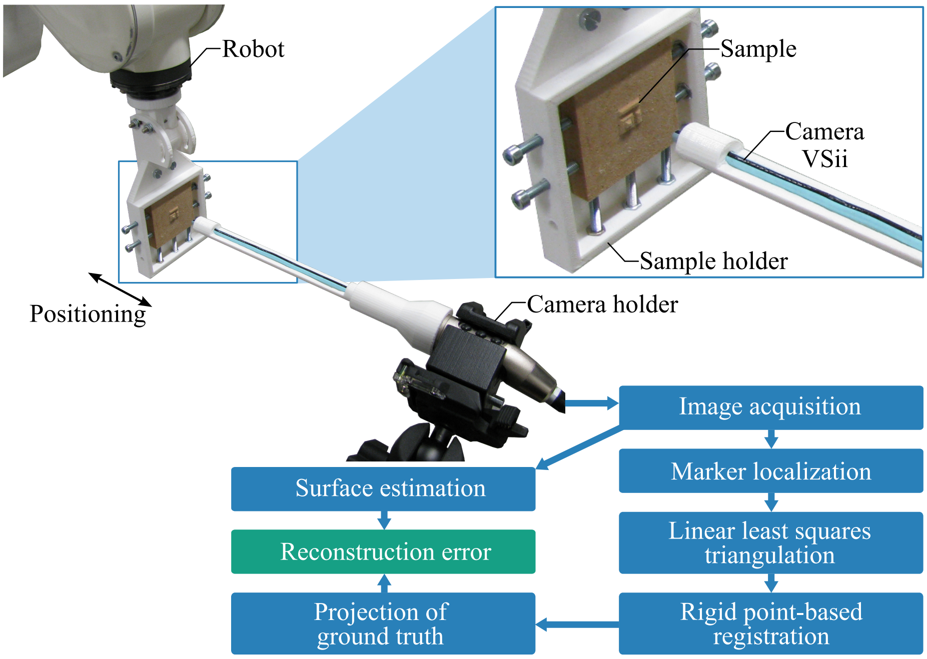


Figure 3.13: Experimental study design for surface reconstruction accuracy assessment. The sample is positioned with a robot while acquiring stereo images with the cameras listed in Table 3.1.

sample is positioned at varying distances (steps of 1 mm) using a robot (KR5sixx R850, KUKA Roboter GmbH, Augsburg, Germany). Subsequently, registration is performed as described before. To assess the reconstruction accuracy, pixel-wise correspondence is established by projecting the ground truth shape into the stereo camera view. Accomplishing that, the STL-file is uniformly upsampled with MeshLab (ISTI-CNR, Italy) to obtain a high-resolution ground truth point cloud [CCC$^{+}$08]. Ambiguity of the projection is solved by z-Buffering [Cat74]. In detail, pixel $\boldsymbol{p}_\mathrm{L} = (u_\mathrm{L}, v_\mathrm{L})^\mathrm{T}$ is assigned to both its reconstructed surface point ${}_{(\mathrm{L})}\boldsymbol{P} = \left({}_{(\mathrm{L})}x, {}_{(\mathrm{L})}y, {}_{(\mathrm{L})}z\right)^\mathrm{T}$ and to its projected ground truth position ${}_{(\mathrm{L})}\boldsymbol{P}_\mathrm{GT} = \left({}_{(\mathrm{L})}x_\mathrm{GT}, {}_{(\mathrm{L})}y_\mathrm{GT}, {}_{(\mathrm{L})}z_\mathrm{GT}\right)^\mathrm{T}$. The latter is obtained from aforementioned upsampling and mapping applying the registration transform ${}^\mathrm{L}\boldsymbol{T}_\mathrm{S}$. The pixelwise Euclidean reconstruction error $e_\mathrm{3D}$ is defined by

$$e_\mathrm{3D} = \left\| \, {}_{(\mathrm{L})}\boldsymbol{P}(\boldsymbol{x}_\mathrm{L}) - {}_{(\mathrm{L})}\boldsymbol{P}_\mathrm{GT}(\boldsymbol{x}_\mathrm{L}) \, \right\|_2 \, . \tag{3.11}$$

For comparison, the root mean square error (RMSE), the mean with standard deviation (abbreviated by SD), the median and the maximum error are computed.

Regarding the cameras µRALP, MS8×, and MS16×, the stereo matching algorithm has been parametrized equally to the conducted *in vitro* validation study discussed in Section 3.1. Due to significantly reduced spatial resolution (see Table 3.1), the cost window size for camera MS4× as well as for device SOM are reduced to $11 \times 11$ pixels to avoid over-smoothing and to maintain sharp edges. For the VSii endoscope, the cost and the bilateral filter kernel sizes are set to $19 \times 19$ and $15 \times 15$ pixels, respectively, due to increased image noise affecting reconstruction quality.

### 3.2.3 Results

This section initially presents the results of the sample measurement and the image-based registration. After discussing the µRALP camera accuracy, a comparison of all devices is provided.

#### Sample Measurement

The intramodal registration yields an $\mathrm{FLE} = 9.1\,\mu\mathrm{m}$. Even though this estimate of the CMM measurement uncertainty seems very small, it has been shown that such a small error is feasible if the measurements are performed in a small volume, here $(12 \times 12 \times 2)\,\mathrm{mm}^3$, while just slightly varying the CMM arm configuration (up to $5\,\mu\mathrm{m}$ are reported [COOS10]). As illustrated in Table 3.2, the initial specimen measurement error is quantified by $(17.2 \pm 11.6)\,\mu\mathrm{m}$. Subsequent to plane segmentation and model update, the error is reduced to $(9.4 \pm 7.6)\,\mu\mathrm{m}$. This is approved by the additional validation measurement yielding an error of $(8.9 \pm 6.9)\,\mu\mathrm{m}$. Due to the correlation with the $\mathrm{FLE}$, the CMM is considered as a reliable tool for ground truth acquisition.

Table 3.2: Sample surface error before and after the model update ($n = 200$ data points).

| Measurement | RMSE in µm | mean in µm | SD in µm |
|---|---|---|---|
| Initial model | 20.8 | 17.2 | 11.6 |
| Updated model | 12.1 | 9.4 | 7.6 |
| Validation | 11.2 | 8.9 | 6.9 |

#### Camera Calibration and Image-Based Registration Results

The calibration results are listed in Table 3.3. Except for MS$16\times$, the errors are below $0.3$ pixels. However, MS$16\times$ provides a high spatial resolution; thus, its error is considered to be adequate.

The registration results are also summarized in Table 3.3. The average $\mathrm{FRE}_{\mathrm{SL}}$ is below $0.1$ mm, except for camera VSii. Since it has a small baseline, triangulation of the localized markers is very sensitive to image noise resulting in a reduced depth resolution compared to the other devices.

Table 3.3: Camera calibration re-projection error and average registration error $\mathrm{FRE}_{\mathrm{SL}}$.

| | µRALP | VSii | MS$4\times$ | MS$8\times$ | MS$16\times$ | SOM |
|---|---|---|---|---|---|---|
| RMSE in pixels | 0.25 | 0.19 | 0.28 | 0.29 | 0.65 | 0.16 |
| $\mathrm{FRE}_{\mathrm{SL}}$ in mm | 0.088 | 0.297 | 0.091 | 0.081 | 0.084 | 0.050 |

### Reconstruction Accuracy with the µRALP Camera

Prior to the overall comparison of the camera devices, the results obtained with the custom-made µRALP stereo camera are briefly discussed. In Figure 3.14, the reconstruction error is shown in the form of boxplots for a distance range of 21 to 40 mm. Measurements below a distance of 21 mm are not possible with the designed specimen since its lateral dimensions would exceed the field of view.

Regarding the mentioned distance range, the spatial resolution of the camera drastically drops from 0.024 mm/pixel (at 21 mm) to 0.047 mm/pixel (at 40 mm). The decrease from initially 214,763 (at 21 mm) to only 37,409 (at 40 mm) reconstructed surface points underlines this observation. Furthermore, objects at a distance greater than 31 mm are no longer in the camera focus. As a consequence, the camera's depth of field (DOF) is defined from 21 to 31 mm (see Figure 3.14). In this range, the overall RMSE amounts to 0.094 mm (see Table 3.4).

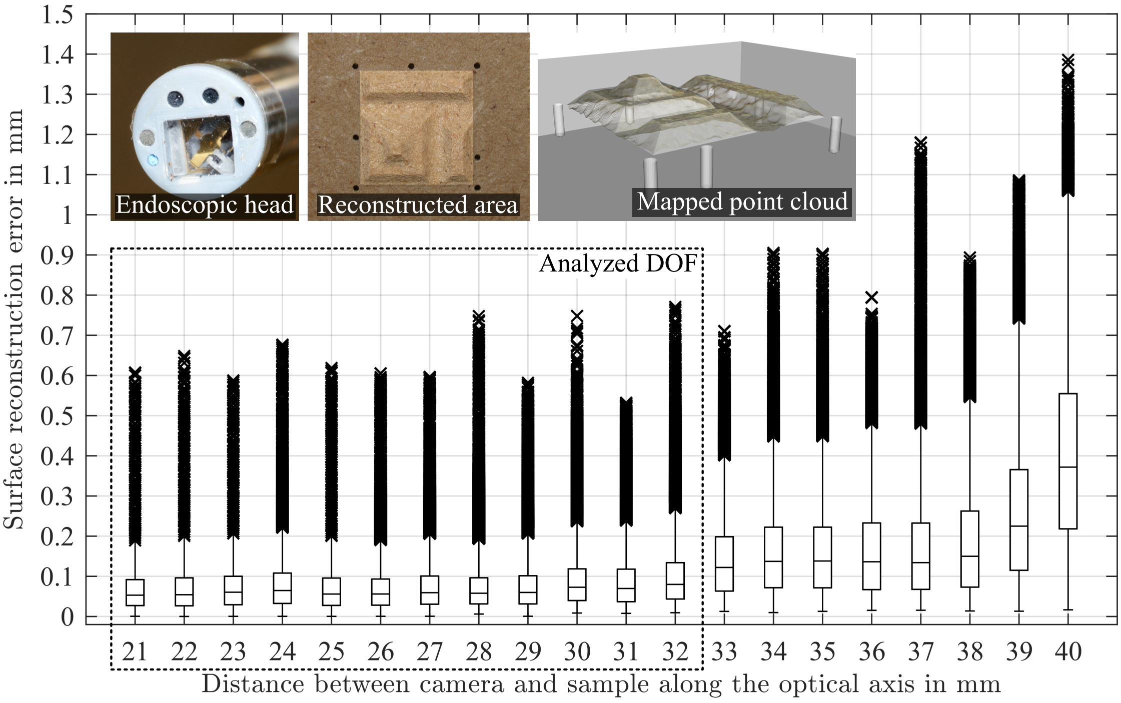


Figure 3.14: Box plot of the µRALP camera reconstruction results. Outliers are marked by ×.

### Overall Comparison of the Cameras

Similar to the µRALP camera, the reconstruction accuracy is assessed for each camera in its DOF. The results are summarized in Figure 3.15 and related quantitative values are provided in Table 3.4.

The overall RMSE is below 0.18 mm while the microscopic setting MS16× yields the most accurate measurement mainly due to its high spatial resolution and its proximally located low-noise CMOS sensor. Similar to the image-based registration results, the camera VSii provides less accurate reconstruction compared with the other cameras due to its small baseline and thus, limited depth resolution. This observation is underlined by the maximum distance error of 1.2 mm.

In comparison to that, the maximum error of the microscope settings does not exceed 0.5 mm (see Table 3.4). When considering the median error, the custom-made cameras µRALP and SOM perform equivalent to MS8× and MS4× even though they produce outliers of higher magnitude, potentially due to increased CMOS noise. Because of this, it seems to be obvious to compare µRALP with MS8× and SOM with MS4×. Furthermore, it needs to be clarified whether the higher magnification MS16× outperforms MS8×. Having a closer look onto the results, it can be found that the data do not meet the assumption of normality. Due to outliers, the computed error distributions of all cameras show a positive skew (non-symmetric distribution). Although the medians and

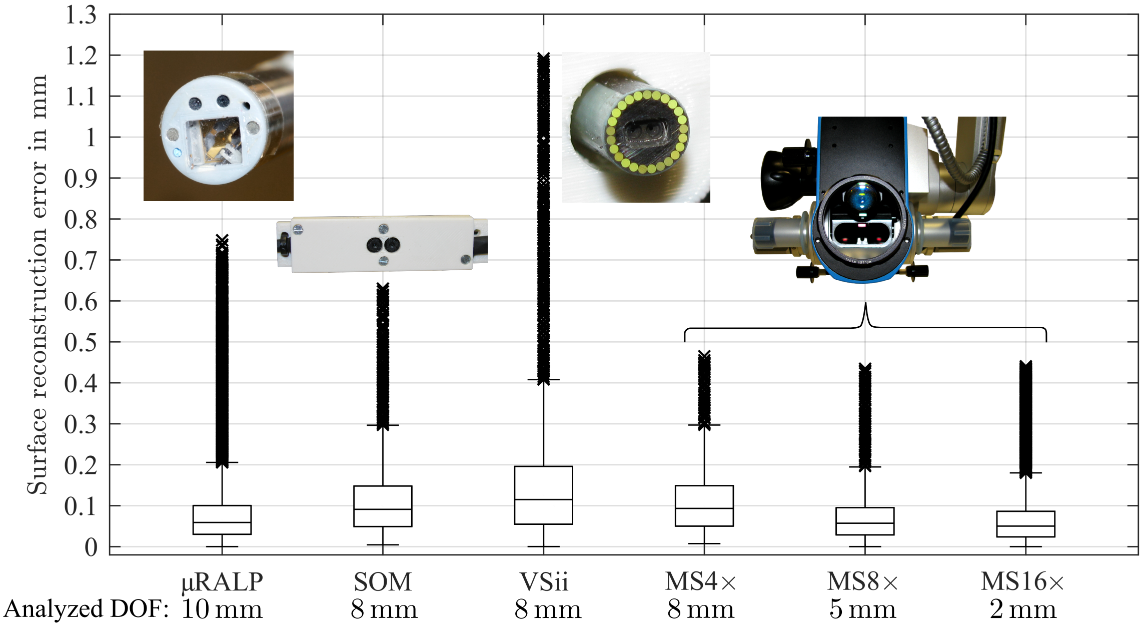


Figure 3.15: Box plot of the reconstruction results for each camera. Outliers are marked by the symbol ×.

Table 3.4: Reconstruction error for analyzed stereo imaging devices.

| | µRALP | SOM | VSii | MS4× | MS8× | MS16× |
|---|---|---|---|---|---|---|
| RMSE in mm | 0.094 | 0.128 | 0.176 | 0.127 | 0.083 | 0.077 |
| Mean* in mm | 0.074 | 0.107 | 0.139 | 0.107 | 0.068 | 0.061 |
| SD* in mm | 0.058 | 0.070 | 0.108 | 0.068 | 0.049 | 0.047 |
| Median in mm | 0.059 | 0.091 | 0.115 | 0.093 | 0.057 | 0.050 |
| Lower quartile in mm | 0.029 | 0.049 | 0.055 | 0.050 | 0.029 | 0.024 |
| Upper quartile in mm | 0.099 | 0.148 | 0.196 | 0.149 | 0.095 | 0.086 |
| Max. in mm | 0.748 | 0.632 | 1.192 | 0.465 | 0.438 | 0.440 |
| Outlier rate in % | 3.3 | 1.5 | 2.3 | 1.3 | 1.9 | 2.1 |
| Avg. runtime in ms | 25 | 20 | 20 | 20 | 27 | 70 |

*Values are provided for further comparison but data basis is not normally distributed (non-symmetric).

the interquartile ranges are similar when comparing µRALP with MS8×, SOM with MS4×, and MS8× with MS16×, applying the non-parametric Brown-Forsythe test for skewed distributions reveals statistically unequal variances in all three cases (significance of $p = 0$, $p = 1 \times 10^{-15}$ and $p = 0$) [BF74]. In other words, the amount and the magnitude of outliers strongly determine the quality of the estimated surface. In this context, the microscopic setting produces results of higher reliability than equivalent custom-made cameras. Furthermore, there is a significant difference between magnification setting MS8× and MS16× although quantitative difference seems relatively small.

To illustrate these findings, Figure 3.16 depicts the color-coded distance errors for each camera. It clearly reveals significant deviations between the reconstructed surface and associated ground truth data. For instance, the µRALP-based reconstruction corresponds well to the surface obtained with MS8× but comprises outliers of higher magnitude (see Figure 3.16a and 3.16e).

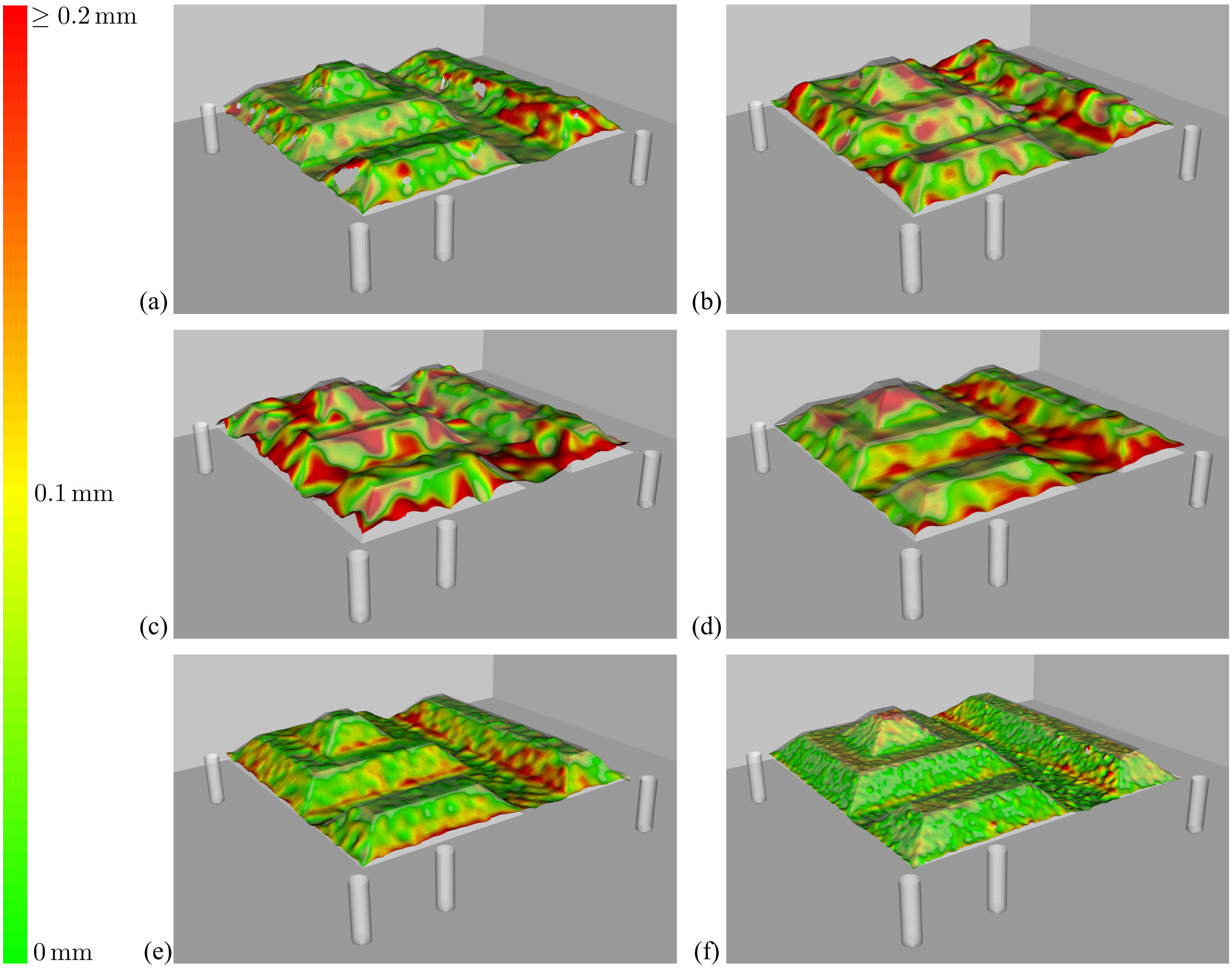


Figure 3.16: Color-coded distance map of the reconstructed surface (examples for each camera) obtained from (a) µRALP, (b) SOM, (c) VSii, (d) MS4×, (e) MS8×, and (f) MS16×. The color bar to the left specifies the magnitude of the distance error with respect to the ground truth.

In terms of computational efficiency, all settings except for MS$16\times$ guarantee average matching rates of at least $37$ frames per second (runtime of $27$ ms, see Table 3.4). The maximum frame rate achieved is $50$ frames per second (runtime of $20$ ms) for the cameras SOM, VSii and MS$4\times$. The configuration MS$16\times$ runs at $14$ frames per second and thus, near real time is achieved.

Regarding the focus requirements of laser microsurgery, the microscopic setting at higher magnification most likely provides the best option for distance-based laser focus adjustment (see Figure 3.16d–f). However, the sub-millimeter accuracy of the µRALP and the SOM camera reveals that both sensors have the potential to be integrated into a miniaturized laser system that provides focus adaptation based on tissue surface information. To achieve this, further effort has to be made with respect to elimination of outliers. By contrast, ablation might be less efficient if the miniaturized, bendable endoscope VSii is utilized as vision system. As illustrated especially in Figure 3.16c, the reconstructed surface is characterized by severe outliers (exceeding $1$ mm) that potentially impede distance-based focus adjustment.

## 3.3 Conclusion

In this chapter, a stereo-based surface reconstruction algorithm that is robust to illumination changes and disparity outliers, has been developed and evaluated. Real-time performance of the time-critical cost aggregation step is achieved by implementing an efficient separable convolution scheme on the GPGPU shared memory architecture. The method is validated on *in vitro* images taken from the Open-CAS database demonstrating that two state-of-the-art methods are outperformed in terms of runtime and accuracy.

In addition, a direct comparison of surface reconstruction applied to different stereo-optical settings is presented. The findings reveal that reconstruction accuracy correlates with the spatial resolution of the camera. The most accurate results have been achieved with the microscopic configurations, i.e., at higher magnification, and with low-cost, custom-made camera setups. To be more specific, the two cameras µRALP and SOM yield reconstruction accuracies in sub-millimeter scale. In particular, the µRALP sensor provides an increased DOF compared to the similar setting MS$8\times$ and thus strongly motivates its integration into an endoscopic laser scalpel. This camera solution meet the requirements of distance-based spot and focus positioning onto the tissue surface. By contrast, the bendable endoscope VSii with its small baseline and thus, limited depth resolution, has to be treated with caution when being considered for application in laser microsurgery. The expected ablation quality might suffer from severe reconstruction outliers. However, VSii provides a good tradeoff between accuracy, outer dimensions of optical components, and accessibility to surgical sites. In particular, it has by far the smallest diameter and thus, facilitates to access cavities that are hard-to-reach or that cannot be examined by direct-line-of-sight as required for microscopes (e.g. natural orifice transluminal endoscopic surgery).

# 4 Registration of Stereo Vision with Laser and OCT

The state-of-the-art TLM technology does not provide intraoperative measurement to guide the surgeon and to automatically adapt the laser focus in real-time with respect to the surgical scene. Thus, surgeons prevalently tolerate greater spot sizes to the benefit of an increased depth of focus as long as adequate (but not optimal) cutting efficiency can be achieved. However, this conflicts with procedures that strongly require microspot scanning, e.g., the resection of vocal fold nodules. Here, the collateral thermal damage to the vocal fold ligaments that are just a few hundred micrometers below the topmost epithelium layer has to be minimized.

Addressing this challenge, Section 4.1 presents a methodology to incorporate tissue surface information, that are computed according to Section 3.1, into the Er:YAG laser setup of Section 2.5. Based on a trifocal model and registration of laser and stereo camera, distance-based focus adjustment, as indicated in Figure 4.1a, is demonstrated by ablation on *ex vivo* tissue. Furthermore, the ablation quality is discussed considering OCT-based analysis of the cutting width and depth.

In addition to using OCT for ablation monitoring, Section 4.2 proposes a two-step registration of stereo vision and OCT data. As shown in Figure 4.1b, this allows for segmenting and visualizing subepithelial vocal fold lesions (see Chapter 5). Considering the OCT setup described in Section 2.5, a reference sample with a distinct shape is scanned while being reconstructed by the stereo-based algorithm. The registration is performed taking salient surface features into account. The results are discussed including two concepts for visualizing the OCT scan aligned to the live camera view.

The methods and results of this chapter are published in [SKK$^+$15, BKS$^+$15].

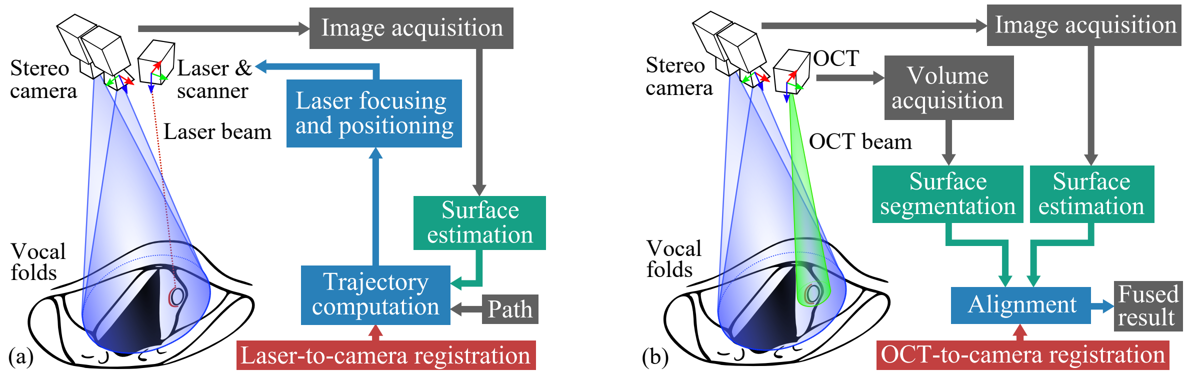


Figure 4.1: Concept for integrating (a) the ablation laser and scanner as well as (b) the OCT into a surgical stereo vision setup.

## 4.1 Tissue Surface Information for Distance-Based Laser Focus Adjustment

A variety of endoscopic and robot-assisted approaches to laser surgery exists. However, establishing consistent image-based alignment of the laser focus with the tissue surface has not been discussed so far even though improper focus adaptation poses the risk of an increased thermal damage zone.

A solution overcoming this limitations is provided by computational stereo methods. If the ablation systems comprises an additional aiming laser, stereo-optical triangulation of the visible spot yields its position in object space. However, the detection and the tracking of the spot on glossy tissue is a challenging task, i.e. at high scanning speed. Thus, this dissertation proposes an alternative approach based on dense surface reconstruction and laser-to-camera registration. The estimated depth information does not only allow for online focus adaptation, but also for intraoperative navigation and augmented reality.

In this section, a work flow for distance-based laser focus adaptation by means of tissue surface information is introduced. A methodology for laser-to-camera registration is presented and demonstrated by laser workspace highlighting in the live view as well as by laser ablation trials. Furthermore, experiments are conducted on tissue substitutes and *ex vivo* biological tissue. Special emphasis is given to quantitative analysis of the ablation width and depth. The results demonstrate the potential of the proposed image-based laser focus adjustment for being integrated into endoscopic or stereo-microscopic laser surgery.

### 4.1.1 Trifocal Model of Stereo Vision and Laser

According to the data flow in Figure 4.1a, the surgical scene is observed with a stereo camera, that might be either endoscopic or microscopic, while scanning the tissue surface with an ablation laser. This trifocal configuration facilitates three-dimensional laser spot positioning and the computation of a virtual laser view for ablation planning from the laser's perspective.

Another feature is laser workspace rendering directly in the live view, i.e., superimposing the intersection with the surgical site, as shown in Figure 4.2a. This instantly provides visual feedback on the scanning and focusing range, i.e., when aligning the laser system with respect to the target.

#### Trifocal Setup

The trifocal model is established for the laser setup outlined in Section 2.5. Regarding scene depth estimation, the miniaturized stereo camera module µRALP from Section 3.2 is attached non-coaxially to the laser. The maximum overlap of the camera field of view and the laser workspace is achieved at a distance of approx. 30 mm with respect to the camera. Simple incision planning is performed directly in the live video with a graphics tablet (Wacom Co. Ltd.). The laser interface, image acquisition and processing are implemented with the Robot Operating System (ROS) as high-level control layer [QGC$^{+}$09]. Furthermore, specimens are positioned using a manually adjustable table with millimeter scale.

OCT measurements are adduced for quantitative evidence of the incision quality and additional qualitative analysis is accomplished by microscopic examination (Stereo Discovery.V8, Carl Zeiss Microscopy). Furthermore, quantitative validation of the laser-to-camera registration, i.e., the lateral deviation between planning and ablation, is conducted by pose estimation of an ablated planar pattern with respect to the calibrated left camera.

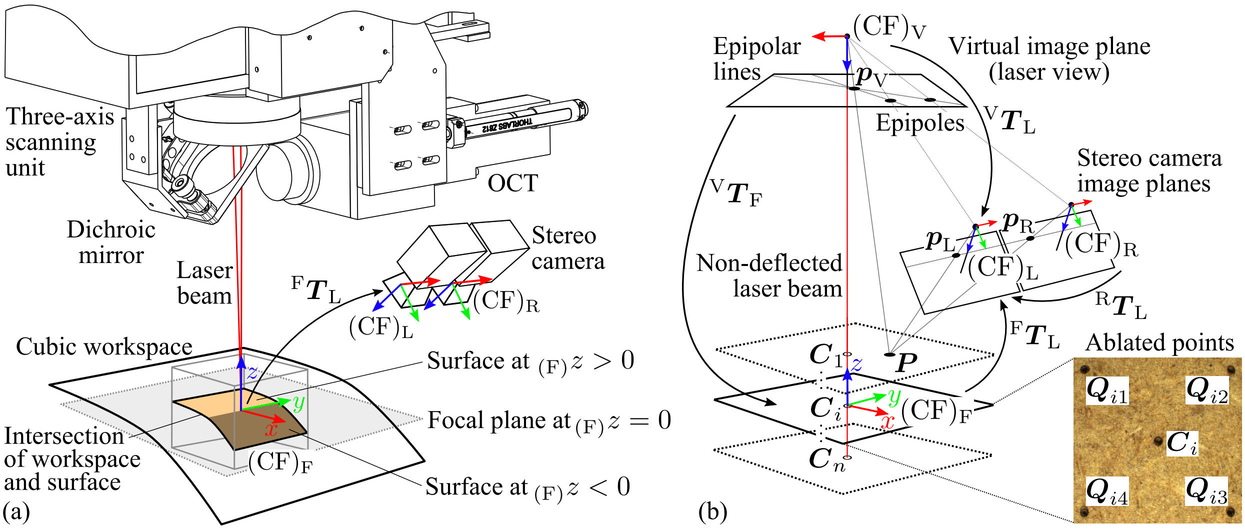

Three-axis scanning unit
OCT
Dichroic mirror
Laser beam
${}^{\mathrm{F}}\boldsymbol{T}_\mathrm{L}$
Stereo camera
$(\mathrm{CF})_\mathrm{L}$
$(\mathrm{CF})_\mathrm{R}$
Cubic workspace
Surface at ${}_{(\mathrm{F})}z > 0$
Focal plane at ${}_{(\mathrm{F})}z = 0$
Surface at ${}_{(\mathrm{F})}z < 0$
Intersection of workspace and surface
$z$
$y$
$x$
$(\mathrm{CF})_\mathrm{F}$
(a)
Epipolar lines
$(\mathrm{CF})_\mathrm{V}$
$\boldsymbol{p}_\mathrm{V}$
Virtual image plane (laser view)
Epipoles
${}^{\mathrm{V}}\boldsymbol{T}_\mathrm{L}$
${}^{\mathrm{V}}\boldsymbol{T}_\mathrm{F}$
Stereo camera image planes
$\boldsymbol{p}_\mathrm{L}$
$\boldsymbol{p}_\mathrm{R}$
$(\mathrm{CF})_\mathrm{R}$
$(\mathrm{CF})_\mathrm{L}$
Non-deflected laser beam
${}^{\mathrm{R}}\boldsymbol{T}_\mathrm{L}$
${}^{\mathrm{F}}\boldsymbol{T}_\mathrm{L}$
Ablated points
$\boldsymbol{C}_1$
$\boldsymbol{P}$
$\boldsymbol{C}_i$
$\boldsymbol{C}_n$
$(\mathrm{CF})_\mathrm{F}$
$\boldsymbol{Q}_{i1}$
$\boldsymbol{Q}_{i2}$
$\boldsymbol{C}_i$
$\boldsymbol{Q}_{i4}$
$\boldsymbol{Q}_{i3}$
(b)

Figure 4.2: The setup for experimental validation of laser focus adaptation and laser workspace highlighting is shown in (a). Related laser-to-camera registration facilitating image based incision planning and a virtual laser view are depicted in (b).

### Laser-to-Camera Registration

Distance-based focus adjustment and laser workspace highlighting requires registration between the imaging device and the laser. In an initial step, a planar, wooden medium-density fiberboard (MDF) is positioned perpendicular to the laser $z$-axis (see coordinate frame $(\mathrm{CF})_\mathrm{F}$ shown Figure 4.2a). By ablating a grid of five points with an outer dimension of $(9 \times 9)\,\mathrm{mm}^2$ at different heights in $z$-direction (see Figure 4.2b), an OCT-based spot ablation test can be performed. In these, the resultant laser ablations are analyzed in terms of crater depth and diameter. The latter is minimal at the beam waist (diameter approx. $330\,\mu\mathrm{m}$) that is located at $z = {}_{(\mathrm{F})}z = 0$. If only one ablated plane is considered, a systematic deviation between the laser axis and the estimated focal plane at coordinate frame $(\mathrm{CF})_\mathrm{F}$ would remain. Thus, an extended registration based on a principal component analysis (PCA) taking $n$ of aforementioned ablation patterns into account is proposed.

The planar MDF surface is reconstructed with the method from Section 3.1 and is parametrized with a plane segmentation based on the random sample consensus (RANSAC) outlier rejection scheme [FB81]. The computation is done with respect to the left camera frame $(\mathrm{CF})_\mathrm{L}$. Next, the outer corner points ${}_{(\mathrm{L})}\boldsymbol{Q}_{ij}$ as well as the center points ${}_{(\mathrm{L})}\boldsymbol{C}_i$ are extracted manually for each $z$-position with $i \in \{1,...,n\}$ and $j \in \{1,...,4\}$. The center points ${}_{(\mathrm{L})}\boldsymbol{C}_i = \left({}_{(\mathrm{L})}x_i,{}_{(\mathrm{L})}y_i,{}_{(\mathrm{L})}z_i\right)^\mathrm{T}$ localized in $(\mathrm{CF})_\mathrm{L}$ define the intersection between the laser in its non-deflected position ${}_{(\mathrm{F})}\boldsymbol{C}_i = \left(0,0,{}_{(\mathrm{F})}z_i\right)^\mathrm{T}$ and current plane $i$ as illustrated in Figure 4.2b. Initially, mean compensation is performed considering the average position ${}_{(\mathrm{L})}\overline{\boldsymbol{C}} = \left({}_{(\mathrm{L})}\overline{x},{}_{(\mathrm{L})}\overline{y},{}_{(\mathrm{L})}\overline{z}\right)^\mathrm{T}$ of the entire center points yielding

$$ {}_{(\mathrm{L})}\widehat{\boldsymbol{C}}_i = {}_{(\mathrm{L})}\boldsymbol{C}_i - {}_{(\mathrm{L})}\overline{\boldsymbol{C}}. \tag{4.1} $$

Applying the PCA, the laser axis, denoted by ${}_{(\mathrm{L})}\boldsymbol{n}$, is obtained by the direction of the greatest variance in the center point distribution. In other words, the axis is given by the eigenvector corresponding to the largest eigenvalue of the covariance matrix

$$ \boldsymbol{\Sigma}_\mathrm{C} = \frac{1}{n}\,{}_{(\mathrm{L})}\widehat{\boldsymbol{C}}\,{}_{(\mathrm{L})}\widehat{\boldsymbol{C}}^\mathrm{T}, \tag{4.2} $$

where matrix ${}_{(\mathrm{L})}\widehat{\boldsymbol{C}} = \left({}_{(\mathrm{L})}\widehat{\boldsymbol{C}}_1, \ldots, {}_{(\mathrm{L})}\widehat{\boldsymbol{C}}_n\right)$ concatenates the mean compensated center points.

In order to map a surface point ${}_{(\mathrm{L})}\tilde{\boldsymbol{P}} = \left({}_{(\mathrm{L})}x,{}_{(\mathrm{L})}y,{}_{(\mathrm{L})}z,1\right)^\mathrm{T}$, as shown in Figure 4.2b, from coordinate frame $(\mathrm{CF})_\mathrm{L}$ to the laser frame $(\mathrm{CF})_\mathrm{F}$, a transformation ${}^\mathrm{F}\boldsymbol{T}_\mathrm{L}$ is required such that

$$ {}_{(\mathrm{F})}\tilde{\boldsymbol{P}} = {}^\mathrm{F}\boldsymbol{T}_\mathrm{L}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{P}}\,. \tag{4.3} $$

The translation component of ${}^\mathrm{F}\boldsymbol{T}_\mathrm{L}$ is given by the point ${}_{(\mathrm{L})}\boldsymbol{C}_i$ that has been localized at the focal plane, i.e., at ${}_{(\mathrm{L})}z = 0\,\mathrm{mm}$. The rotation is computed in two steps. The initial rotation is defined by an axis-angle representation determined as cross product ${}_{(\mathrm{F})}\boldsymbol{n} \times {}_{(\mathrm{L})}\boldsymbol{n}$ where ${}_{(\mathrm{F})}\boldsymbol{n} = (0,0,1)^\mathrm{T}$ defines the laser axis in $(\mathrm{CF})_\mathrm{F}$. Subsequently, the workspace orientation is aligned by a rotation about ${}_{(\mathrm{F})}\boldsymbol{n}$ applying a point-based, singular value decomposition-based registration between the ablated and the planned corner points ${}_{(\mathrm{F})}\boldsymbol{Q}_{ij}$. As a result, the transformation ${}^\mathrm{F}\boldsymbol{T}_\mathrm{L}$ is obtained that maps the point ${}_{(\mathrm{L})}\tilde{\boldsymbol{P}}$ to the task space of the three-axis laser scanning and ablation unit.

### Laser View Synthesis

As outlined above, each three-dimensional point ${}_{(\mathrm{F})}\boldsymbol{P}$, provided that corresponding pixels $\boldsymbol{p}_\mathrm{L}$ and $\boldsymbol{p}_\mathrm{R}$ are correctly matched in the stereo view, can be classified as inlier or outlier of the laser workspace. A solution of practical importance is to highlight those points directly in the live view.

However, if the camera and the laser cannot be aligned coaxially, occlusions and indentations at depth discontinuities might occur that are not visible from the camera view and thus cannot be highlighted as mentioned above. This is circumvented by computing a synthetic view from the laser perspective applying epipolar geometry constraints (see Figure 4.2b). A virtual pinhole camera is positioned along the estimated laser axis. In detail, a translational shift ${}^{\mathrm{V}}\boldsymbol{T}_{\mathrm{F}}$ is added to the registration transform ${}^{\mathrm{F}}\boldsymbol{T}_{\mathrm{L}}$ yielding transform

$$ {}^{\mathrm{V}}\boldsymbol{T}_{\mathrm{L}} = {}^{\mathrm{V}}\boldsymbol{T}_{\mathrm{F}}\,{}^{\mathrm{F}}\boldsymbol{T}_{\mathrm{L}} \tag{4.4} $$

that positions the virtual camera along the laser axis. Regarding the view synthesis, epipolar geometry can be applied as described in the following. If the intrinsic and extrinsic parameters of the stereo camera are known, the fundamental matrices ${}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{L}} \in \mathbb{R}^{3\times3}$ and ${}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{R}} \in \mathbb{R}^{3\times3}$, that relate corresponding pixels between the virtual and the stereo view, can easily be calculated [HZ04]. The resultant epipolar constraint, expressed in homogeneous pixel coordinates, is given by

$$ \tilde{\boldsymbol{p}}_{\mathrm{V}}^{\mathrm{T}}\,{}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{L}}\,\tilde{\boldsymbol{p}}_{\mathrm{L}} = \tilde{\boldsymbol{p}}_{\mathrm{V}}^{\mathrm{T}}\,{}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{R}}\,\tilde{\boldsymbol{p}}_{\mathrm{R}} = 0 \tag{4.5} $$

and implies that the virtual camera pixel $\tilde{\boldsymbol{p}}_{\mathrm{V}}$ is located on the epipolar line ${}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{L}}\,\tilde{\boldsymbol{p}}_{\mathrm{L}}$ for a given image point $\tilde{\boldsymbol{p}}_{\mathrm{L}}$ in the left camera view [HZ04]. The same applies to the epipolar line ${}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{R}}\,\tilde{\boldsymbol{p}}_{\mathrm{R}}$ and the corresponding right image point $\tilde{\boldsymbol{p}}_{\mathrm{R}}$. Intersecting the epipolar lines from the left and the right view according to the cross product

$$ s \cdot \tilde{\boldsymbol{p}}_{\mathrm{V}} = \left({}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{L}}\,\tilde{\boldsymbol{p}}_{\mathrm{L}}\right) \times \left({}^{\mathrm{V}}\boldsymbol{F}_{\mathrm{R}}\,\tilde{\boldsymbol{p}}_{\mathrm{R}}\right) \tag{4.6} $$

yields an estimate of the virtual image point $\tilde{\boldsymbol{p}}_{\mathrm{V}} = (u_{\mathrm{V}}, v_{\mathrm{V}}, 1)^{\mathrm{T}}$. If two surface points are projected to the same coordinate $\boldsymbol{p}_{\mathrm{V}}$, $z$–buffering with respect to $(\mathrm{CF})_{\mathrm{V}}$ is applied by selecting the pixel having the minimal distance to the epipole. The latter is defined by the projection of the virtual camera center onto the left view. As a consequence, the virtual laser view comprises only surface points that are visible, i.e., accessible, by the Er:YAG laser. The trifocal transfer in Equation 4.6 can easily be incorporated into laser visual servoing without the need of estimating tissue surface [ATDH13]; however, this requires to robustly detect the visible laser spot in the surgical scene.

In this thesis, the virtual view is computed by sub-pixel sampling of the estimated disparity map using bi-linear interpolation. Compared to common ray-casting methods, the proposed view synthesis approach is highly parallelizable and thus enables efficient GPGPU implementation.

### 4.1.2 Laser Ablation Experiments

Based on the laser-to-camera registration and laser workspace highlighting, ablations with and without focus adjustment are planned and performed on different samples. The related experimental design for quantitative validation of the proposed methods is presented in the following.

### Validation of the Laser-to-Camera Registration

The calibrated left camera is used as for validating the laser-to-camera registration. In detail, a printed $9 \times 8$ circle grid at $(8 \times 7)\,\mathrm{mm}^2$, as illustrated in Figure 4.3a, is attached to an MDF sample and positioned within the laser workspace at different levels $-4.4\,\mathrm{mm} \leq z \leq 4.4\,\mathrm{mm}$. For each grid point ${}_{(\mathrm{G})}\boldsymbol{P}$ with reference to the grid frame $(\mathrm{CF})_{\mathrm{G}}$ its corresponding pixel $\tilde{\boldsymbol{p}}_{\mathrm{L}} = (u_{\mathrm{L}}, v_{\mathrm{L}}, 1)^{\mathrm{T}}$ is detected in the calibrated left view. The related external transformation $\left({}^{\mathrm{L}}\boldsymbol{R}_{\mathrm{G}} \,|\, {}_{(\mathrm{L})}\boldsymbol{t}_{\mathrm{G}}\right) \in \mathbb{R}^{3\times4}$ is computed by an iterative Levenberg-Marquardt optimization that minimizes the grid point re-projection error [Bra00]. The resulting mapping is defined by

$$\tilde{\boldsymbol{p}}_{\mathrm{L}} = \boldsymbol{K}_{\mathrm{L}} \left({}^{\mathrm{L}}\boldsymbol{R}_{\mathrm{G}} \,|\, {}_{(\mathrm{L})}\boldsymbol{t}_{\mathrm{G}}\right) {}_{(\mathrm{G})}\tilde{\boldsymbol{P}} = \boldsymbol{K}_{\mathrm{L}} \left({}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},1}\, {}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},2}\, {}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},3} \,|\, {}_{(\mathrm{L})}\boldsymbol{t}_{\mathrm{G}}\right) {}_{(\mathrm{G})}\tilde{\boldsymbol{P}}, \tag{4.7}$$

where $\boldsymbol{K}_{\mathrm{L}} \in \mathbb{R}^{3\times3}$ defines the left camera matrix and ${}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},i}$ denotes a column of the rotation matrix ${}^{\mathrm{L}}\boldsymbol{R}_{\mathrm{G}} \in \mathbb{R}^{3\times3}$. The point ${}_{(\mathrm{G})}\boldsymbol{P}$ is given by homogeneous coordinates ${}_{(\mathrm{G})}\tilde{\boldsymbol{P}} = \left({}_{(\mathrm{G})}x, {}_{(\mathrm{G})}y, {}_{(\mathrm{G})}z, 1\right)^{\mathrm{T}}$ with respect to $(\mathrm{CF})_{\mathrm{G}}$. Due to grid planarity (${}_{(\mathrm{G})}z = 0$), Equation 4.7 simplifies to

$$\tilde{\boldsymbol{p}}_{\mathrm{L}} = \boldsymbol{K}_{\mathrm{L}} \left[{}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},1}\, {}^{\mathrm{L}}\boldsymbol{r}_{\mathrm{G},2} \,|\, {}_{(\mathrm{L})}\boldsymbol{t}_{\mathrm{G}}\right] \begin{pmatrix} {}_{(\mathrm{G})}x \\ {}_{(\mathrm{G})}y \\ 1 \end{pmatrix} = {}^{\mathrm{L}}\boldsymbol{H}_{\mathrm{G}} \begin{pmatrix} {}_{(\mathrm{G})}x \\ {}_{(\mathrm{G})}y \\ 1 \end{pmatrix}, \tag{4.8}$$

where the homography ${}^{\mathrm{L}}\boldsymbol{H}_{\mathrm{G}} \in \mathbb{R}^{3\times3}$ (see Figure 4.3b) describes a plane-to-plane projective transformation [HZ04]. Circle grid images are detected for varying $z$–levels. Subsequently, the grid pattern is ablated onto a MDF sample of same height as the one comprising the printed grid. The same image-based corner detection is applied to determine $\tilde{\boldsymbol{p}}_{\mathrm{L,a}} = (u_{\mathrm{L,a}}, v_{\mathrm{L,a}}, 1)^{\mathrm{T}}$ that corresponds to the ablated point ${}_{(\mathrm{G})}\tilde{\boldsymbol{P}}_{\mathrm{a}} = \left({}_{(\mathrm{G})}x_{\mathrm{a}}, {}_{(\mathrm{G})}y_{\mathrm{a}}, 0, 1\right)^{\mathrm{T}}$. Considering the inverse mapping

$$s \begin{pmatrix} {}_{(\mathrm{G})}x_{\mathrm{a}} \\ {}_{(\mathrm{G})}y_{\mathrm{a}} \\ 1 \end{pmatrix} = {}^{\mathrm{L}}\boldsymbol{H}_{\mathrm{G}}{}^{-1}\, \tilde{\boldsymbol{p}}_{\mathrm{L,a}}\,, \tag{4.9}$$

the distance error $\left\|{}_{(\mathrm{G})}\boldsymbol{P} -_{(\mathrm{G})} \boldsymbol{P}_{\mathrm{a}}\right\|_2$ is computed to assess the registration quality.

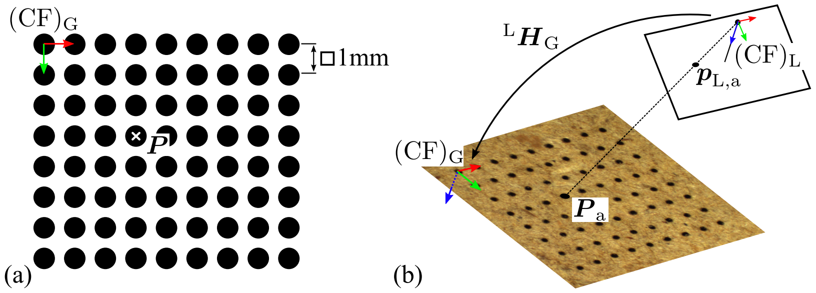

$(\mathrm{CF})_{\mathrm{G}}$
□1mm
$\boldsymbol{P}$
(a)
${}^{\mathrm{L}}\boldsymbol{H}_{\mathrm{G}}$
$(\mathrm{CF})_{\mathrm{L}}$
$\boldsymbol{p}_{\mathrm{L,a}}$
$(\mathrm{CF})_{\mathrm{G}}$
$\boldsymbol{P}_{\mathrm{a}}$
(b)

Figure 4.3: Registration validation by using (a) a circle grid that is ablated onto (b) a MDF sample.

### Laser Focusing on Medium-Density Fiberboards (MDF)

Validation of the laser-to-camera registration and of the distance-based laser focus adjustment is carried out on MDF specimens as shown in Figure 4.4a. In the latter case, laser cutting is compared for trials without (NAF – no autofocus) and with autofocus (AF – autofocus).
After successful registration, the planar specimen is positioned parallel ($\phi = 0°$) to the focal plane ($z = 0\,\text{mm}$). Incisions with and without focus adaptation are performed for eight distances with respect to the focal plane. Due to the reduced OCT scanning range, these distances are within a range of $-3.2\,\text{mm} \leq z \leq 4.7\,\text{mm}$. In addition, another MDF specimen that is tilted with respect to the focal plane ($\phi = 21°$) is ablated in order to demonstrate focus adaption on a varying depth. Here, the incisions start at the focal plane and pass a range of $0\,\text{mm} \leq z \leq 3.7\,\text{mm}$.
Regarding the parallel configuration at eight different heights, 16 datasets ($8\times$NAF, $8\times$AF) are quantitatively assessed. For each trial, a straight cutting line with a length of 10 mm is defined by two points using the graphics tablet (Bamboo CTH-470, Wacom Co., Ltd., Japan).

### Laser Focusing on Soft Tissue and Bone

In addition to the MDF experiments, distance-based laser focus adjustment is investigated on *ex vivo* soft tissue and bone (see Figure 4.4b and 4.4c). Six incisions ($3\times$NAF, $3\times$AF) are carried out on poultry tissue and porcine femur. The incisions cover a depth range of approximately $0\,\text{mm} \leq z \leq 3.6\,\text{mm}$. In order to obtain a significant ablation depth on the femur bone, each laser scan is repeated three times. For the two *ex vivo* series, the incision length is 10 mm. An additional suction attached to the laser removes the ablated particles but at the same times causes dehydration of the tissue. Thus, the tissue is slightly moistened with saline solution before activating the laser.

### Laser and Scanning Parameters

The laser parameters and scanning velocity are summarized in Table 4.1. To minimize thermal damage on the soft tissue and bone, ablation is performed with a reduced diode current and an increased scanning speed compared to the MDF trials.

Table 4.1: Er:YAG laser and scanner parameters during experiments

| | Diode current in A | Pulse duration in µs | Pulse frequency in Hz | Scanning velocity in mm/s |
|---|---|---|---|---|
| MDF specimen | 200 | 200 | 50 | 1.0 |
| Soft tissue & bone | 120 | 200 | 50 | 4.0 |

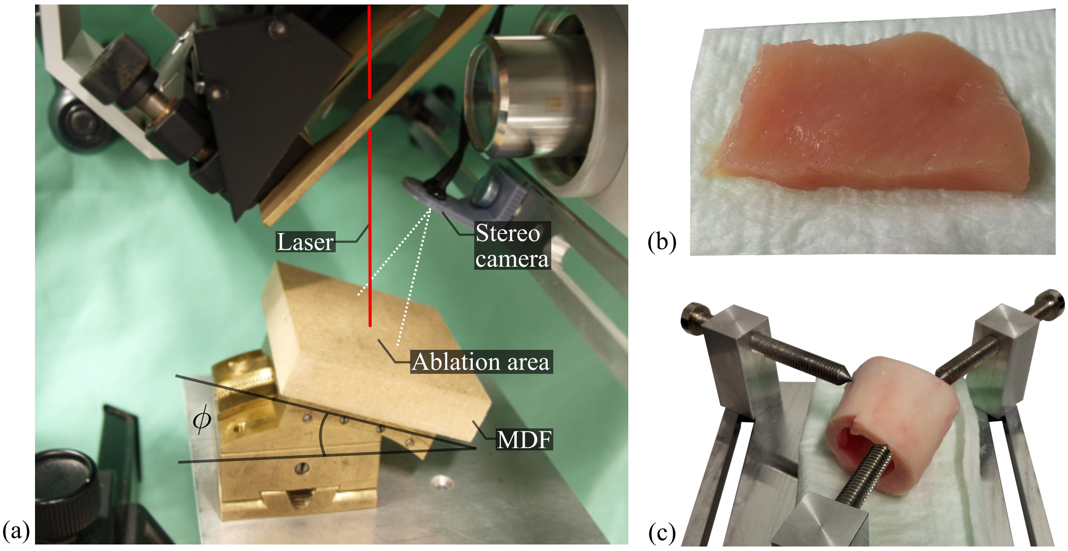


Figure 4.4: Specimens used during the laser ablation trials. Laser ablation is performed on (a) tissue substitutes (MDF) that are positioned parallel and tilted with respect to the focal plane of the laser, (b) *ex vivo* poultry tissue, and (c) porcine femur (bone).

### OCT Measurements of Ablation Geometry

The ablation geometry, i.e., incision width and depth, are analyzed with OCT scans. If tissue is properly aligned with respect to the laser focus, the incision depth should be maximal while its width should be minimal. To measure this, the ablated surface is segmented with a recent method adapted to OCT measurements [FSK$^+$12] based on active contours [XP98].

## 4.1.3 Results

In the following, the results of the laser-to-camera registration, the workspace highlighting, and the distance-based laser focusing are presented.

### Laser-to-Camera Registration

Quantitative assessment of the laser-to-camera registration is conducted by analyzing the ablated grids in the calibrated view. The resulting ablation errors are illustrated in Figure 4.5a in the form of box plots. The maximum deviation is $0.2\,\text{mm}$ (at $z = 3\,\text{mm}$) whereas the medians do not exceed $0.075\,\text{mm}$. The highest correlation of planning and ablation is achieved in the range of $-1\,\text{mm} \leq z \leq 1\,\text{mm}$, close to the focal plane at $z = 0$. Respective medians are about $0.05\,\text{mm}$. Since a simple parallel beam model is assumed, there is a slight increase in the ablation error for $|z| > 2\,\text{mm}$. To compensate for this deviation, a more complex model would be required that includes a height-dependent lateral scaling factor. However, due to the high ablation accuracy achieved with the simple model, an extension of the model is not considered in this thesis.

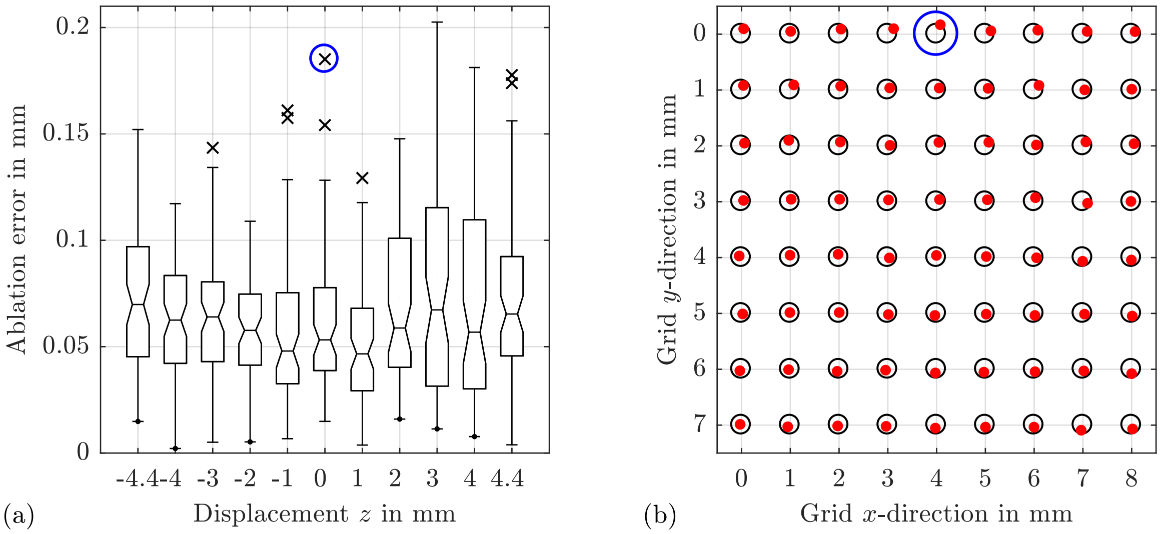


Figure 4.5: Ablation errors of the registration validation shown as boxplots (a). Outliers are denoted by a cross. For $z = 0$ mm, the ablated grid shown in (b). The planned grid is defined by the black circles whereas the ablated points are highlighted by the red dots. A concentric alignment denotes an error equal to zero. The maximum error for $z = 0$ mm is marked by a blue circle.

The ablation results for $z = 0$ (at the focal plane) are exemplarily shown in Figure 4.5b. The maximum ablation error of $0.18$ mm is highlighted by the blue circle. The observed model inaccuracies are illustrated by the slight increase of the scaling in radial direction.

### Visualization of the Laser Workspace

Qualitative evidence is given by visual inspection of the laser workspace superimposed to the live view in real-time. As Figure 4.6a shows the scene directly after performing the registration, it can be seen that the lateral workspace of $(10 \times 10)$ mm$^2$ is correctly aligned to the ablated grid. Any pixel that is not included in the workspace is grayed out (see Figure 4.6f–h). If the sample is positioned with $|z| > 5$ mm, the loss of workspace highlighting indicates that the target is out of the focal range. This provides instant feedback for manual repositioning. Regarding the intraoperative use, the surgeon can easily position the surgical laser correctly within a certain distance. For further investigation on superimposing the workspace for visual guidance, please refer to Section 5.1.

For a scene with varying depth, such as depicted in Figure 4.6b–e, the overlay provides visual feedback on the surface shape even in the mono view. Another feature is the virtual view from the laser perspective. Significant depth discontinuities, as illustrated in Figure 4.6c, that cause occlusions in the left image are localized in the laser view (see Figure 4.6e). If the laser passes this region, laser focus adaption has to be treated with caution because of the missing depth data. However, appropriate solutions can be found, e.g., by interpolating depth from the spatial neighborhood or a linear sub-sampling of the laser path.

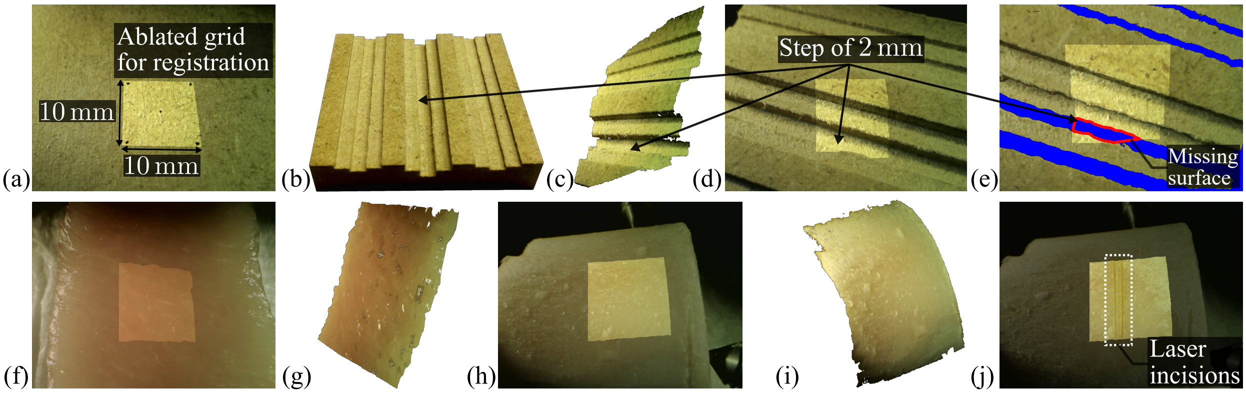


Figure 4.6: Live visualization with workspace highlighting and virtual laser view. For registration, a grid of five points (black dots) is ablated onto the MDF surface as shown in (a). Workspace highlighting is illustrated in (a,d) using (b) a MDF sample with milled steps of 2 mm. Its reconstructed surface is shown in (c). The virtual laser view in (e) reveals a lack of surface information (marked blue) from the laser perspective. The red border denotes the missing area inside the laser workspace. Regarding the laser trials, the workspace highlighting, and the computed surface of the soft tissue sample are shown in (f-g). Results on the porcine femur are presented in (h–j).

### Automatic Laser Focus Adjustment on MDF

Incisions without (NAF – no autofocus) and with laser focus adjustment (AF – autofocus) are performed on the MDF sample for eight distances to the focal plane. The OCT-based scans are segmented for computing the depth and the width of the incision. For NAF, the mean incision depth significantly drops from 500 µm to below 350 µm (see Figure 4.7a) if the sample is positioned out of focus ($z \geq 3$ mm). By contrast, AF demonstrates an almost constant depth for the entire $z$-range (SD below 55 µm). Regarding the incision width, similar results are observed (see Figure 4.7b). For AF, the mean incision width is kept between 325 µm and 360 µm (SD below 39 µm) indicating optimal cutting. For NAF, i.e., at $z = 4.7$ mm, the width significantly slopes upwards to a maximum of 564 µm.

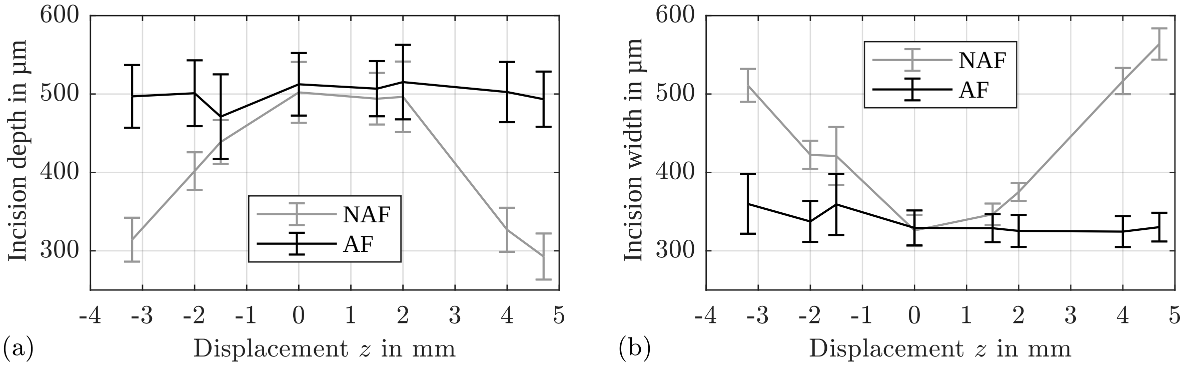


Figure 4.7: Comparison of the ablation results on MDF with (AF) and without focusing (NAF) at different $z$-levels. Mean and SD are shown in (a) for the incision depth and in (b) for the incision width.

Constant incision quality is achieved for $0\,\mathrm{mm} \leq z \leq 2\,\mathrm{mm}$ even without focusing. Thus, the laser focus is assumed to be approx. at $z = 1\,\mathrm{mm}$ and not at $z = 0$. This deviation is attributed to the problem of correctly estimating the center of the focal range, i.e., the exact position of the beam waist. This can be addressed by the further advanced knife-edge technique [KFS+16].

To conclude, distance-based laser focus adjustment facilitates constant cutting geometries, regardless of the tissue distance. The incisions carried out on the tilted MDF sample, as presented in Figure 4.8a and 4.8d, demonstrate the superior cutting performance when using AF.

### Automatic Laser Focus Adjustment on *Ex Vivo* Tissue

Validation is carried out for the soft tissue and bone samples. Incisions are performed in a depth range of $0\,\mathrm{mm} \leq z \leq 3.6\,\mathrm{mm}$ demonstrating that improved ablation geometry is achieved with AF. Two cuttings are exemplarily illustrated in Figure 4.8b and 4.8c.

Aforementioned findings are approved by visual inspection with a microscope. Regarding porcine femur, the surface is ablated with a reduced depth if the laser is not focused (NAF). Especially at $z = 3.6\,\mathrm{mm}$, the effect of cutting seems to disappear, as illustrated in Figure 4.8f. For the soft tissue scenario that most likely refers to the targeted vocal fold laser surgery, precise and narrow incisions can be reproduced if the distance-based laser focusing (AF) is utilized.

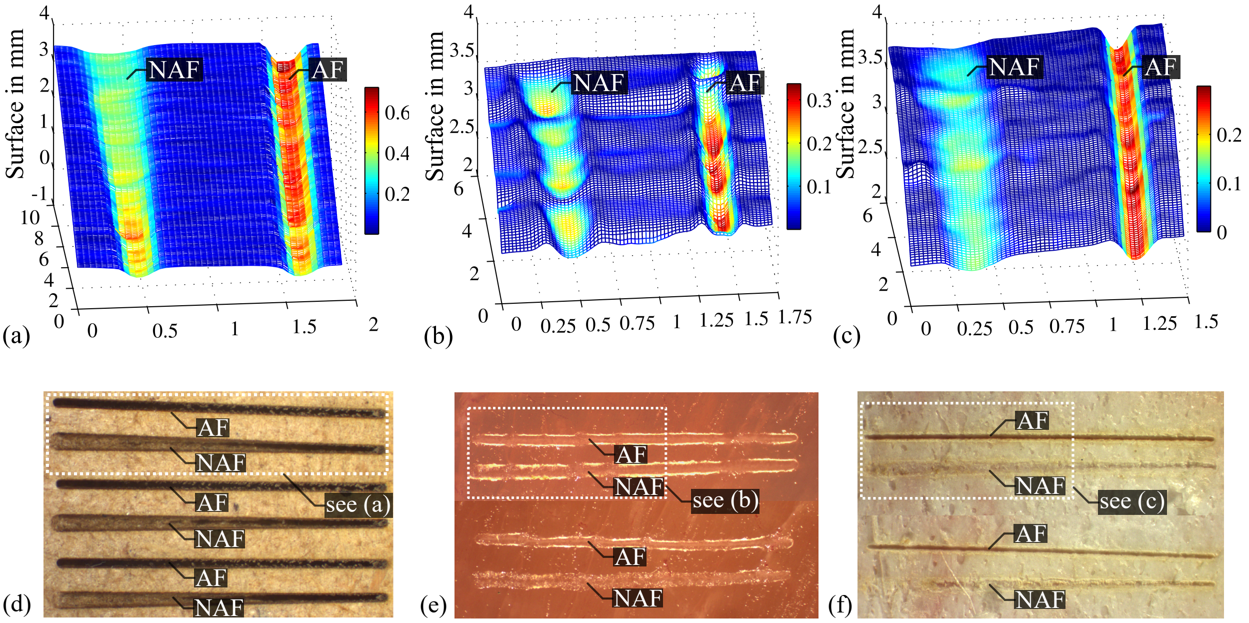


Figure 4.8: OCT-based segmentation of laser cutting performed on (a) MDF in the tilted configuration, (b) poultry tissue, and (c) porcine femur. Corresponding microscopic views are shown below (d-f).

## 4.2 Fusion of Stereo Vision and OCT

As outlined in Section 1.2, OCT provides a valuable alternative to diagnosis of the laryngeal submucosal space. However, the simultaneous use of stereo white light and OCT imaging, i.e., the visualization on two different displays, is non-intuitive. This is because of an increased mental workload when establishing correspondence between the two views or when transferring the surgical plans from one modality to the other. As a consequence, fusing stereo vision with OCT in a single view, e.g., in the form of an image overlay, has a high clinical relevance.

### 4.2.1 OCT Surface Segmentation and Registration

In this section, a framework for automatic surface segmentation and feature-based registration of OCT with stereo vision is presented (see Figure 4.1b). Evaluation is conducted on a custom-made reference sample and the results are finally demonstrated by two augmented reality concepts.

#### Bimodal Imaging Setup

The bimodal setup consists of an SD-OCT (see Section 2.5) and the stereo module SOM (see Section 3.2) both mounted in a converged alignment, as shown in the schematic drawing of Figure 4.9a. The camera is positioned at a distance of $40\,\text{mm}$ above the scene surface. The OCT scan range is set to $(15 \times 15 \times 2.76)\,\text{mm}^3$ at a resolution of $256 \times 256 \times 1024$ voxels, resulting in a depth resolution of $2.69\,\mu\text{m}$. The camera parameters are listed in Table 3.1.

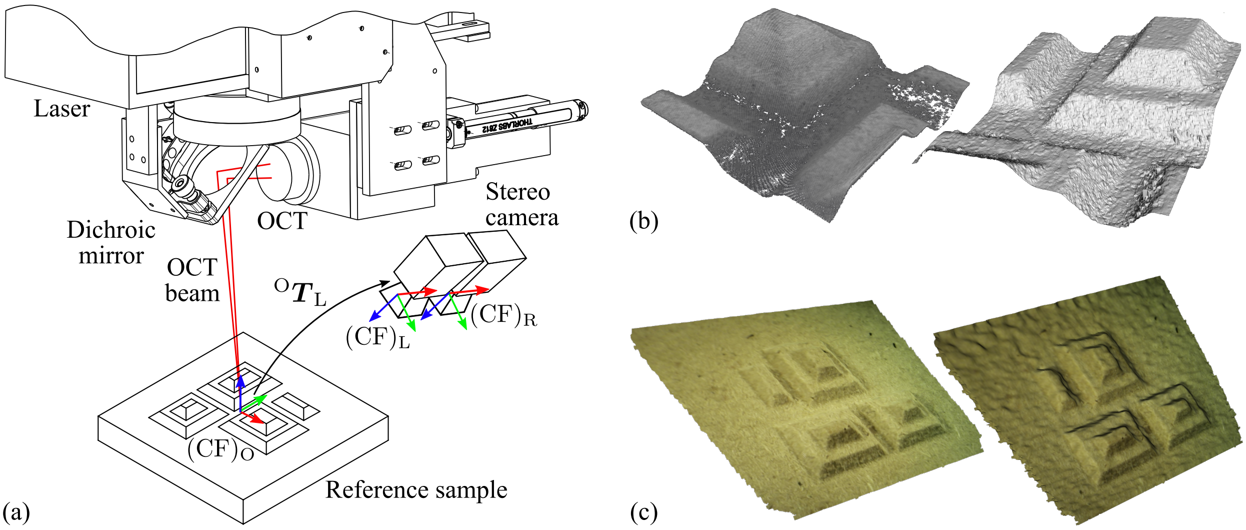

Laser
Dichroic mirror
OCT beam
OCT
Stereo camera
$^{\mathrm{O}}\boldsymbol{T}_{\mathrm{L}}$
$(\mathrm{CF})_{\mathrm{L}}$
$(\mathrm{CF})_{\mathrm{R}}$
$(\mathrm{CF})_{\mathrm{O}}$
Reference sample
(a)
(b)
(c)

Figure 4.9: Imaging setup with OCT and stereo camera (a). For registration, the surface of a reference sample is computed by (b) processing the OCT volume (left) yielding a triangulated surface mesh (right), as well as by (c) the stereo matching method yielding the surface point cloud (left) and its triangulated mesh (right).

### Sample Design

Registration of OCT and stereo white light imaging is a challenging task due to the different optical characteristics. OCT does not provide texture information as white light imaging while stereo vision is not able to acquire volumetric, i.e., tomographic image data. The common ground are surface features. Due to the very low axial OCT scan range of $2.76\,\mathrm{mm}$ that limits the possible sample height, a reference sample similar to the one described in Section 3.2 is designed.

As shown in Figure 4.9a, a structure of truncated pyramids with a height of $2\,\mathrm{mm}$ and an extend of $(20 \times 20)\,\mathrm{mm}^2$ is made of MDF. A schematic drawing is shown in Figure 4.9a. As described in the following section, an OCT and a stereo image are acquired to estimate the transform ${}^{\mathrm{O}}\boldsymbol{T}_{\mathrm{L}}$.

### Surface Segmentation in OCT Image Data

A near real time OCT surface segmentation is achieved by a sequence of binary thresholding, connected component labeling for image denoising, and final relabeling. Exemplary results of this work flow are illustrated in Figure 4.10 for an OCT B-scan.

As shown in Figure 4.10a and 4.10b, the raw OCT-scan is initially binarized considering a empirically found threshold. The main surface layer is detected by connected component labeling using the quadtree technique that simultaneously eliminates speckles [Sam81]. Exploring all spatial neighbors of a voxel in a propagation scheme, elements of similar appearance are assigned to a label. The resulting labels are sorted with respect to their size. During the relabeling phase, small regions are classified as noise speckles and thus, can be successfully eliminated with the result that the volume representing the surface layer remains (see Figure 4.10c).

Finally, the surface is extracted by iterating through the relabeled volume for each A-scan voxel stack until the first non-zero element is found. As a result, an accurate representation of the object surface, as shown in Figure 4.9b, is created.

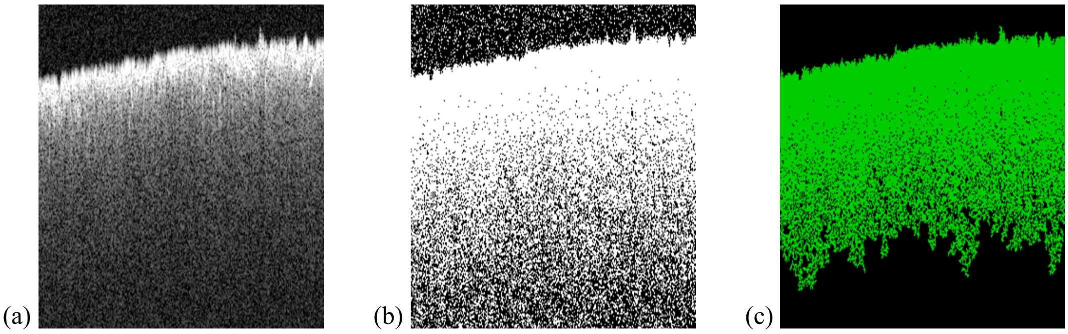


Figure 4.10: Surface segmentation of (a) the raw OCT image (B-scan) is achieved by (b) binarization followed by (c) connected component labeling.

### Registration of the Segmented Surfaces

The alignment between OCT and stereo camera is computed by a two-step registration of the object surface segmented from the OCT and the stereo image data (see Figure 4.9c). For the latter, the reconstruction is obtained according to Section 3.1.

Initially, the surfaces are coarsely aligned with a feature-based method, followed by an fine registration with the iterative closest point (ICP) algorithm [BM92]. The work flow is illustrated in Figure 4.11a. Both point clouds are down-sampled by partitioning the image space into a voxel-grid in order to speedup the computation process. The cloud points within a voxel are consolidated into its center of gravity. Next, features based on the Fast Point Feature Histograms (FPFH) are computed [RBB09]. This method represents the fast version of the Point Feature Histograms (PFH) [RBMB08]. FPFH are descriptive, local surface features encoding the spatial geometry, such as the surface normal distribution. Unique features are required to establish correspondence between the voxel grids. Those are determined with a persistence analysis [RBMB08]. In detail, distinct FPFH features are found by calculating the Kullback-Leibler divergence (KLD) with respect to the mean histogram computed from all features. If the KLD is greater than the standard deviation of the assumed Gaussian distribution, the feature point is considered to be unique.

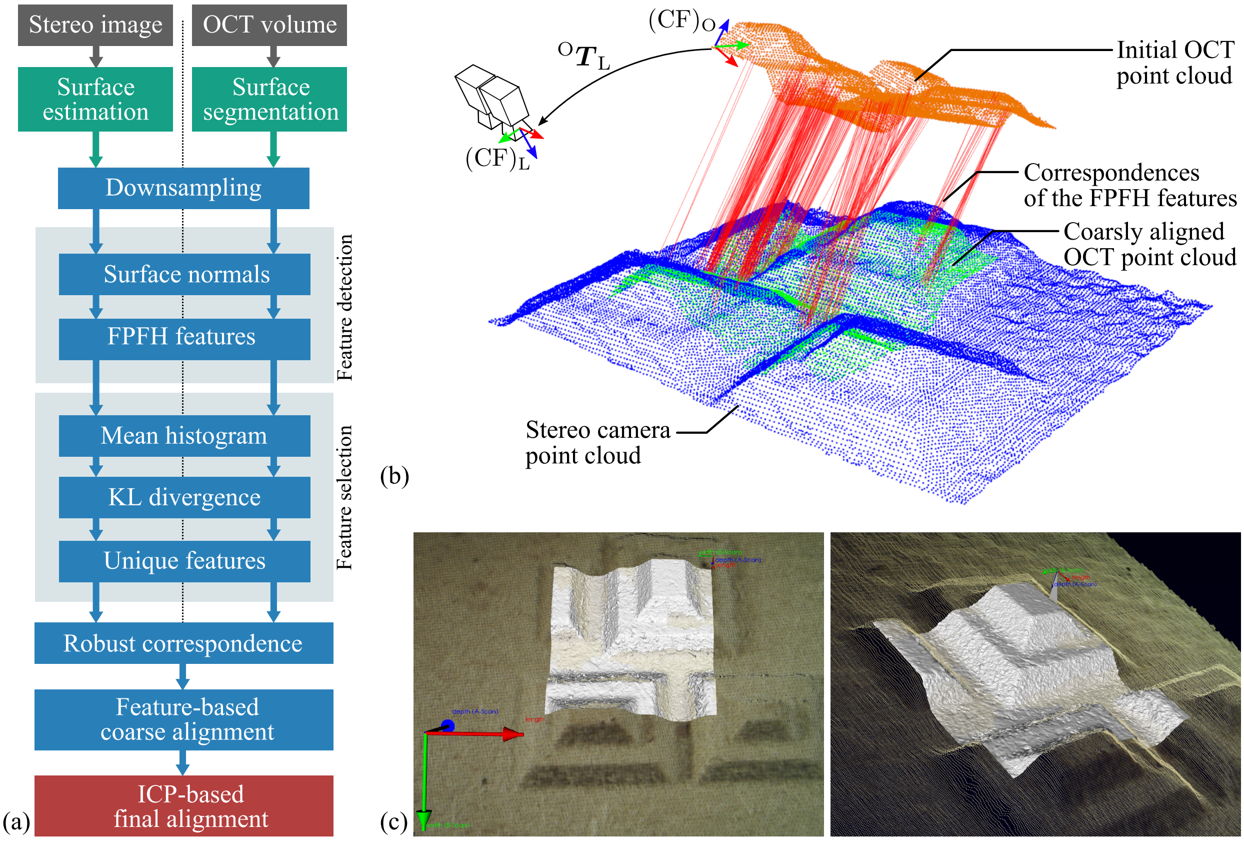


Figure 4.11: Sequence of the registration (a). Correspondences are established for (b) coarse alignment of the two surfaces before performing (c) an ICP-based fine registration.

The transform $^{\mathrm{O}}\boldsymbol{T}_{\mathrm{L}}$, as shown in Figure 4.11b, is estimated by a rigid, point-based registration considering the most descriptive features and a random sample consensus (RANSAC) outlier rejection scheme [FB81]. This provides an initial guess for the subsequent ICP-based fine registration on the two raw (non down-sampled) point clouds. For correspondence estimation, nearest neighbor search is applied. The algorithm is implemented in C++ using the Point Cloud Library (PCL) [RC11]. The final alignment of two surfaces is depicted in Figure 4.11c.

### 4.2.2 Results

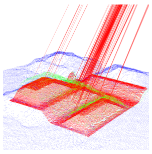
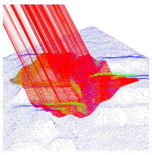
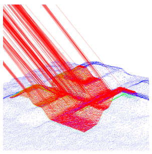
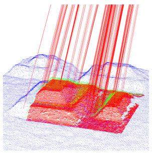
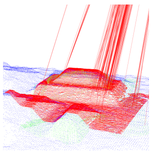
Accuracy assessment is provided for five registrations performed on the reference sample from reasonable perspectives (see Table 4.2). For each trial, the root mean square error (RMSE) of both the coarse and the fine registration are computed from the Euclidean distance of the correspondences.

Table 4.2 lists the RMSE of the five registration trials. Furthermore, the point clouds of the OCT segmentation after coarse (green) and fine registration (red) as well as the surface cloud obtained from the stereo view (blue) are shown. The red lines represent the correspondences of the unique features used for the coarse registration.

Within the five measurements, different sections of the pyramidal structure have been considered for registration. Measurements 1–3 utilize a centered region with a high variance in the surface structure resulting in a registration error less than 0.18 mm. By contrast, measurement 4 covers only one of the larger pyramids and the small one, potentially leading to the increased registration error of 0.61 mm. For measurement 5, the OCT coverage is not as centric as for measurements 1–3, but compared with measurement 4, the surface features are more distinct yielding a lower RMSE of 0.41 mm. As most of the correspondences are found on the edges of the truncated pyramids, it can be concluded that a high variance of the surface structure is mandatory to precisely register the OCT with respect to the stereo camera. Otherwise, the ICP registration does not converge into the correct minimum due to imprecise coarse alignment, as demonstrated by measurement 4.

The average computational time is 2.8 s for the OCT surface segmentation, 73.74 s for the coarse and 7.5 s for the fine registration (Intel®Core™i7-3770K, 16GB RAM, Ubuntu 12.04 LTS).

Table 4.2: RMSE in mm after coarse and after fine registration.

| Measurement | <br>1 | <br>2 | <br>3 | <br>4 | <br>5 | Mean |
|---|---|---|---|---|---|---|
| Coarse alignment | 0.43 | 0.22 | 0.22 | 0.59 | 0.79 | 0.45 |
| Final alignment | 0.18 | 0.11 | 0.12 | 0.61 | 0.41 | 0.29 |

The results are demonstrated by two augmented reality concepts. In the first example shown in Figure 4.12a, distance information calculated from the OCT surface data, here the sample height, are color-encoded and superimposed to the live camera view. In the second example, texture information of the camera image is mapped onto the highly accurate OCT surface data (see Figure 4.12b). Augmented reality for laryngeal surgery is discussed in detail in Section 5.2.

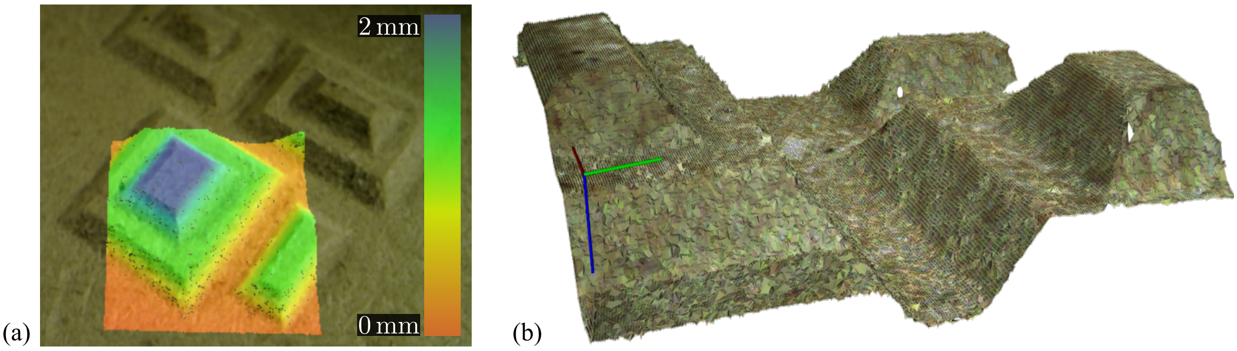


Figure 4.12: Image overlay of (a) color-coded distance in the live view and (b) texture on the OCT surface.

## 4.3 Conclusion

In this chapter, results on aligning laser workspace information and OCT image data with the live stereo view are presented. In the first case, a method for laser-to-camera registration is proposed. The distance between the tissue surface and laser focus is determined online enabling accurate laser focus adaption. Conducted laser experiments demonstrate that the ablation geometry can be kept uniform within a depth range of almost 8 mm. Real-time surface reconstruction combined with straightforward trifocal model of laser and stereo camera furthermore facilitate another feature which is the virtual laser view. The latter can easily reveal areas with uncertain depth information from the laser's perspective, i.e., at depth discontinuities that are not visible in the live camera image. When ablating those regions, warning signals can be visualized to the surgeon.

The presented framework for registration of OCT and stereo vision demonstrates successful fusion of the two modalities. Its practical relevance is underlined when thinking of commercial stereo microscopes with OCT functionality [LKO$^+$13]. For instance, mapping texture information directly onto the volumetric OCT scan provides a global context of the surgical scene and thus can simplify the navigation through the volume data. On the other side, structural information of the subcutaneous space computed from OCT data can be superimposed to the live view, as shown in Section 5.2, without compromising depth perception in case of stereoscopic visualization.

# 5 Development of a Visualization and Planning Interface

Consistent alignment of the laser focus with respect to the tissue is inevitable when aiming at tissue-preserving ablation. This requirement can be addressed by taking tissue surface information into account. In practice, the laser focal range and the lateral scanning range, even if adjustable by motorization, are limited to a few millimeters. Consequently, the surgical laser has to be pre-positioned manually. To assist this task, Section 5.1 presents a method for superimposing laser workspace information to the live view considering the laser-to-camera registration discussed in Section 4.1.

Following the work flow outlined in Section 1.4, the surgeon then elaborates a dissection plan based on an optical biopsy with *in situ* OCT imaging. Benign or malignant lesions, i.e., located in the subepithelial space, are segmented and visualized to the surgeon. In this context, Section 5.2 introduces an OCT-based augmented reality concept considering the algorithm from Section 4.2. A segmentation framework is developed and demonstrated with a phantom replicating the optical properties of laryngeal tissue. A color gradient ranging from maximum function preservation to maximal radicality is presented providing assistance during the resection planning.

Once a dissection strategy is established, the incision is defined and laser cutting is initiated via the planning interface, as outlined in Figure 5.1. In Section 5.3, a comparative study on six stylus-tablet-based strategies for visualization and incision planning is discussed. User performance is assessed by means of planning accuracy, completion time, and ease of use.

The contents of this chapter are published in [SLK$^+$14, SLK$^+$15, CGS$^+$15, SKL$^+$16].

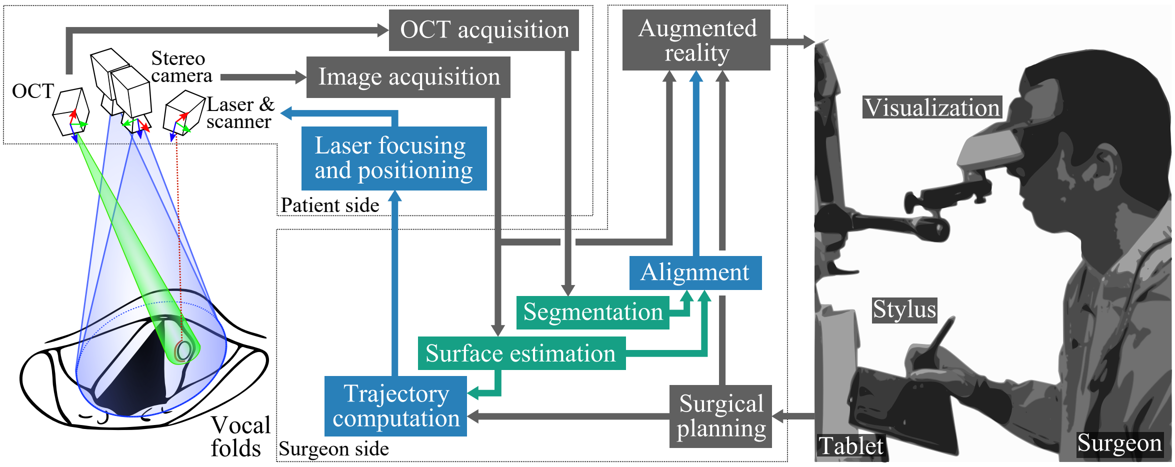


Figure 5.1: Work flow for visualization and planning.

## 5.1 Color-Encoded Distance for Interactive Laser Focus Positioning

In addition to precise control of the laser beam direction, efficacy of the ablation process depends on accurate laser focus positioning as demonstrated in Section 4.1. However, establishing consistent distance between laser focus and tissue surface is challenging.

A promising approach is represented by color-encoded distance visualization. Color-gradients, denoted as chromadepth, are superimposed to the live view. The refraction of visible light, i.e., hues ranging from red ($\lambda = 700\,\mathrm{nm}$, close scene objects) via orange, yellow, green to blue ($\lambda = 450\,\mathrm{nm}$, farther objects), leads to different focuses in the human eye and thus enable depth perception without the need of eyeglasses [Ste87]. Clinical application has been discussed for angiography [RSH06, WFW$^+$12], neurosurgery [SPP$^+$10, KOCD$^+$12], laparoscopy [ZIN$^+$13], or orthopedic surgery [DBW11]. However, the color coding context is either static meaning that the user does not alter the distance relation, or is applied to preoperative data only. By contrast, user interaction, that decide on the success or the failure of the positioning task, can benefit from visual feedback.

This section addresses interactive focus alignment of the laser to master the limited focal range and scanning workspace. Based on laser-to-camera registration, the reconstructed surface is virtually intersected with the laser workspace providing guidance during focus positioning tasks. Three visualization concepts are presented. The performance is measured and discussed for a user study conducted with ten subjects. For quantitative evidence, the different concepts are investigated on the laser setup outlined in Section 2.5. Qualitative results are obtained from laryngeal *in vivo* and cadaver sequences.

### 5.1.1 Visual Augmentation of the Live View

The concept of interactive focus positioning is evaluated using the surgical laser setup schematically shown in Figure 5.2a. The attached camera (see Section 3.2) comprises two miniaturized modules with an image resolution of $640 \times 480$ pixels. The cameras are mounted at a baseline of $6.3\,\mathrm{mm}$. Laser-to-camera registration is performed according to Section 4.1 while the target surface is reconstructed as outlined in Section 3.1.

Target positioning is achieved with a manually adjustable stage with millimeter scale attached to the laser setup. At the same time, the intersection of reconstructed surface and laser workspace is projected to the live view, as illustrated in Figure 5.2b, providing visual feedback for interactive laser focus positioning on the surface.

Any point ${}_{(\mathrm{F})}\boldsymbol{P} = \left({}_{(\mathrm{F})}x, {}_{(\mathrm{F})}y, {}_{(\mathrm{F})}z\right)^{\mathrm{T}}$ that is considered as inlier of the laser workspace is mapped from the frame $(\mathrm{CF})_{\mathrm{F}}$ to the frame $(\mathrm{CF})_{\mathrm{L}}$ applying the registration transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{F}}$. Furthermore, correspondence between object and image space is established by the function

$$s\,\tilde{\boldsymbol{p}}_{\mathrm{L}} = {}^{\mathrm{L}}\boldsymbol{M}_{\mathrm{F}}\,{}_{(\mathrm{F})}\tilde{\boldsymbol{P}} = {}^{\mathrm{L}}\boldsymbol{M}_{\mathrm{L}}\,{}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{F}}\,{}_{(\mathrm{F})}\tilde{\boldsymbol{P}}, \tag{5.1}$$

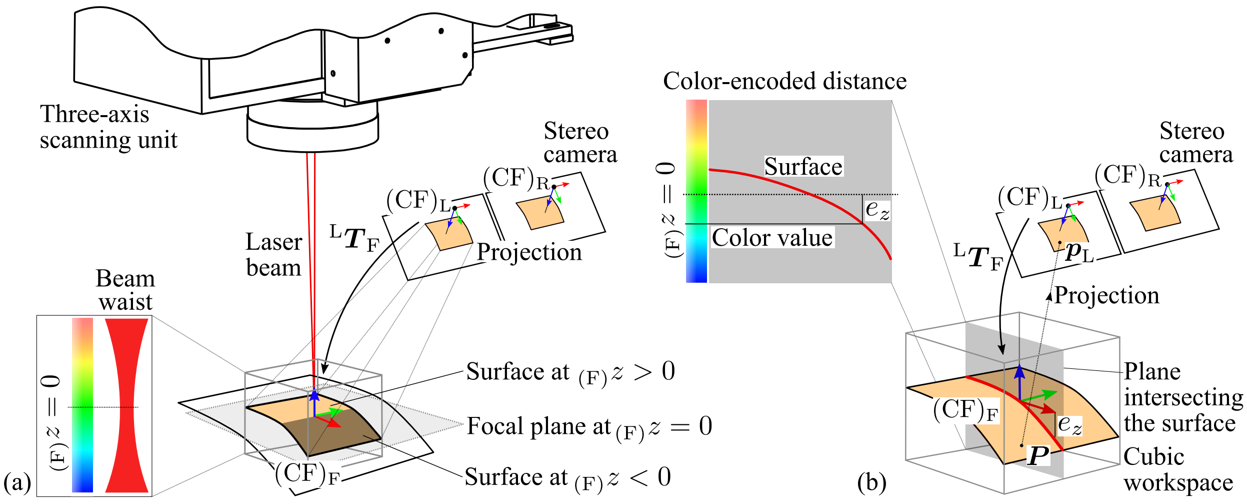

Three-axis scanning unit
Laser beam
Beam waist
${}_{(\mathrm{F})}z = 0$
${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{F}}$
$(\mathrm{CF})_{\mathrm{L}}$
$(\mathrm{CF})_{\mathrm{R}}$
Stereo camera
Projection
Surface at ${}_{(\mathrm{F})}z > 0$
Focal plane at ${}_{(\mathrm{F})}z = 0$
Surface at ${}_{(\mathrm{F})}z < 0$
$(\mathrm{CF})_{\mathrm{F}}$
(a)
Color-encoded distance
Surface
Color value
$e_z$
${}_{(\mathrm{F})}z = 0$
Stereo camera
$(\mathrm{CF})_{\mathrm{L}}$
$(\mathrm{CF})_{\mathrm{R}}$
${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{F}}$
$\boldsymbol{p}_{\mathrm{L}}$
Projection
$(\mathrm{CF})_{\mathrm{F}}$
$e_z$
$\boldsymbol{P}$
Plane intersecting the surface
Cubic workspace
(b)

Figure 5.2: The trifocal setup of stereo camera and laser is depicted in (a). As illustrated in (b), the distance $e_z$ to the focal plane ${}_{(\mathrm{F})}z = 0$ is color-encoded for every point inside the cubic workspace.

where ${}^{\mathrm{L}}\boldsymbol{M}_{\mathrm{L}}$ is the projection matrix obtained from camera calibration, as described in Section 2.3. The pixel position $\tilde{\boldsymbol{p}}_{\mathrm{L}} = (u_{\mathrm{L}},v_{\mathrm{L}},1)^{\mathrm{T}}$ and related tissue surface point ${}_{(\mathrm{F})}\tilde{\boldsymbol{P}} = \left({}_{(\mathrm{F})}x, {}_{(\mathrm{F})}y, {}_{(\mathrm{F})}z,1\right)^{\mathrm{T}}$ are described in homogeneous coordinates. Equation 5.1 considers equality up to a scale factor $s$.

Three concepts for superimposing laser workspace information to the live view are proposed. As shown in Figure 5.3a, the first concept, denoted by the gray mask (GM), is applied by graying out any surface region that is not included in the workspace. In addition, two transparent overlays based on chromadepth with either symmetric (SCC) or asymmetric (ACC) color gradients are implemented highlighting the distance

$$e_z = \left| {}_{(\mathrm{F})}z \right| \tag{5.2}$$

between the laser focal plane and the tissue surface located inside the cubic workspace. Surface points located at the focal plane, i.e., the beam waist at ${}_{(\mathrm{F})}z = 0$ (see Figure 5.2a), are colored green. Hues blue or red indicate surface areas inside the cubic workspace that are close to the axial workspace borders at $|{}_{(\mathrm{F})}z| = 5\,\mathrm{mm}$. A labeled color bar as shown in Figure 5.2b is displayed in the live view and linearly maps the hue value to a metric distance. Color-encoding is applied to the ${}_{(\mathrm{F})}z$ - direction only since the focusing aspect is the most critical part when aiming at efficient ablation with minimal tissue trauma. In the lateral direction, tissue located outside the workspace is excluded from color coding in order to maintain an unobstructed view onto the scene. The color-encoded overlay provides feedback on the distance and on the shape of the tissue surface even in a mono view. Furthermore, a change in color or even the disappearance of the overlay indicate that the target is out of focus and has to be re-positioned with respect to the laser focus. Otherwise, laser ablation potentially causes increased tissue trauma.

To compute the projective relation and the distance information for each pixel in real time, the image

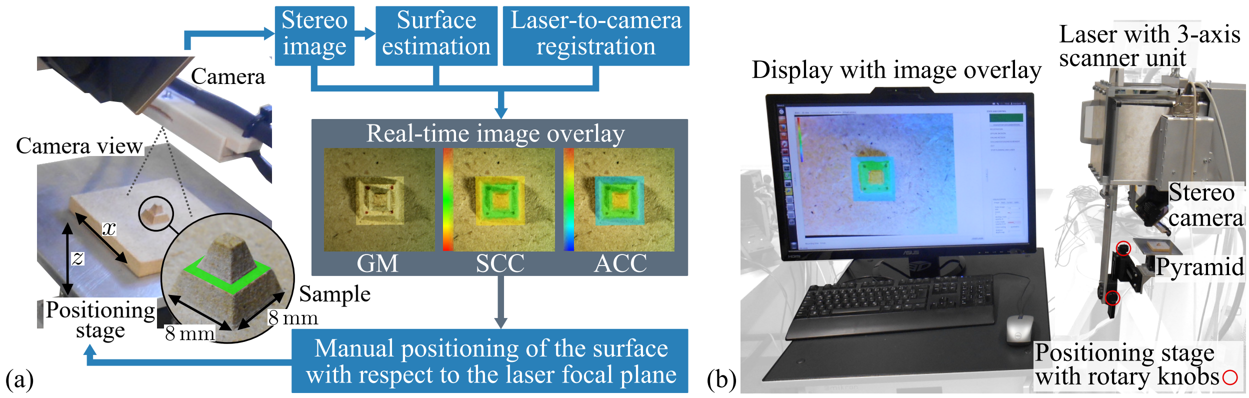

Stereo image
Surface estimation
Laser-to-camera registration
Camera
Camera view
$x$
$z$
Real-time image overlay
GM
SCC
ACC
Sample
8 mm
8 mm
Positioning stage
Manual positioning of the surface with respect to the laser focal plane
(a)
(b)
Display with image overlay
Laser with 3-axis scanner unit
Stereo camera
Pyramid
Positioning stage with rotary knobs ○

Figure 5.3: The work flow for interactive laser focus positioning is depicted in (a). A truncated, stepped pyramid of known shape is used to quantify positioning accuracy. Proposed overlay concepts deploy either a gray mask (GM), symmetric (SCC) or asymmetric color coding (ACC) for highlighting the laser workspace. The user study setup with live visualization, the surgical laser unit, and the positioning stage are shown in (b).

data is processed deploying CUDA on a GeForce GTX Titan (Nvidia Corporation, Santa Clara, CA, USA). Furthermore, OpenGL-based alpha compositing of the image and the color-gradient is applied for the augmented visualization.

### 5.1.2 User Study Design

Interactive focus positioning accuracy is assessed in a user study. A milled, stepped pyramidal structure has to be positioned with the aim that its intermediate plateau, as highlighted green in Figure 5.3a, is aligned with the focal plane at ${}_{(\mathrm{F})}z = 0$. The sample dimensions with a square base of $(8 \times 8)\,\mathrm{mm}^2$ and a height of $6\,\mathrm{mm}$ are chosen such that it can entirely be covered by the cubic laser workspace. Positioning is done by manually adjusting the translational stages of the sample holder (see Figure 5.3b). In order to focus the user attention to this specific task, all user trials are conducted performing neither incision planning nor laser ablation. Similar to an endoscopic scenario, where the laser has to be positioned in axial and lateral direction with respect to the vocal folds, the movements of the sample are restricted to the ${}_{(\mathrm{F})}x$ (lateral) and ${}_{(\mathrm{F})}z$ – direction (axial). The simplicity of the setup focuses the user's interest solely on the augmented scene visualization. The pyramid is mounted on the positioning table and the position of the intermediate plateau is automatically determined by image-based detection of the colored corner points.

To evaluate the eligibility of the concepts in terms of focus positioning assistance, experiments with ten volunteers having a background in medical engineering were carried out. Subjects were asked to move the pyramidal plateau from an out of workspace position into the center of the laser ablation volume, i.e., to ${}_{(\mathrm{F})}x = 0$ (lateral center) and ${}_{(\mathrm{F})}z = 0$ (focal plane) where color coding of SCC and ACC alters to green. During the conducted experiments, the three overlay concepts were tested randomly counterbalancing learning effects while positioning was repeated five times.

For quantitative evidence, the task completion time $t$ and the final positioning errors in $_{(\mathrm{F})}x$ and $_{(\mathrm{F})}z$ – direction, denoted by $e_x$ and $e_z$, respectively, were measured. In total, $50$ measurements were acquired per overlay concept ($150$ in total).

### 5.1.3 Results

Figure 5.4a–d show the position of the pyramidal plateau plotted over time achieved by two subjects. Regarding the positioning in $_{(\mathrm{F})}x$ – direction (see Figure 5.4a and 5.4c), the two subjects were able to align the pyramid to the workspace for all overlay concepts resulting in $e_x \leq 0.13\,\mathrm{mm}$. For the axial $_{(\mathrm{F})}z$ – direction, the performance based on SCC and ACC clearly differ compared with GM (see Figure 5.4b and 5.4d). When using GM and when starting from an out-of-focus position (see Figure 5.4e), the two subjects first had to locate the lower and the upper workspace border at $\left|_{(\mathrm{F})}z\right| = 5\,\mathrm{mm}$ before they were able to align the target as close as possible to the focal plane. As a result, positioning time increased by a few seconds in comparison to the results obtained with SCC or ACC. Although the completion time is slightly different for ACC and SCC, positioning in the $_{(\mathrm{F})}z$ – direction is highly accurate. Regarding the two subjects, the remaining distance to the focal plane is given by $e_z \leq 0.23\,\mathrm{mm}$ (SCC) and $e_z \leq 0.09\,\mathrm{mm}$ (ACC). The worst result for GM with an error of $e_z = 2.45\,\mathrm{mm}$ is measured for Subject $1$.

Table 5.1: Results of the user study with 10 subjects

| | GM | SCC | ACC |
|---|---|---|---|
| $x$-pos. error median $\tilde{e}_x$ in mm | 0.21 | 0.18 | 0.24 |
| $z$-pos. error median $\tilde{e}_z$ in mm | 1.15 | 0.24 | 0.36 |
| completion time median $\tilde{t}$ in s | 21.4 | 20.2 | 20.1 |

The overall performance of the ten subjects is shown in Figure 5.4f–h summarizing the results in terms of positioning error in lateral and axial direction, $e_x$ and $e_z$, respectively, as well as completion time $t$. As depticted in Figure 5.4f, positioning in the lateral direction yields an error median of $\tilde{e}_x \leq 0.24\,\mathrm{mm}$ (see Table 5.1). By contrast, the axial positioning accuracy is drastically reduced when GM is used (see Figure 5.4g).

Regarding statistical evidence, the finding of differences in the variances is as important as the finding of differences in the mean or the median. Thus, the three error distributions of $e_z$ are compared by applying the non-parametric Brown-Forsythe test [BF74]. It reveals that the variances of SCC and ACC are significantly lower compared to the variance of GM. Associated $p$ – values are $p = 2 \times 10^{-7}$ and $p = 5 \times 10^{-8}$, respectively. Most of the subjects were not able to position the pyramidal structure accurately in the focal plane when GM is used. The superior performance of SCC and ACC is substantiated by the increased error median of GM that amounts to $\tilde{e}_z = 1.15\,\mathrm{mm}$. Finally, it seems obvious to statistically compare SCC and ACC although both concepts provide similar error medians of $\tilde{e}_z = 0.24\,\mathrm{mm}$ and $\tilde{e}_z = 0.36\,\mathrm{mm}$, respectively. Since equality of the

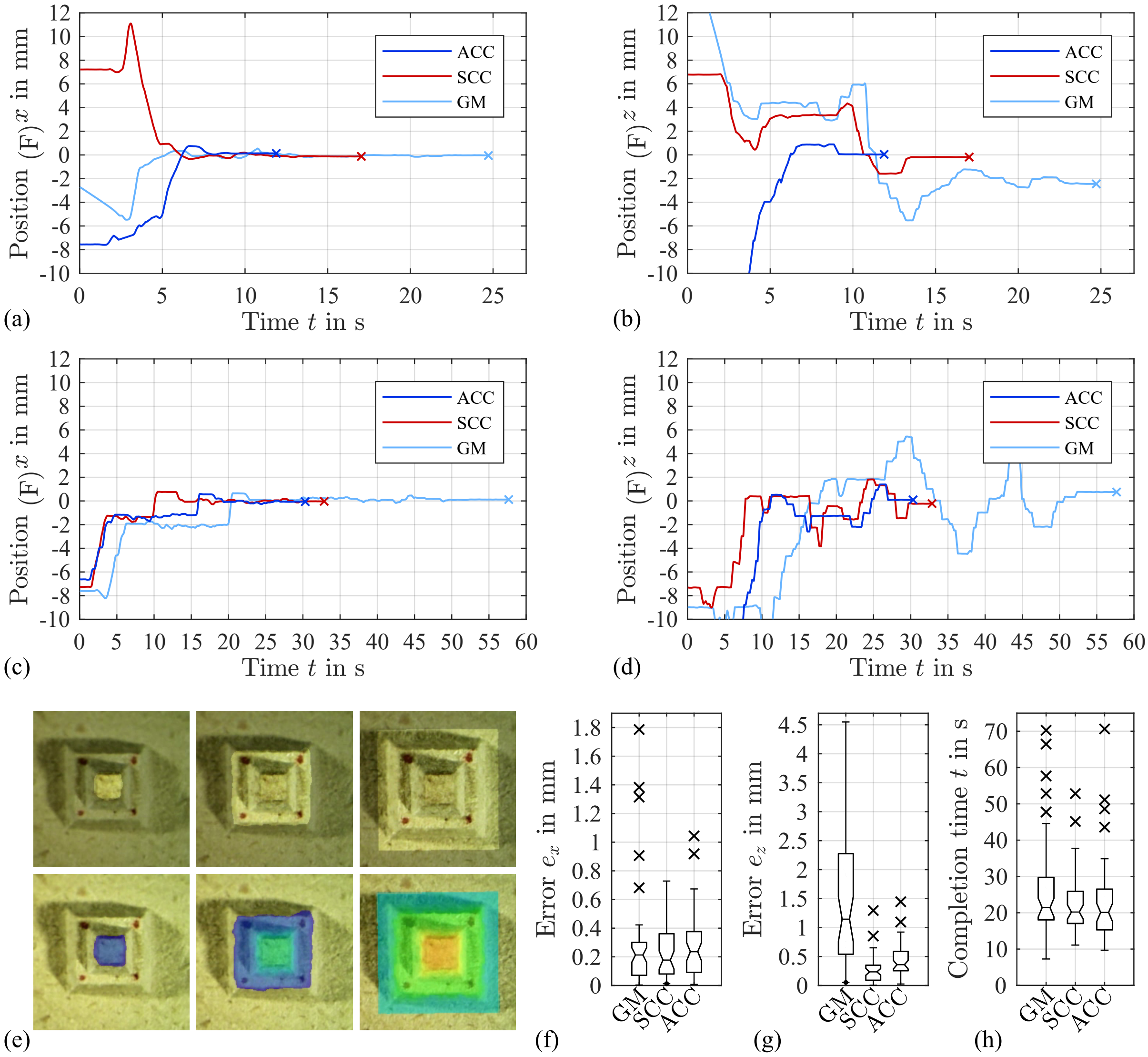


Figure 5.4: Positioning results in $_{(F)}x$ – direction and $_{(F)}z$ – direction shown for two subjects: Subject 1 (a,b) and Subject 2 (c,d). Task completion is marked by symbol ×. Image overlays GM and ACC are demonstrated in (e) for moving the pyramid from an out-of-focus position into the laser focal range (from left to right). Final results of the user study are shown as box plots for the positioning error in $_{(F)}x$ – direction (f) and $_{(F)}z$ – direction (g) as well as the completion time (h).

variances cannot be rejected (Brown-Forsythe test gives $p = 0.298$) and since the sample sizes are equal, the non-parametric Mann-Whitney $U$ test can be applied revealing significantly different error medians ($p = 2.6 \times 10^{-3}$). In other words, the subjects were able to position the pyramid more accurately when they used SCC. However, no significant differences are observed regarding the completion time. As listed in Table 5.1, the median time values are similar.

### 5.1.4 Transfer to an Endoscopic Scenario

In addition to the user study with aforementioned setup, the visualization concepts are transferred to an intraoperative scenario and evaluated under clinical conditions in collaboration with the surgeons involved in the μRALP project (see Figure 5.5) [uRA15]. Regarding the developed endoscopic prototype, the focal range of the integrated laser is limited to a few millimeters. The endoscopic laser beam is deflected by an integrated micro-robotic mirror unit and controlled by trifocal visual servoing [KSKO15, RTR+16]. As testing the developed endoscopic laser system was restricted to cadaver trials, the distance-encoding concepts are additionally assessed on *in vivo* videos acquired with a commercial stereo endoscope.

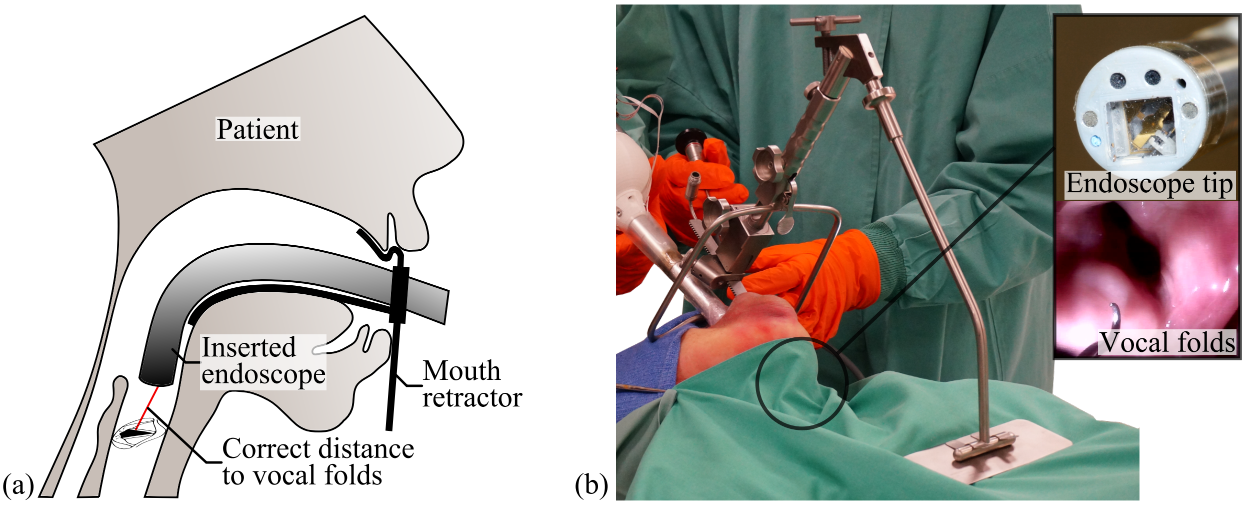


Figure 5.5: Intraoperative scenario of vocal fold surgery with (a) inserting the μRALP laser endoscope to a human cadaver providing endoscopic images of the larynx. The endoscopic tip shown in (b) integrates a stereo camera and a fiber-guided laser with deflection unit.

#### Registration of the Endoscopic Laser

In contrast to the laser setup utilized in Section 5.1.1, the intraoperative scenario demands for a modified registration of the trifocal setup (see Figure 5.6a). The reason is that laser spot positioning is performed by micro-mirror rotation around two axes for deflecting the laser beam. Consequently, the laser workspace has to be redefined by a pyramid that is truncated at the maximum allowable distance with respect to the focal plane. As shown in Figure 5.6b, the lateral boundaries are defined by planes $\boldsymbol{\pi}_i$ with $i \in \{1, \ldots, 4\}$. Due to manufacturing and assembly inaccuracies, the position of the mirror pivot point and the direction of the non-deflected laser beam are not exactly known; thus, they have to be determined to guarantee accurate color-encoding of the workspace in the live view.

In the following, a method is proposed for computing the required homogeneous transform ${}^{\mathrm{P}}\boldsymbol{T}_{\mathrm{L}}$ that maps image data from camera coordinate frame $(\mathrm{CF})_{\mathrm{L}}$ to coordinate frame $(\mathrm{CF})_{\mathrm{P}}$ of the pivot

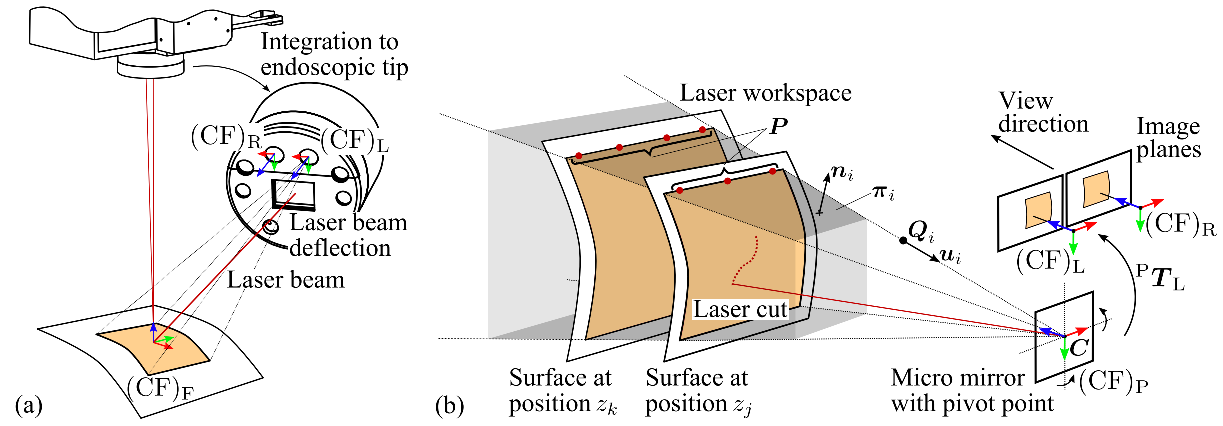


Figure 5.6: Trifocal configuration integrated into the endoscope (a). The laser workspace is modeled by a truncated pyramid (gray colored) as a result of laser scanning around two axes of the micro-mirror (b). The red dots are the laser spot positions $\boldsymbol{P}$ sampling the workspace border at plane $\pi_i$.

point with origin $\boldsymbol{C}$ describing the center of rotation. The registration with respect to coordinate frame $(\mathrm{CF})_{\mathrm{P}}$ consists of three steps: (1) segmentation of planes ${}_{(\mathrm{L})}\boldsymbol{\pi}_i$, (2) computation of the direction vectors ${}_{(\mathrm{L})}\boldsymbol{u}_i$ of the plane-plane intersections, and (3) estimation of the pivot point by minimizing the projection error of non-intersecting lines. In order to compute the plane ${}_{(\mathrm{L})}\boldsymbol{\pi}_i$, several spot positions ${}_{(\mathrm{L})}\boldsymbol{P}$ are sampled at varying distances while deflecting the laser beam into the lateral direction. The associated object coordinates are obtained from triangulation of the laser spot positions detected in the left and right camera view (according to Section 2.3.4). Plane ${}_{(\mathrm{L})}\boldsymbol{\pi}_i$ is then parametrized with a RANSAC-based segmentation [FB81]. Subsequently, the intersection line ${}_{(\mathrm{L})}\boldsymbol{u}_i$ between ${}_{(\mathrm{L})}\boldsymbol{\pi}_i$ and its adjacent plane ${}_{(\mathrm{L})}\boldsymbol{\pi}_j$, both defined by normal vectors ${}_{(\mathrm{L})}\boldsymbol{n}_i$ and ${}_{(\mathrm{L})}\boldsymbol{n}_j$, respectively, is computed by

$$ {}_{(\mathrm{L})}\boldsymbol{u}_i = \frac{{}_{(\mathrm{L})}\boldsymbol{n}_i \times {}_{(\mathrm{L})}\boldsymbol{n}_j}{\left\| {}_{(\mathrm{L})}\boldsymbol{n}_i \times {}_{(\mathrm{L})}\boldsymbol{n}_j \right\|_2} \tag{5.3} $$

pointing to the mirror pivot point ${}_{(\mathrm{L})}\boldsymbol{C}$. As a result, the matrix $(\boldsymbol{I} - {}_{(\mathrm{L})}\boldsymbol{u}_i \, {}_{(\mathrm{L})}\boldsymbol{u}_i^{\mathrm{T}})$ projects every point to the plane that passes through the origin and is orthogonal to ${}_{(\mathrm{L})}\boldsymbol{u}_i$. In particular, the intersection of ${}_{(\mathrm{L})}\boldsymbol{u}_i$ with this plane is

$$ {}_{(\mathrm{L})}\boldsymbol{Q}_{\mathrm{p},i} = (\boldsymbol{I} - {}_{(\mathrm{L})}\boldsymbol{u}_i \, {}_{(\mathrm{L})}\boldsymbol{u}_i^{\mathrm{T}}) \, {}_{(\mathrm{L})}\boldsymbol{Q}_i, \tag{5.4} $$

where ${}_{(\mathrm{L})}\boldsymbol{Q}_i$ denotes an arbitrary point on the line ${}_{(\mathrm{L})}\boldsymbol{u}_i$. In order to estimate the rotation center ${}_{(\mathrm{L})}\boldsymbol{C}$ as the nearest point to the most-likely non-intersecting lines ${}_{(\mathrm{L})}\boldsymbol{u}_i$, the projection error described by the sum of squared distances

$$ e\left({}_{(\mathrm{L})}\boldsymbol{C}, {}_{(\mathrm{L})}\boldsymbol{Q}_i, {}_{(\mathrm{L})}\boldsymbol{u}_i\right) = \sum_{i=1}^{4} \left\| (\boldsymbol{I} - {}_{(\mathrm{L})}\boldsymbol{u}_i \, {}_{(\mathrm{L})}\boldsymbol{u}_i^{\mathrm{T}}) {}_{(\mathrm{L})}\boldsymbol{C} - {}_{(\mathrm{L})}\boldsymbol{Q}_{\mathrm{p},i} \right\|_2^2 \tag{5.5} $$

is minimized with respect to the laser pivot point ${}_{(\mathrm{L})}\boldsymbol{C}$. The unique solution in the least squares sense is found by

$$ {}_{(\mathrm{L})}\boldsymbol{C} = \left( \sum_{i=1}^{4} (\boldsymbol{I} - {}_{(\mathrm{L})}\boldsymbol{u}_i \, {}_{(\mathrm{L})}\boldsymbol{u}_i^{\mathrm{T}}) \right)^{+} \left( \sum_{i=1}^{4} {}_{(\mathrm{L})}\boldsymbol{Q}_{\mathrm{p},i} \right) . \tag{5.6} $$

The laser axis in zero position is defined as the central ray of the lines ${}_{(\mathrm{L})}\boldsymbol{u}_i$ that passes through ${}_{(\mathrm{L})}\boldsymbol{C}$. Finally, the mirror position and its orientation are concatenated to the homogeneous transform ${}^{\mathrm{P}}\boldsymbol{T}_{\mathrm{L}}$ describing the transfer of a tissue surface point ${}_{(\mathrm{L})}\boldsymbol{P}$ detected in the frame $(\mathrm{CF})_{\mathrm{L}}$ to the coordinates ${}_{(\mathrm{P})}\boldsymbol{P}$ in the mirror frame $(\mathrm{CF})_{\mathrm{P}}$.

In the surgical scenario, point ${}_{(\mathrm{P})}\boldsymbol{P} = ({}_{(\mathrm{P})}x, {}_{(\mathrm{P})}y, {}_{(\mathrm{P})}z)^{\mathrm{T}}$ is considered as workspace inlier if positioned at proper distance, i.e., if ${}_{(\mathrm{P})}z$ is within the laser focal range, and furthermore if the condition

$$ {}_{(\mathrm{P})}\boldsymbol{n}_i^{\mathrm{T}} \, {}_{(\mathrm{P})}\boldsymbol{P} - {}_{(\mathrm{P})}d_i \leq 0 \tag{5.7} $$

is fulfilled for all four planes $\boldsymbol{\pi}_i$ each defined by normal $\boldsymbol{n}_i$ and distance $d_i$ from the origin.

### Image Sequences

Based on the findings described in Section 5.1.3, color coding concepts SCC and ACC are applied to endoscopic image data. The vocal fold images are obtained from two laryngoscopic settings taking aforementioned pyramidal workspace model into account. Two videos were captured with a stereo endoscope (VSii, Visionsense, Petach-Tikva, Israel) in an *in vivo* laryngeal intervention, conducted by Prof. Giorgio Peretti from the Department of Otorhinolaryngology, University of Genoa, Italy. The data collection was part of the µRALP project [uRA15]. Those videos are denoted as SEQ1 and SEQ2. Sequence SEQ1 contains axial endoscopic movements through the laryngoscope (see Figure 5.7a). Fine positioning in the axial and the lateral direction with respect to the tissue surface is shown in sequence SEQ2 (see Figure 5.7d).

Figure 5.5b shows the sequence SEQ3 representing a first human specimen trial with the flexible µRALP laser endoscope [KSKO15, RTR$^+$16]. The endoscope contains an integrated chip-on-the-tip stereo camera (MO-BS0804P, MISUMI Electronics Corp., Taiwan). The tip prototype and an endoscopic image from the cadaver are shown in Figure 5.5b. In comparison to the setup described in Section 5.1.1, the endoscopic laser requires focusing at a shorter distance. The prototype utilized in this dissertation consists of a two lens focusing system integrated to the endoscopic head. Since its Rayleigh length is estimated to approx. 2 mm [KFS$^+$16], the color coding range is reduced to $\left|{}_{(\mathrm{F})}z\right| \leq 2\,\mathrm{mm}$ for the endoscopic sequences shown in Figure 5.7. In the lateral direction, sequences SEQ1–3 demonstrate that the laser cannot reach the entire surgical field of view. The workspace is currently restricted strongly by the micro-mirror deflection angle.

### Results of the Endoscopic Scenario

For sequence SEQ1 representing axial movements of the endoscope inside the patient's larynx, the proposed color gradient ACC provides adequate visual feedback on the position of the laser focal range (see Figure 5.7a–c). In addition, chromadepth color coding enhances depth perception of underlying tissue structure. Far distance is encoded by the color blue whereas close objects are highlighted in red. Fine positioning with respect to the vocal fold tissue is illustrated by SEQ2 in Figure 5.7d–f. The benefit of the color-encoding is underlined by the positioning example of the virtually defined target (white circle). Even though the two endoscopic views in Figure 5.7d and Figure 5.7e appear similar, ACC reveals a distance deviation of approx. 2 mm in axial direction. The target would be fully focused for optimal laser ablation if positioned as shown in Figure 5.7e. In addition to the overlay ACC, the symmetric color-gradient SCC provides clear recognition of the laser focal plane intersecting with the curved shape of the tissue whereas the depth perception effect is less distinct (see Figure 5.7f).

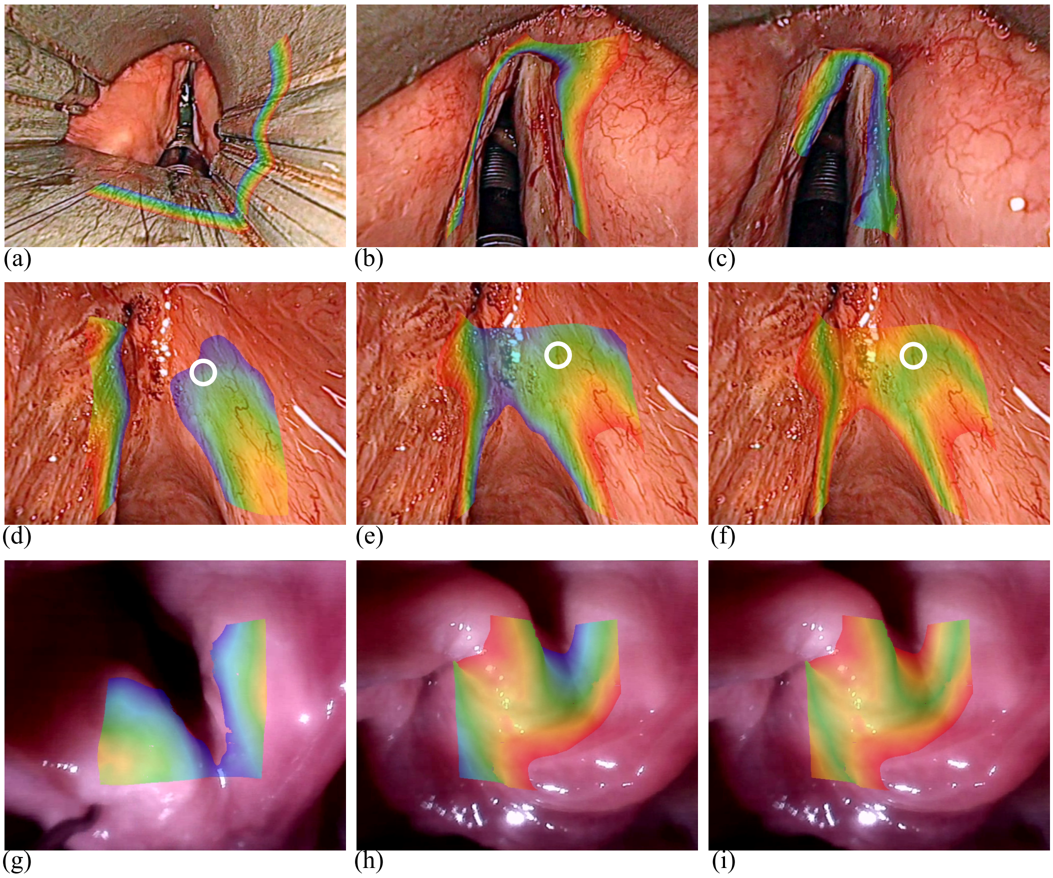

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 5.7: Asymmetric color coding ACC during insertion to patient larynx is shown on sequence SEQ1 (a-c), fine positioning of laser focus to a virtual target (white circle) deploying ACC and SCC on sequence SEQ2 (d-f), and the human specimen sequence SEQ3 (g-i). SEQ1–2 were provided by Prof. Giorgio Peretti, Department of Otorhinolaryngology, University of Genoa, Italy.

Figure 5.7g–i illustrate the application of the color coding to SEQ3 representing vocal fold tissue of a human cadaver. As proof of concept, the method is able to highlight the focus range even on sparsely textured tissue (no vessels visible due to loss of blood circulation) and even when deploying low cost cameras integrated to the endoscopic tip. The symmetric overlay SCC in Figure 5.7i provides clear visual feedback on the green-colored focal plane similar to Figure 5.7f.

## 5.2 OCT Imaging and Visualization of Subepithelial Tissue Structure

OCT facilitates high-resolution, cross-sectional tomographic imaging of the subepithelial space, such that benign or malignant entities can be localized. Especially in laryngeal laser interventions, the surgeon will benefit from a framework fusing high-resolution, OCT-based tumor imaging with white light stereo vision. Volumetric OCT data can be segmented providing tumor information that is intuitively visualized in the live view. Despite its clinical application, recent research focuses on the design and fabrication of phantoms mimicking the optical, mechanical, and structural properties of biological tissues [LKK$^+$12]. In this context, the use of a realistic phantom enables continuous evaluation of the required image processing and visualization without the necessity of *in vivo* trials in early development stages.

This section introduces an OCT-based augmented reality concept based on the registration presented in Section 4.2. A segmentation framework is developed and demonstrated by means of a phantom modeling the optical properties of laryngeal tissue. To assist during the resection planning, a color gradient ranging from maximum function preservation to maximal radicality is proposed.

### 5.2.1 Phantom Design and Segmentation in OCT Images

The upper section of the vocal folds consists of the thin translucent epithelium with an intact basement membrane followed by the superficial layer of the lamina propria with the elastic vocal ligaments [KGvG$^+$08]. Below, there is the thyroarytenoid muscle that commonly cannot be imaged with OCT due to limited penetration depth [KGvG$^+$08].

In this section, two scenarios are considered that require volumetric imaging to determine the extent of the abnormal changes in the layer-based structure of the vocal folds. The first scenario is the invasive carcinoma which is characterized by the basement membrane being no longer recognizable within the OCT scan. The second type is associated to translucent, subepithelial cysts or polyps which are characterized by high absorption of light from the OCT.

#### Design and Fabrication of the Soft Tissue Phantom

Due to the manifold properties of human tissue, i.e., in its micro and macroscopic structure, its optical characteristics significantly vary when analyzed with OCT. Thus, phantom design aims at finding an optimal configuration of a bulk matrix acting as scaffold and additives defining the desired optical characteristics of laryngeal tissue.

The phantom is designed as a single layer rectangular block with the dimensions of $(60 \times 30 \times 5)\,\mathrm{mm}^3$. The carrier bulk structure is composed by a two-part room-temperature-vulcanization silicone (SF45 2k-Silikon, Silikonfabrik, Germany). Basic scattering and absorption properties are iteratively determined by adding titanium dioxide ($TiO_2$) powder (titanium (IV) oxide, Sigma-Aldrich, St. Louis, Missouri, US) at a concentration of 5 mg/ml.

To mimic the absorption entity, separate silicone-based samples with carbon nanoparticles (Carbon graphitized, Sigma Aldrich, US) added at a concentration of 10 mg/ml had been fabricated. Including them into the bulk material simulates the transition between the vocal fold epithelium and a cyst located below. By contrast, further specimens with an increased $TiO_2$ concentration, i.e., 100 mg/ml, had been made. Those were used to replicate the transition zone between an intact subepithelial basement membrane and its loss simulating invasive cancer. Even though the materials are well discussed in literature [CKS11, LKK+12], they do not provide realistic texture information. Thus, skin-colored pigment powder (skin tone powder European, Silikonfabrik, Germany) resembling human tissue appearance has been added at a quantity of up to 0.1 g of the entire sample. Finally, the absorbing and scattering samples are embedded into the not yet completely cured bulk structure. For further details, the reader is kindly referred to [CGS+15].

### Image-Based Segmentation Algorithm

The image-based segmentation of the tissue-mimicking phantom consists of two steps: (1) segmentation of the up-most epithelium layer surface according to Section 4.2, and (2) localization of the embedded subepithelial scattering or absorption structures.

As depicted in Figure 5.8, a segmentation framework was implemented. It follows the assumption that the most descriptive information about the pathological structure is located within the subep-

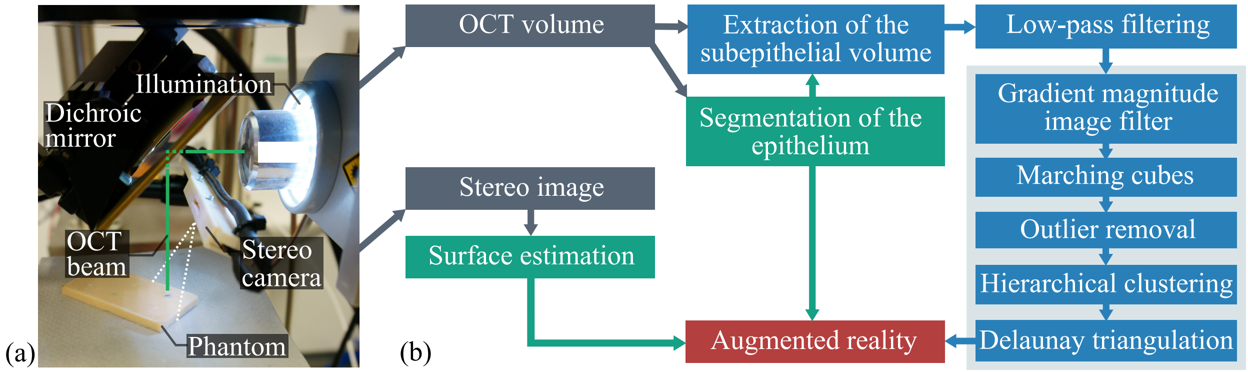


Figure 5.8: Image acquisition setup with OCT and stereo camera (a). Phantom images are processed as shown in (b). The proposed segmentation work flow is highlighted in blue.

ithelium layers and has distinct optical characteristics. The detected epithelial surface layer provides an initialization for the second segmentation step, which involves the extraction of the subepithelial volume. Subsequently, the entire sub-volume is low-pass filtered in the frequency-domain after applying the fast Fourier transform. As a result, background speckles are suppressed. The gradient magnitude image filter is applied to segment the potential boundaries of the volume characterized by differing scattering or absorption properties due to the embedded specimens causing abrupt changes in the $TiO_2$ or carbon particle concentration. In order to generate a polygonal mesh out of the gradient information, the marching cubes algorithm is applied with a threshold set by the user [LC87]. This is necessary because the degree of the intensity change on the detected borders varies between the data sets due to different penetration depth and inhomogeneous additive concentration.

Morphological filtering eliminates outliers such as air cells remaining in the bulk after curing. Over-segmentation is avoided by a ratio-based hierarchical clustering with subsequent Delaunay triangulation of the final polygonal surface mesh. The size ratio between two segments and their distance defines their corresponding cluster. This allows larger segments to be further apart, but still be assigned to the same clustered like small and closer segments.

The image processing pipeline is implemented on Ubuntu $12.04$ LTS deploying the Insight Segmentation and Registration Toolkit [ITK17], and the Visualization Toolkit [VTK17].

### 5.2.2 Visualization of the Subepithelial Segmentation

Volumetric images of the fabricated phantoms are acquired with the SD-OCT device (see Section 2.5) and aligned to the stereo view considering the registration of OCT and stereo camera according to Section 4.2. The scanning range has been set to $(10 \times 10 \times 2.76)\,\mathrm{mm}^3$ with an image definition of $256 \times 256 \times 1024$ voxels, resulting in a depth resolution of $2.69\,\mu\mathrm{m}$. Subsequent to OCT imaging of the tissue phantom, the developed subepithelial segmentation algorithm is demonstrated by two visualization concepts that are described in the following.

The first one is dedicated to projecting the contour of the segmented subepithelial structure, i.e., the polygonal mesh, to the left and right image view. The distance with respect to the reconstructed surface point cloud is computed and color-encoded in the live view. In detail, visual augmentation is achieved by mapping the distance value onto spectral color information, i.e., considering hues going from red (close distance) to blue (far distance). A highlighted color bar illustrates the distance in millimeters. In the context of laser surgery, the projection of the outer contour provides clear resection margins.

In the second concept, the distance is not only highlighted in depth direction, i.e., the shortest distance with respect to the epithelium layer, but also in the lateral direction. On the one hand, this augmentation concept makes the surgeon aware of uncertainties in the detection of subepithelial pathologies. On the other hand, it intuitively provides visual feedback on how to find an adequate resection margin as the color gradient ranges from maximum function preservation (red) to maximal radicality (green).

### 5.2.3 Results

In the following, the results of the phantom design and the subepithelial segmentation are presented.

#### OCT Phantom

Multiple phantoms have been designed to determine the impact of selected additives and pigments on OCT scattering. A phantom mimicking the vocal fold tissue is depicted in Figure 5.9a. Small fragments of $TiO_2$ and carbon, that are embedded into the bulk, are shown in Figure 5.9b.
As shown in Figure 5.9c–f, the OCT scans of the phantom are compared with related human vocal fold images provided by the literature [ARV$^+$06, KGvG$^+$08]. A two layer-based structure has been achieved by utilizing a $TiO_2$ concentration discontinuity. The related OCT scan of the phantom is shown in Figure 5.9c. In particular, the epithelium (e) and the basal membrane (bm) are successfully replicated according to Figure 5.9d. Similar to that, the use of carbon nanoparticles facilitates reduction of scattering (see Figure 5.9e). This allows resembling for instance the transition between the epithelium layer and a vocal fold cyst, as depicted in Figure 5.9f. Beside this area, the main bulk structure shows desired scattering properties due to the minimal concentration of $TiO_2$ added to the silicone.

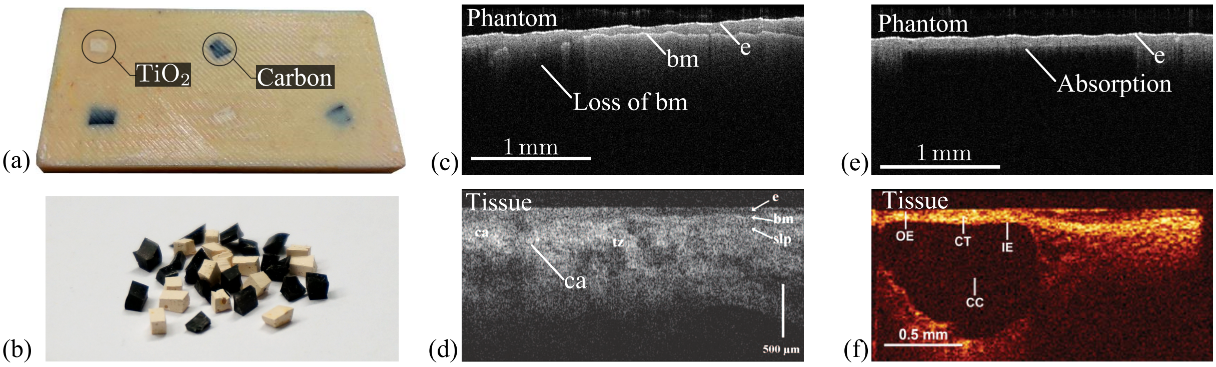


Figure 5.9: The phantom shown in (a) is made of silicone bulk and includes fragments of $TiO_2$ and carbon depicted in (b). OCT scans in (c) and (e) demonstrate tissue-mimicking properties of invasive cancer and cysts, respectively, that occur within the subepithelial layer of vocal fold tissue (e–epithelium, bm–basal membrane). OCT imaging of the human vocal folds of (d) squamous cell carcinoma (e–epithelium, b–basement membrane, slp–superficial lamina propria, tz–transition zone, ca–cancer) (reprinted from [ARV$^+$06], © 2006, with kind permission from John Wiley and Sons), and (f) of a retention cyst (CC–cyst) (reprinted from [KGvG$^+$08], © 2008, with kind permission from John Wiley and Sons).

### Image-Based Segmentation Algorithm

The segmentation of an embedded carbon bit is exemplarily illustrated in Figure 5.10a–c. The segmentation algorithm, with minimal user interaction, successfully determines the subepithelial absorbing entity. Related visualization with color-encoded distance is illustrated in Figure 5.10d–f. In this context, the results of six OCT-based segmentations that were superimposed to the live view are shown in Figure 5.11. In Figure 5.11c, a larger air cell is detected as false positive whereas, in general, smaller speckles are successfully eliminated by morphological filtering.

The first attempt of injecting a liquid into the viscous silicone, e.g., for resembling a cyst, did not result in adequate specimens since the liquid instantly migrated upwards during curing. Trials on embedding different types of non-silicon fragments generated impenetrable layers for OCT imaging. Finally, using already cured fragments of $TiO_2$ and carbon turned out to be successful for resembling the anatomical properties in an abstract manner. Further investigation is required not only in terms of replicating the optical properties of laryngeal tissue (without air cells) but also fabricating multiple layers in an anatomical arrangement with the results that the surface mimics the three-dimensional shape of vocal folds.

Regarding the proposed segmentation algorithm, promising results are achieved, although the current OCT penetration depth is limited to approx. $0.5\,\mathrm{mm}$. As a result, a few of the embedded

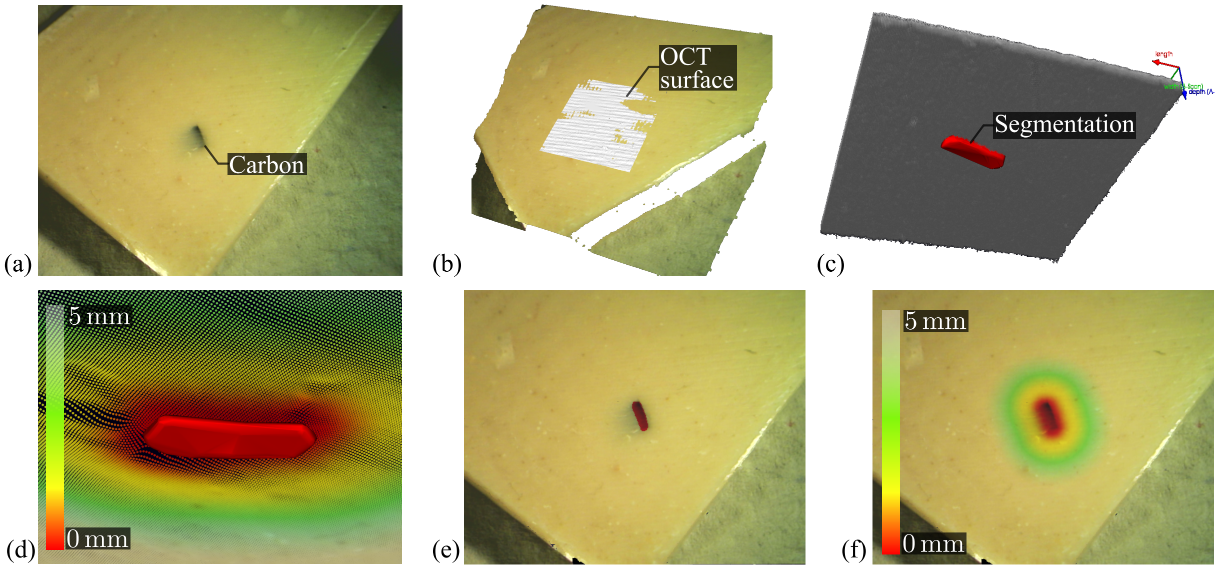
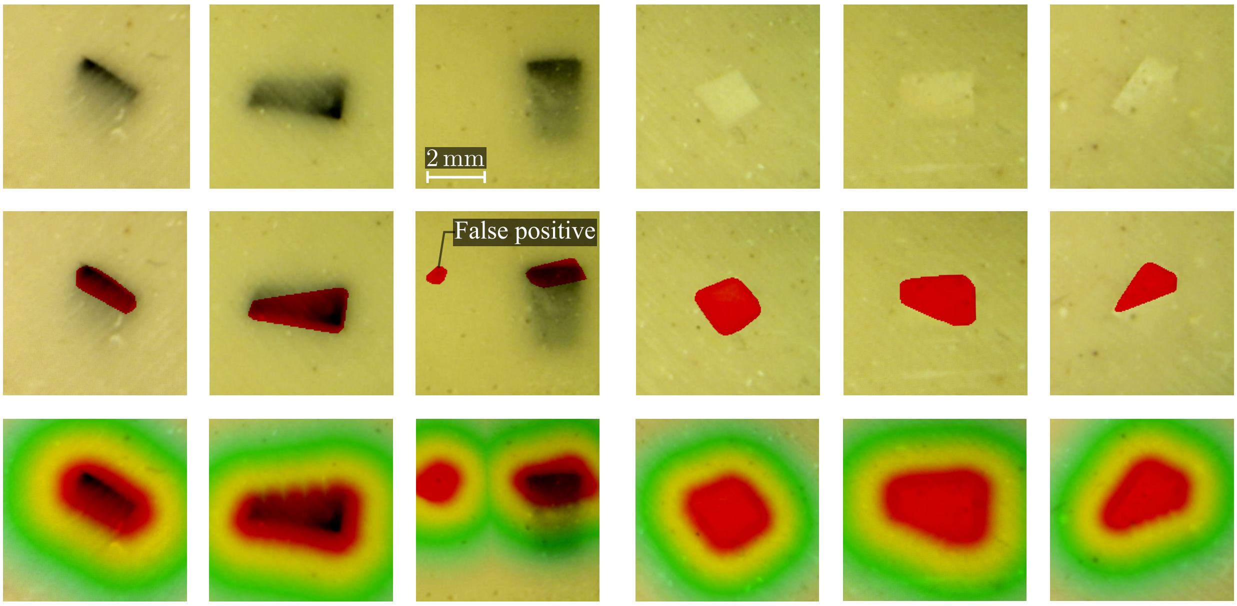


Figure 5.10: The phantom embedding a carbon bit is depicted in (a). The OCT surface overlaid to the camera-based reconstruction shown in (b) demonstrates successful registration of the two modalities. The subepithelial lesion embedded in the OCT scan (c) is superimposed as polygonal mesh to the camera-based reconstruction as illustrated in (d) (view from below the surface). Augmented visualization is achieved by (e) projecting the contour and distance to the epithelium layer or by (f) the extended color-coded distance.

$TiO_2$ and carbon fragments cannot be segmented in their entirety. Instead, only sub-volumes are precisely detected (see Figure 5.11c). Furthermore, stereo-based surface reconstruction requires sufficient texture information. However, the concentration of the color pigment to be added strongly impairs the desired penetration depth. Thus, a tradeoff has to be found with respect to the tissue-like appearance and the visibility of the embedded entities shining through the surface layer, i.e., the carbon samples (see Figure 5.11a–c).

However, the implemented framework including registration, segmentation and multimodal fusion is successfully demonstrated by the results shown in Figure 5.11. In the context of laryngeal laser surgery, the augmented reality concept provides visual feedback during resection planning allowing for accurately defined safety margins of the incision. Due to the limited penetration depth, the segmentation and visualization concept has to be applied in an iterative manner. In other words, an OCT image is acquired once before the next resection is planned and executed to gradually expose the subepithelial lesion.

In future work, the segmentation algorithm has to be extended in order to highlight the loss of the basement membrane as indicator of invasive squamous cell carcinoma. To provide realistic tissue phantoms for laser surgery, e.g., for surgical training purposes, further materials enabling simultaneous laser ablation of the detected pathologies need to be investigated.



(a) (b) (c) (d) (e) (f)

Figure 5.11: Magnified camera view (top row) and superimposed segmentation (middle and bottom row) with carbon (a-c) and $TiO_2$ (d-f).

## 5.3 Tablet-Based Strategies for Incision Planning in Laser Microsurgery

The literature review in Section 1.2 has shown that laser control is superior in precision and ergonomics when employing a stylus-tablet-based planning interface in both microscopic and endoscopic laser microsurgery.

However, those studies focus on comparing the conventional micromanipulator with the tablet interface when continuously tracing a path. By contrast, a detailed benchmark on concepts for planning with a stylus, i.e., for various visualization techniques, has not been discussed so far. When introducing computational stereo methods, real time augmentation is feasible not only in the monoscopic view, but also in the stereo or even a virtual laser view. For instance, both the incision line and the highlighted laser workspace (see Section 5.1) can be correctly aligned to the left and the right live view without compromising depth perception.

In this section, a comparative study of six stylus-tablet-based concepts for visualization and incision planning in laser microsurgery are presented. Ablations are planned and performed while user performance is assessed by means of accuracy, completion time, and ease of use.

### 5.3.1 Incision Planning Strategies

Six incision planning strategies are implemented utilizing the laser setup with attached stereo camera, as shown in Figure 5.12a. The scene is visualized either monoscopic, stereoscopic, or as synthesized laser view (see Section 4.1). For stereoscopic visualization, a 3D monitor (VG278, ASUS, Taiwan) and active shutter glasses (3D Vision 2, Nvidia Corporation, Santa Clara, CA, USA) are used. As illustrated in Figure 5.12b and 5.12c, the framework enables incision planning with a writing interface either on a pen display or on a graphics tablet (DTU-2231 and Bamboo CTH-470, Wacom Co., Ltd., Japan). The laser interface, the image acquisition and related processing are implemented deploying the Robot Operating System (ROS) as high-level control layer [QGC$^+$09].

In the user study, ablations are performed on tissue substitutes, namely medium density fiberboards

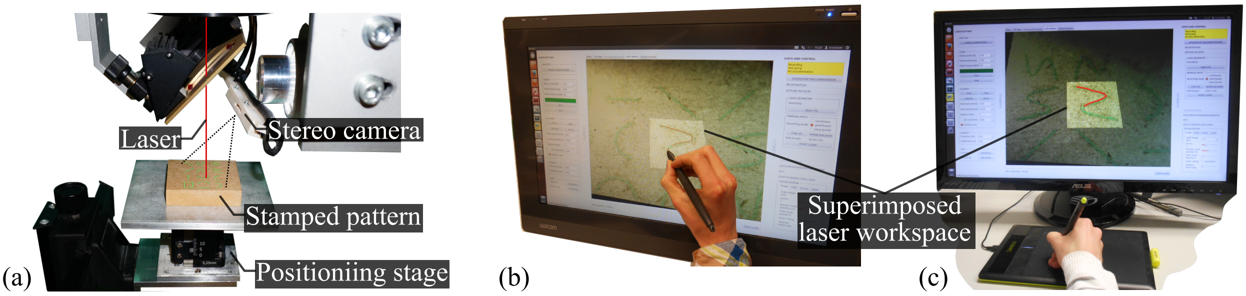


Figure 5.12: The surgical laser setup with stereo camera and stamped pattern in (a) is operated either with (b) a pen display or (c) a graphics tablet and a (2D/3D) monitor.

(MDF), comprising 13 stamped straight lines and curves (see Figure 5.13a). The reference strategy is defined by planning in the mono view whereas the path is continuously drawn on the graphics tablet. The moving cursor on the screen is directly mapped to the laser workspace considering the laser-to-camera registration (see Section 4.1). The obtained path is characterized by a vector of $M$ points (see Figure 5.13b). The same planning strategy is applied to the stereoscopic and the synthesized laser view.

In contrast to continuous drawing, a point-based planning that concatenates straight line segments to a curved shape is proposed as the fourth concept. Subsequent to defining a point $\boldsymbol{p}_k$, the following partial path is spanned to the current cursor position as shown in Figure 5.13c. The point $\boldsymbol{p}_{k+1}$ is set by the user connecting $\boldsymbol{p}_k$ and $\boldsymbol{p}_{k+1}$ by a straight line that fits best to the desired trajectory, i.e., to the stamped curve. Compared to continuous planning, the number of required user inputs with the stylus is reduced ($N \ll M$).

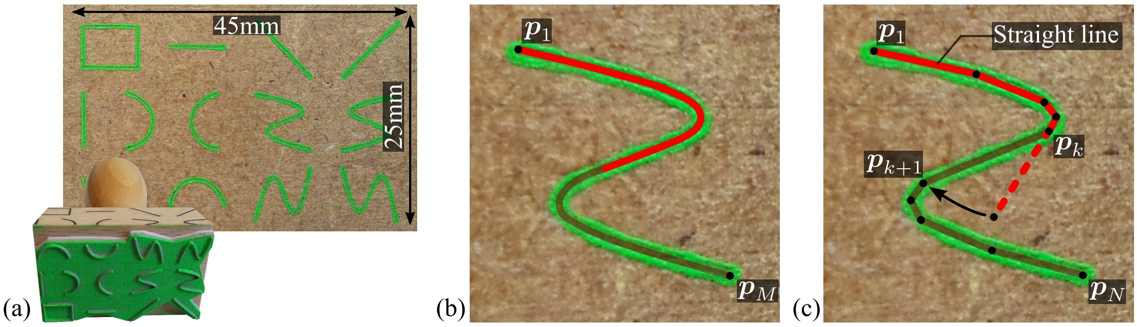


Figure 5.13: The stamp and the colored pattern used for path tracing are shown in (a). The path planning is either performed by (b) continuous or (c) point-based drawing with the stylus. In the latter case, the planned incision is concatenated by straight line segments of user-defined length.

The fifth concept consists of simultaneous path drawing and ablation using the mono view and the graphics tablet, denoted by real-time teleoperation. Improper movement of the stylus instantly leads to irreversible and inaccurate laser ablation. To investigate the impact of hand-eye-coordination, the sixth strategy deploys the pen display with continuous path drawing in the mono view (see Figure 5.12c); thus, the tablet and the visualized pattern are within the same field of view. The user can keep an eye on both, the path to be planned and the moving stylus tip. Hence, more accurate planning is expected.

The six stylus-tablet-based strategies are summarized as follows:

- Continuous path planning on the graphics tablet using the mono (MV), the virtual laser (LV), or the stereo view (SV) with subsequent ablation
- Point-based path planning (PB) on the graphics tablet using the mono view with subsequent ablation

- Continuous path planning on the graphics tablet using the mono view with real-time ablation (RT)
- Continuous path planning on the pen display (PD) with subsequent ablation

### 5.3.2 User Study Design

In the user study, six subjects with a background in medical engineering were asked to plan multiple laser ablation paths. The strategies were performed in a random order for counterbalancing learning effects. To provide equal conditions to left and right-handed participants, stamped curves were mirrored, rotated and duplicated (see Figure 5.13a). The results are discussed in terms of the root mean square error (RMSE) and the maximum distance error (MDE) between the desired and the executed trajectory. In this context, image-based path detection is accomplished by color-based thresholding and subsequent thinning of the stamped and the ablated patterns [ZS84]. Additionally, the completion time required for drawing and ablation is measured. The laser power is set with the result that clearly visible cuttings are achieved (diode current $I_d = 120\,\mathrm{A}$, pulse duration $t_p = 220\,\mu\mathrm{s}$, pulse frequency $f_p = 50\,\mathrm{Hz}$).

The usability was assessed by carrying out an After Scenario Questionnaire (ASQ) subsequently to each strategy. Finally, the System Usability Scale Questionnaire (SUS) was answered by the subjects in order to rate the overall stylus-tablet-based planning framework. The ASQ consists of hypotheses for assessing user satisfaction in terms of ease of use and task completion time [Lew91]. Scoring is defined by seven-point Likert scale (1 - strongly disagree, 7 - strongly agree). A high rating correlates with increased user satisfaction. The evaluated ASQ hypotheses are

Hyp–1: “Overall, I am satisfied with the ease of completing the tasks in this scenario”, and
Hyp–2: “Overall, I am satisfied with the amount of time it took to complete the tasks”.

Additionally, the user can add comments. The SUS, that is based on five-point Likert scale (1 - strongly disagree, 5 - strongly agree), gives a maximum score of 100 [Bro96]. A score of 70 defines average rating whereas a score of below 50 indicates unacceptable system usability.

### 5.3.3 Results

Even though ablation was performed on all curves, a limited number of them could not be detected as consistent paths due to the chosen detection threshold. To not mix automatic and manual segmentation within the evaluation, aforementioned datasets were excluded from the conducted 468 incisions (78 per strategy). For the PD trials, only 72 measurements were evaluated (see Table 5.2).

Regarding the quantitative analysis, the Kolmogorov-Smirnov test reveals that some of the datasets do not meet the assumption of normality. Nonetheless, the sample sizes and the variances are similar and thus, the non-parametric Wilcoxon rank-sum test (significance level $p = 0.05$) is

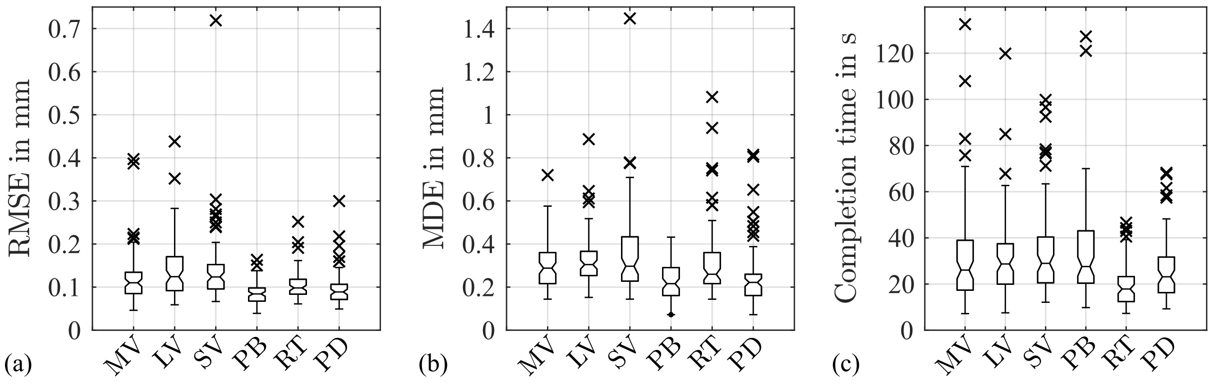


Figure 5.14: Box plots of the path tracing showing (a) the RMSE, (b) the MDE, (c) the task completion time.

Table 5.2: Path tracing results with medians of the root mean square error (RMSE), maximum distance error (MDE), and completion time. Significance in comparison with MV is given in brackets ($p$-value).

| | MV | LV | SV | PB | RT | PD |
|---|---|---|---|---|---|---|
| Detected ablations | 78 | 74 | 75 | 77 | 75 | 72 |
| RMSE in mm | $0.110$ | $0.124$ $(0.03)$ | $0.120$ $(0.02)$ | $0.084$ $(1 \times 10^{-6})$ | $0.099$ $(0.12)$ | $0.089$ $(5 \times 10^{-4})$ |
| MDE in mm | $0.288$ | $0.305$ $(0.31)$ | $0.297$ $(0.08)$ | $0.216$ $(3 \times 10^{-6})$ | $0.260$ $(0.45)$ | $0.222$ $(2 \times 10^{-4})$ |
| Completion time in s | 26.0 | 28.7 (0.48) | 29.0 (0.13) | 27.6 (0.28) | 17.8 $(7 \times 10^{-6})$ | 23.1 (0.17) |

applied to compare the outcomes of the reference strategy MV with the other concepts.

The ablation errors and related completion times are shown in the box plots in Figure 5.14 and are listed in Table 5.2. The reference strategy MV is measured with a RMSE median of $0.11$ mm, a MDE median of $0.288$ mm, and a completion time median of $26$ s. Deploying the laser (LV) or the stereoscopic view (SV) is less accurate (RMSE of $0.124$ mm and $0.120$ mm, respectively, $p \leq 0.03$) whereas path tracing with point-based planning (PB) or the pen display (PD) significantly increases the ablation accuracy by approximately $24\,\%$. The corresponding RMSE medians are given by $0.084$ mm ($p = 1 \times 10^{-6}$) and $0.089$ mm ($p = 5 \times 10^{-4}$), respectively. This observation is substantiated by the significantly decreased MDE medians ($p \leq 2 \times 10^{-4}$). Regarding the real time teleoperation (RT), the increased ablation accuracy does not reach statistical significance compared to MV ($p > 0.05$). But in contrast to the other strategies, the completion time is drastically reduced to $17.8$ s ($p = 7 \times 10^{-6}$) since planning and laser cutting are performed simultaneously. To illustrate the outcomes, exemplary cuttings conducted by one of the users are shown in Figure 5.15.

The quantitative results are emphasized by the conducted ASQ. In especially for strategy PB, a score of $6.67$ (out of seven) indicates a superior ease of use compared with MV while the user subjectively perceive a reduced task completion time (see Figure 5.16). Similar user feedback (scoring $> 6$) is obtained from PD demonstrating improved hand-eye-coordination compared to MV. However, a few subjects claim that in some situations the target visualized on the pen display is occluded by their writing hand. Thus, tracing the remaining path becomes more difficult. As

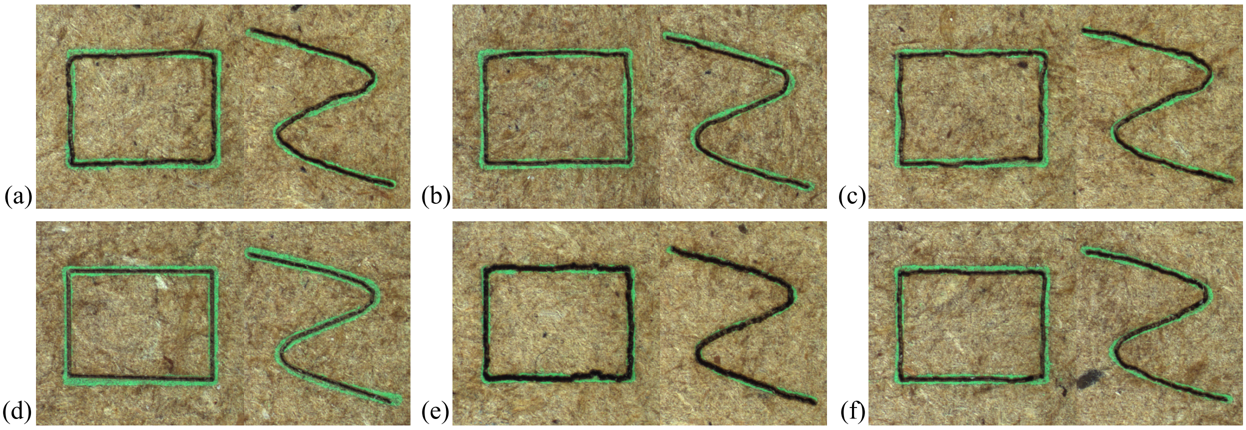


Figure 5.15: Laser cuttings (black lines) of one of the subjects tracing the stamped curves (green lines) when deploying strategy (a) MV (mono view, tablet), (b) LV (laser view, tablet), (c) SV (stereo view, tablet), (d) PB (point-based, tablet), (e) RT (real-time, tablet), and (f) PD (pen display). The shown microscopic photographs were acquired subsequently to the trials.

already observed in the quantitative measurements, RT is well accepted by the users due to the short amount of time that is required for planning and laser cutting (score of $6.83$).

Regarding the scenario SV, a reduced ease of use and thus, lower acceptance is revealed for the current implementation with active shutter glasses. The score of only $4.83$ correlates with the observed decrease in ablation accuracy. Moreover, the visualization LV is rated as good as strategy MV in terms of completion time and ease of use, even in the presence of slight jitter of the scene due to image noise that affects the synthesis of the live laser view.

Regarding the overall usability assessed with the SUS, the stylus-tablet-based planning framework is rated with an average score of $86.25$ demonstrating high acceptance among the involved subjects.

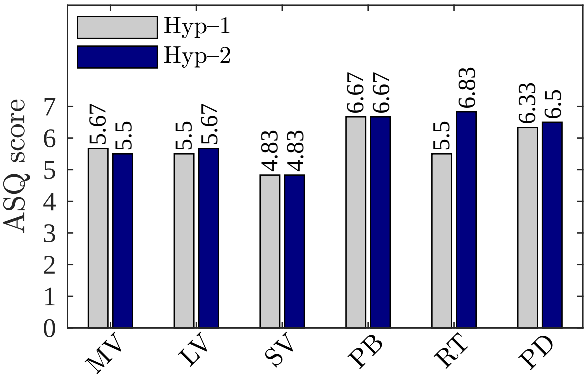


Figure 5.16: Results of the ASQ describing the overall satisfaction regarding criteria ease of use (Hyp–1) and completion time (Hyp–2). High scoring correlates with an increased user satisfaction.

## 5.4 Conclusion

In this chapter, a framework for incision planning based on augmented reality visualization of relevant laser and tissue information is presented.

To assist the surgeon during laser focus positioning, the surgical live view is superimposed with color-encoded distance information highlighting the intersection of the tissue surface with the laser workspace. For the considered color coding concepts, user studies were performed. The subjects were able to correctly align the target with the laser focal range. However, workspace highlighting with the gray mask (GM) does not provide depth information in axial direction. Thus, a few subjects had to determine the upper and lower workspace boundary before being able to position the surface near to the focal plane. By contrast, symmetric (SCC) and asymmetric color coding (ACC) directly project the laser focal range onto the live view demonstrating that accurate positioning is feasible.

If the system does not comprise motorized focusing due to spatial limitations, i.e., in endoscopic laser surgery, correct distance to the tissue has to be established manually. Strategies SCC and ACC clearly reveal accurate alignment of the laser focus as a prerequisite to atraumatic ablation. Regarding vocal fold surgery, a laser endoscope can be inserted to the larynx through the patient's mouth and positioned at proper distance to the lesion. As proof of concept, color-encoded distance information overlaid to human *in vivo* and human specimen images has been discussed.

Furthermore, the fabrication and the experimental evaluation of an optical laryngeal tissue phantom for OCT and the associated implementation of visualization strategies has been presented. The proposed work flow combines the advantages of OCT imaging (high resolution and non-invasive optical penetration) and surgical stereo vision (real-time scene depth computation). The color-encoded overlay of the segmented structure focuses the surgeon's interest solely on the live camera view without taking care of further imaging modalities and displays. This might increase the acceptance within the prospective user group. However, further optimization of the phantom fabrication and OCT-based segmentation is required. A major problem with the phantom OCT data arises from false positive segmentation of air bubbles as well as reduced penetration depth that might be attributed to the additive concentration.

Regarding incision planning, the results clearly demonstrate that even when making use of a tablet and familiar hand writing, significant differences exist among the analyzed strategies. Highly accurate path alignment is achieved with the pen display (PD) and point-based planning (PB) while no significant difference is indicated in completion time. However, prospective interfaces for laser surgery require an optimal setting of the input mode and associated visualization. For instance, combining the strategies PD and PB has the potential to become a powerful planning strategy due to the improved hand-eye-coordination and the simple technique for path definition.

# 6 Motion Compensation in Planning and Laser Ablation

Precise resection of the lesion in a soft tissue environment is a challenging task. When considering a strategy for path planning and subsequent ablation, as discussed in Section 5.3, cutting accuracy does not solely depend on the surgeon's dexterity, but also on tissue deformation occurring due to respiration artifacts or manipulation with instruments. Additionally, misalignment of the laser path and focus is evoked by the non-stiff mechanical fastening of the laser system with respect to the patient; thus, motion externally applied to the microscope most likely results in positional deviation of the laser spot. Deformations and camera motion are difficult to handle, especially when aiming at function preservation with resection margins of less than 1 mm. Consequently, deformation tracking for online adaptation of the laser spot position is mandatory.

Despite all advances in laser phonomicrosurgery including tablet-based planning interfaces or even vision-guided laser control, as discussed in Section 1.2, motion compensation for laser ablation on deforming tissue has not been addressed so far. To overcome this drawback, Section 6.1 presents a stereoscopic, non-rigid tracking scheme and its evaluation on *in vivo* data with comparison to state-of-the-art methods. In Section 6.2 and Section 6.3, practical demonstration is discussed for motion compensation during incision planning and laser ablation, respectively, whereas the tablet-based concept and the surgical laser setup are utilized (see illustration in Figure 6.1).

The methods and results of this chapter are published in [SLKO16, SKKO17].

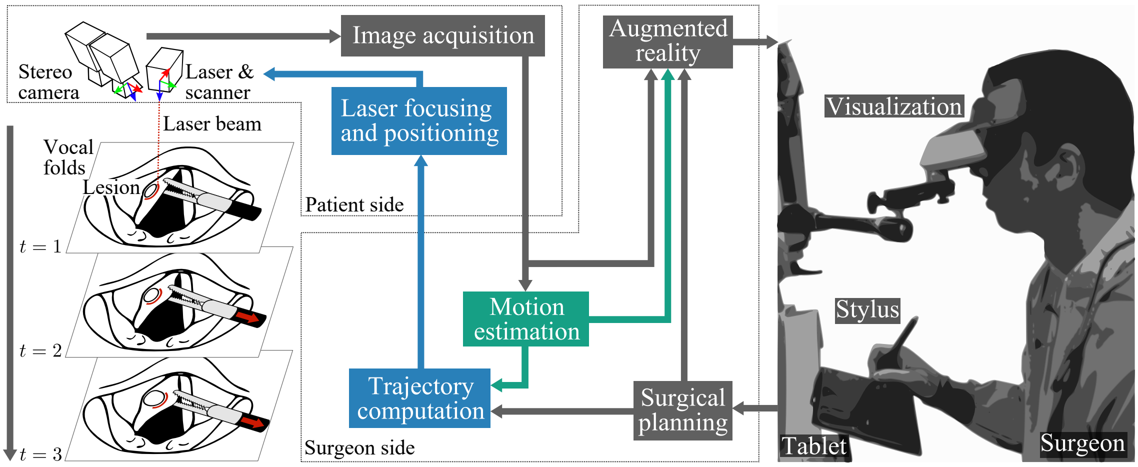


Figure 6.1: Work flow for application of motion compensation in laser phonomicrosurgery. While exposing the vocal fold lesion (oval structure) by pulling with the forceps, tissue motion is tracked to adapt online the ablation scan pattern (red line). Vision-guided laser control is intended to be integrated into a framework with stylus-tablet-based planning and augmented reality visualization.

## 6.1 Stereo Vision-Based Tracking of Soft Tissue Motion

This section presents a novel method for stereoscopic tracking of soft tissue motion. Extending a single view approach, the epipolar constraint is incorporated into a piecewise affine deformation model to enforce left-right consistency. Specific challenges such as illumination changes, occlusions, or drift are systematically addressed in respect of real time capability.

This tracking follows the idea of splitting the optimization into (1) quasi-deterministic tracking robust to illumination changes as well as partial occlusions and (2) appearance-based mesh refinement to compensate for tracking inaccuracies such as drift. Instead of sequential processing, as described in the original work [ZLH09], concurrent computation of both steps is proposed. Once convergence is reached for the mesh refinement, affine-invariant fusion with respect to the current tracking estimate is performed at minimal computational cost. As a result, the latency-dependent tracking error is drastically reduced. To provide output at control loop frequency, Kalman filter-based upsampling of the motion measurements is used.

The epipolar constraint-based, linear parametrization is applied throughout the entire framework of tracking, mesh refinement, and motion upsampling. The presented method fully exploits stereoscopic constraints to estimate soft tissue motion including changes in depth as a prerequisite to online laser focusing and spot positioning.

### 6.1.1 Monoscopic Non-Rigid Motion Tracking

For a better understanding of the stereoscopic motion model presented in this thesis, this section provides a brief introduction to single view approach [PLF08, ZLH09]. Tracking in the mono view is based on a triangular mesh as illustrated in Figure 6.2a. The area of interest is approximated by a triangular mesh of $N$ vertices $\boldsymbol{s}_j = (u_j, v_j)^{\mathrm{T}}$ concatenated to the vector

$$\boldsymbol{S} = (u_1, \dots, u_N, v_1, \dots, v_N)^{\mathrm{T}} \in \mathbb{R}^{2N}. \tag{6.1}$$

An abritrary feature point $\boldsymbol{p}_m$ inside this region can be described by its constant barycentric coordinates $(\xi_j, \xi_k, \xi_l)^{\mathrm{T}}$ of its adjacent vertices $(\boldsymbol{s}_j, \boldsymbol{s}_k, \boldsymbol{s}_l)$ with the piecewise affine warp function

$$\boldsymbol{W}(\boldsymbol{p}_m, \boldsymbol{S}) = \begin{pmatrix} u_j & u_k & u_l \\ v_j & v_k & v_l \end{pmatrix} \begin{pmatrix} \xi_j \\ \xi_k \\ \xi_l \end{pmatrix}. \tag{6.2}$$

A correspondence set $\{\boldsymbol{p}_m, \boldsymbol{c}_m\}$ between initial feature point $\boldsymbol{p}_m$ and its position $\boldsymbol{c}_m$ in subsequent frames can be established by common feature matching techniques. Local motion between consec-

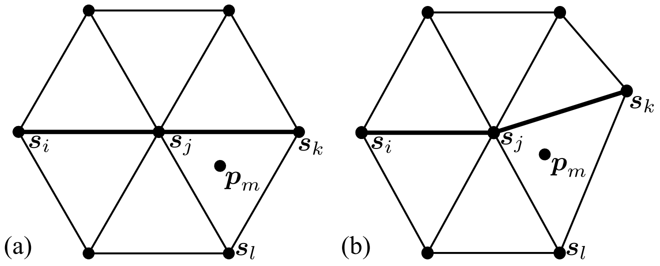


Figure 6.2: Hexagonal element of the mesh in (a) undeformed state and (b) deformed, penalized state.

utive frames is tracked with the pyramidal Lucas-Kanade (LK) method [Bou00]. With the features and the triangular mesh model, the non-rigid tracking problem can be formulated by energy term

$$\varepsilon(\boldsymbol{S}) = \varepsilon_{\mathrm{C}}(\boldsymbol{S}) + \lambda_{\mathrm{D}}\varepsilon_{\mathrm{D}}(\boldsymbol{S}), \tag{6.3}$$

where $\varepsilon_{\mathrm{C}}$ defining the mesh correspondence energy and $\varepsilon_{\mathrm{D}}$ the mesh deformation energy weighted by $\lambda_{\mathrm{D}}$.

The correspondence energy

$$\varepsilon_{\mathrm{C}}(\boldsymbol{S}) = \sum_{\boldsymbol{p}_m \in M} \rho\left(\delta\left(\boldsymbol{c}_m, \boldsymbol{p}_m, \boldsymbol{S}\right), r\right) \tag{6.4}$$

aggregates the weighted Euclidean distances

$$\delta\left(\boldsymbol{c}_m, \boldsymbol{p}_m, \boldsymbol{S}\right) = \|\boldsymbol{c}_m - \boldsymbol{W}(\boldsymbol{p}_m, \boldsymbol{S})\|_2 \tag{6.5}$$

between the features $\{\boldsymbol{c}_m, \boldsymbol{p}_m\}$ in an inlier set $M = \{\boldsymbol{p}_m \mid \delta^2 \leq r^2\}$. The robust estimator $\rho\left(\delta, r\right)$ penalizes less reliable measurements caused by occlusions or mismatches and is chosen according to [ZLH09]. When minimizing Equation 6.3, confidence radius $r$ is initiated with a sufficiently large value, e.g. $500$ pixels, before being decreased by a factor $\eta$ at constant rate. As a result, $\rho$ becomes more selective and detects outliers in the correspondence set having a distance greater than $r$. The correspondence energy $\varepsilon_{\mathrm{C}}(\boldsymbol{S})$ in Equation 6.3 can be formulated in matrix form

$$\varepsilon_{\mathrm{C}}(\boldsymbol{S}) = \boldsymbol{S}^{\mathrm{T}} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A} \end{pmatrix} \boldsymbol{S} - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{S} + c \tag{6.6}$$

with matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, vector $\boldsymbol{b} \in \mathbb{R}^{2N}$ computed from the barycentric coordinates of the feature points, and a constant $c$. For further details, the reader is kindly referred to [ZLH09].

The deformation energy $\varepsilon_{\mathrm{D}}$ regularizes the deformation and is approximated by the sum over the squared second-order derivatives of the mesh vertex coordinates in $u$ and $v$ direction. In detail, every collinear connected triplet $(\boldsymbol{s}_i,\boldsymbol{s}_j,\boldsymbol{s}_k)$ of a hexagon element that is centered at vertex $\boldsymbol{s}_j$ (see Figure 6.2b) contribute to the regularization term

$$\varepsilon_{\mathrm{D}}(\boldsymbol{S}) = \frac{1}{2} \sum_{\boldsymbol{s}_j \in \boldsymbol{S}} k_u^2(\boldsymbol{s}_j) + k_v^2(\boldsymbol{s}_j), \tag{6.7}$$

approximating the squared directional curvature of the surface. The terms $k_u(\boldsymbol{s}_j) = u_i - 2u_j + u_k$ and $k_v(\boldsymbol{s}_j) = v_i - 2v_j + v_k$ are the finite second order differences. A deformation of the mesh induces length differences between two collinear connected edges and thus, an increase in $\varepsilon_{\mathrm{D}}$. Regularization constraint in Equation 6.7 can be formulated as

$$\varepsilon_{\mathrm{D}}(\boldsymbol{S}) = \boldsymbol{S}^{\mathrm{T}} \begin{pmatrix} \boldsymbol{K} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K} \end{pmatrix} \boldsymbol{S} \tag{6.8}$$

with sparse matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ determined by the coefficients of the triplet configurations. For further details, please refer to [PLF08]. Finally, combining Equation 6.6 and 6.8, the total energy function in Equation 6.3 can be reformulated into the unconstrained quadratic optimization problem

$$\varepsilon(\boldsymbol{S}) = \boldsymbol{S}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{S} - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{S} + c\,, \tag{6.9}$$

where matrix $\boldsymbol{U} \in \mathbb{R}^{2N \times 2N}$ is defined as

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} + \lambda_{\mathrm{D}}\boldsymbol{K} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A} + \lambda_{\mathrm{D}}\boldsymbol{K} \end{pmatrix}. \tag{6.10}$$

The progressive finite Newton (PFN) method is applied to align the mesh to the tracked feature set [ZLH09]. During optimization, a fixed number of steps is used to decrease the confidence radius $r$ down to one pixel. Thus, outliers are rejected in a coarse-to-fine manner taking the penalty function in Equation 6.4 into account. Since each step requires only one Newton iteration, the PFN method provides a deterministic solution to the minimization problem. Each step within the PFN is defined by

$$\boldsymbol{S} \leftarrow \boldsymbol{S} - \Delta\boldsymbol{S} = \boldsymbol{S} - \gamma\, \boldsymbol{H}^{-1}(\boldsymbol{S})\, \nabla\varepsilon(\boldsymbol{S}), \tag{6.11}$$

with step-size $\gamma = 1$, gradient

$$\nabla\varepsilon(\boldsymbol{S}) = 2\,(\boldsymbol{U}\boldsymbol{S} - \boldsymbol{b})\,, \tag{6.12}$$

and Hessian matrix

$$\boldsymbol{H}(\boldsymbol{S}) = 2\boldsymbol{U}. \tag{6.13}$$

Algorithm 6.1 summarizes the presented method. Regarding further speed-up of the minimization process, mesh update in Equation 6.11 can be split into two independent matrix operations each solved by LU decomposition [ZLH09].

**Algorithm 6.1:** Non-rigid tracking in the mono view [ZLH09].

```
precompute:
(1) Initialize $\boldsymbol{S}$ and correspondences $\{\boldsymbol{p}_m,\boldsymbol{c}_m\}$
(2) Precompute $\boldsymbol{K}$ and $(\xi_j,\xi_k,\xi_l)\,\forall \boldsymbol{p}_m$
for each image do
    input: Current mesh $\boldsymbol{S}$ from previous image
    (3) Initialize PFN confidence radius $r \leftarrow r_{\mathrm{start}}$
    repeat
        (3) Reject outliers for confidence region $r$
        (4) $\boldsymbol{S} \leftarrow \boldsymbol{S} - \Delta\boldsymbol{S}$ acc. Equation 6.11
        (5) $r \leftarrow \eta r$ with $0 < \eta < 1$
    until $r \leq r_{\mathrm{end}}$
    output: Updated mesh $\boldsymbol{S}$
end
```

## 6.1.2 Robust Stereoscopic Motion Tracking

In this section, the single view mesh model outlined above is extended to stereo vision. This is achieved by incorporating the epipolar constraint that enforces left–right consistency.

### Extension to Stereo Vision

Endoscopic stereo imaging facilitates metric surface measurements by triangulation of image points from the left and right camera view. Common methods for stereo-based motion estimation often consider projective camera geometry [RPL10, WYLP13, YWLP14]. Consequently, the computation of the Jacobian and the Hessian matrix is complex due to the non-linearity of the projective functions. Instead of considering projective geometry, a computationally more efficient solution is found by formulating the problem in disparity space. When the epipolar constraint for a rectified stereo configuration with coplanar image planes (see Figure 6.3) is applied, a linear parametrization can be defined as follows

$$\boldsymbol{q} = (u_1, \dots ,u_N,v_1, \dots ,v_N,d_1, \dots ,d_N)^{\mathrm{T}} \in \mathbb{R}^{3N}, \tag{6.14}$$

where the stacked vertex coordinates and the associated disparities $\boldsymbol{d} = (d_1, \dots ,d_N)^{\mathrm{T}}$ describe the horizontal pixel shift between correspondences in both views. Reference is given with respect to the left camera frame $(\mathrm{CF})_{\mathrm{L}}$ such that $(u_{\mathrm{L},j},v_{\mathrm{L},j})^{\mathrm{T}} = (u_j,v_j)^{\mathrm{T}}$. For a rectified stereo view, as shown in Figure 6.3, corresponding points are enforced to have the same vertical coordinate $v$.

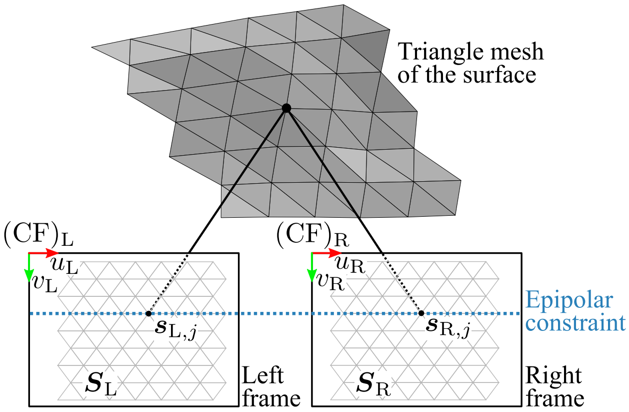


Figure 6.3: Rectified stereo view illustrating the epipolar constraint for left-right consistency of the mesh.

Based on the parametrization $\boldsymbol{q}$, the mesh coordinates in the left and right view, denoted by $\boldsymbol{S}_\mathrm{L}$ and $\boldsymbol{S}_\mathrm{R}$, respectively, are defined by

$$\boldsymbol{S}_i = \boldsymbol{S}_i(\boldsymbol{q}) \quad \text{with } i \in \{\mathrm{L,R}\} \tag{6.15}$$

with $j$-th mesh point

$$\boldsymbol{s}_{i,j}(\boldsymbol{q}) = \begin{cases} (u_j,v_j)^\mathrm{T} & \text{if } i = \mathrm{L} \\ (u_j + d_j,v_j)^\mathrm{T} & \text{if } i = \mathrm{R} \end{cases}. \tag{6.16}$$

Consequently, the piecewise affine warping of the point $\boldsymbol{p}_m$ in the left or right view is given by

$$\boldsymbol{W}_i(\boldsymbol{p}_m,\boldsymbol{q}) = \boldsymbol{W}(\boldsymbol{p}_m,\boldsymbol{S}_i(\boldsymbol{q})) = \begin{pmatrix} \boldsymbol{\xi}_m^\mathrm{T} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^\mathrm{T} \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}), \tag{6.17}$$

where $\boldsymbol{0}_N \in \mathbb{R}^{1\times N}$ is the zero vector and $\boldsymbol{\xi}_m^\mathrm{T} \in \mathbb{R}^{1\times N}$ the vector containing the non-zero barycentric coordinates $(\xi_j,\xi_k,\xi_l)^\mathrm{T}$ with respect to the adjacent mesh vertices of the point $\boldsymbol{p}_m$. The remaining elements in $\boldsymbol{\xi}_m$ are set to zero.

In comparison with the monoscopic approach [ZLH09], stereo-based motion estimation aims at minimizing the mesh alignment error in both the left and right views. Initially, features are matched independently between consecutive frames. Then, left-right consistency is achieved by minimizing the function

$$\varepsilon(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L,R}\}} \varepsilon_\mathrm{C}(\boldsymbol{S}_i(\boldsymbol{q})) + \lambda_\mathrm{D} \sum_{i\in\{\mathrm{L,R}\}} \varepsilon_\mathrm{D}(\boldsymbol{S}_i(\boldsymbol{q})) \tag{6.18}$$

combining the correspondence energy $\varepsilon_{\mathrm{C},i} = \varepsilon_\mathrm{C}(\boldsymbol{S}_i(\boldsymbol{q}))$

$$\varepsilon_{\mathrm{C},i} = \boldsymbol{S}_i(\boldsymbol{q})^\mathrm{T} \begin{pmatrix} \boldsymbol{A}_i & 0 \\ 0 & \boldsymbol{A}_i \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}) - 2\boldsymbol{b}_i^\mathrm{T}\, \boldsymbol{S}_i(\boldsymbol{q}) + c_i \tag{6.19}$$

and the deformation energy $\varepsilon_{\mathrm{D},i} = \varepsilon_{\mathrm{D}}(\boldsymbol{S}_i(\boldsymbol{q}))$

$$\varepsilon_{\mathrm{D},i} = \boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}} \begin{pmatrix} \boldsymbol{K}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_i \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}) \tag{6.20}$$

of the two views, respectively. The total energy in Equation 6.18 can be summarized as

$$\varepsilon(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \left(\boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}} \boldsymbol{U}_i\, \boldsymbol{S}_i(\boldsymbol{q}) - 2\boldsymbol{b}_i^{\mathrm{T}}\, \boldsymbol{S}_i(\boldsymbol{q}) + c_i\right). \tag{6.21}$$

Since Equation 6.21 remains an unconstrained quadratic optimization problem similar to Equation 6.9, the PFN method in Equation 6.11 is adopted to minimize the energy term with respect to $\boldsymbol{q}$. The overall gradient $\nabla\varepsilon(\boldsymbol{q}) = \nabla\varepsilon_{\mathrm{L}}(\boldsymbol{q}) + \nabla\varepsilon_{\mathrm{R}}(\boldsymbol{q})$ requires the two gradients from the left and right view defined by

$$\nabla\varepsilon_i(\boldsymbol{q}) = \begin{pmatrix} \frac{\partial\varepsilon_i}{\partial u_1} \\ \vdots \\ \frac{\partial\varepsilon_i}{\partial d_n} \end{pmatrix} = 2\,\frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}}^{\mathrm{T}} \left(\boldsymbol{U}_i \boldsymbol{S}_i(\boldsymbol{q}) - \boldsymbol{b}_i\right) \tag{6.22}$$

with Jacobian matrix

$$\frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}} = \begin{pmatrix} \boldsymbol{I}_{N\times N} & \boldsymbol{0}_{N\times N} & *_i \\ \boldsymbol{0}_{N\times N} & \boldsymbol{I}_{N\times N} & \boldsymbol{0}_{N\times N} \end{pmatrix} \in \mathbb{R}^{2N\times 3N}. \tag{6.23}$$

Matrix $*_i$ differs for the two views such that

$$*_i = \begin{cases} \boldsymbol{0}_{N\times N} & \text{if } i = \mathrm{L} \\ \boldsymbol{I}_{N\times N} & \text{if } i = \mathrm{R} \end{cases} \in \mathbb{R}^{N\times N} \tag{6.24}$$

defines either zero matrix $\boldsymbol{0}_{N\times N}$ or identity matrix $\boldsymbol{I}_{N\times N}$. Due to linearity of $\boldsymbol{S}_i(\boldsymbol{q})$, the two Jacobian matrices are constant and can be precomputed for efficiency. The overall Hessian matrix $\boldsymbol{H}(\boldsymbol{q}) = \boldsymbol{H}_{\mathrm{L}}(\boldsymbol{q}) + \boldsymbol{H}_{\mathrm{R}}(\boldsymbol{q})$ requires the two Hessians from the left and right view defined by

$$\boldsymbol{H}_i(\boldsymbol{q}) = 2\,\frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}}^{\mathrm{T}} \boldsymbol{U}_i\, \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}}. \tag{6.25}$$

Similar to the monoscopic approach, the PFN method can be used to minimize the energy term in Equation 6.21 by computing each Newton step

$$\boldsymbol{q} \leftarrow \boldsymbol{q} - \Delta\boldsymbol{q} = \boldsymbol{q} - \gamma\, \boldsymbol{H}^{-1}(\boldsymbol{q})\, \nabla\varepsilon(\boldsymbol{q}) \tag{6.26}$$

with step size $\gamma = 1$. Algorithm 6.2 summarizes the non-rigid tracking extended to stereo vision. In addition to each triangle center, salient gradient-based landmarks (commonly five to seven points) are selected and tracked as mesh support points [Shi94].

**Algorithm 6.2:** Stereoscopic non-rigid tracking.

```
precompute:
(1) Initialize $\boldsymbol{S}_{\{\mathrm{L,R}\}}$
(2) Concatenate $\boldsymbol{S}_{\mathrm{L}}$ and $\boldsymbol{S}_{\mathrm{R}}$ to form parameter vector $\boldsymbol{q}$
for each stereo image pair do
    input: Parameter vector $\boldsymbol{q}$ from previous time step
    (3) Initialize PFN confidence radius $r \leftarrow r_{\mathrm{start}}$
    repeat
        (4) Reject outliers for confidence region $r$
        (5) Compute gradient $\nabla\varepsilon(\boldsymbol{q})$
        (6) Compute Hessian $\boldsymbol{H}(\boldsymbol{q})$
        (7) $\boldsymbol{q} \leftarrow \boldsymbol{q} - \Delta\boldsymbol{q}$ according to Equation 6.26
        (8) Update $\boldsymbol{S}_{\{\mathrm{L,R}\}}(\boldsymbol{q})$
        (9) $r \leftarrow \eta r$ with $0 < \eta < 1$
    until $r \leq r_{\mathrm{end}}$
    output: Updated stereo mesh $\boldsymbol{S}_{\{\mathrm{L,R}\}}(\boldsymbol{q})$
end
```

### Illumination-Invariance and Occlusion Detection

To compensate for nonlinear illumination changes during tracking, the rank transform is applied to the endoscopic images, encoding the relative ordering of the color intensities $\boldsymbol{I}(\boldsymbol{p_c}) \in \mathbb{R}^{3\times1}$ in a local proximity $\Omega_{\mathrm{R}}$ of center pixel $\boldsymbol{p}_c$ [ZW94]. The rank transform $\boldsymbol{I}_{\mathrm{R}}$ of image $\boldsymbol{I}$ is defined by

$$\boldsymbol{I}_{\mathrm{R}}(\boldsymbol{p}_c) = \sum_{\boldsymbol{p}_j \in \Omega_{\mathrm{R}}} \zeta\left(\bar{\boldsymbol{I}}(\boldsymbol{p}_c), \boldsymbol{I}(\boldsymbol{p}_j)\right), \tag{6.27}$$

where $\Omega_{\mathrm{R}}$ is a $13 \times 13$ pixel window around $\boldsymbol{p}_c$ and $\zeta$ is the function

$$\zeta(I_1, I_2) = \begin{cases} 0, & \text{if } I_1 \leq I_2 \\ 1, & \text{else} \end{cases} \tag{6.28}$$

evaluating the sign of the pixelwise comparison. In contrast to the original work [ZW94], the intensity mean $\bar{\boldsymbol{I}}(\boldsymbol{p}_c)$ from a $3 \times 3$ neighborhood is computed to compensate for image noise (see Figure 6.4a).

The progressive outlier rejection scheme, as outlined in Section 6.1.1, is able to detect inconsistent feature correspondences. Additionally, the feature similarity of the Lucas-Kanade method is

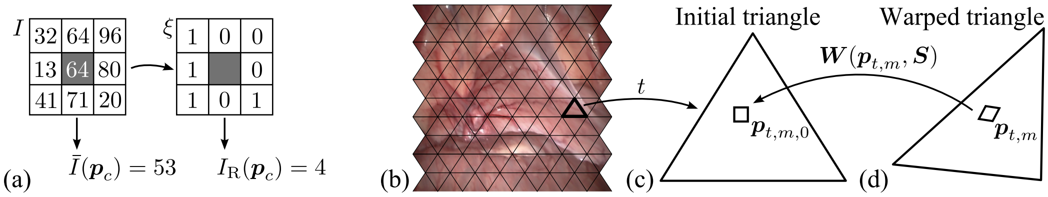


Figure 6.4: Illumination-invariant tracking is achieved by applying the rank transform as exemplarily shown in (a) for a 3 × 3 gray-scale image section. Robustness to occlusions of (b) the triangular mesh model is improved by computing the Mahalanobis distance between (c) the triangle $t$ in the reference frame and (d) its warped configuration as a result of tissue deformation.

analyzed to discard mismatches. However, the two rejection schemes alone do not ensure temporal consistency. In particular occlusions might be detected once but then mistakenly tracked as false positives in subsequent frames even though the occlusion still exists. This may leads to tracking failure. To overcome this limitation, color similarity is analyzed to reject outliers beforehand and thus to relax the regularization term in Equation 6.18. A shortcoming of common color-based methods is the detection of false positives as a result of illumination changes. Illumination invariant image representation by deployed rank transform facilitates discriminative matching but does not respond well to changes in color. To increase sensitivity to appearance changes, the idea of identifying local multivariate outliers is adopted [FRGTA13]. In detail, a cross-channel, pairwise Mahalanobis distance is implemented considering the spatial context of the image texture. For each pixel $\boldsymbol{p}_{t,m,0}$ with $m \in \{1, \ldots, M\}$ in mesh triangle $t$ the distance

$$\Delta\boldsymbol{I}_\mathrm{R}(\boldsymbol{p}_{t,m}) = \boldsymbol{I}_\mathrm{R}(\boldsymbol{W}(\boldsymbol{p}_{t,m}, \boldsymbol{S})) - \boldsymbol{I}_\mathrm{R}(\boldsymbol{p}_{t,m,0}) \tag{6.29}$$

is computed with respect to tracked position $\boldsymbol{p}_{t,m}$ mapped back onto the reference frame with $\boldsymbol{p}_{t,m,0}$ (see Figure 6.4b–d). In this context, vector $\boldsymbol{I}_\mathrm{R} \in \mathbb{R}^{3\times1}$ represents the stacked, rank-encoded color channels. The pairwise Mahalanobis distance is obtained by

$$d^2_\mathrm{MHD}(\boldsymbol{p}_{t,m}) = \Delta\boldsymbol{I}^\mathrm{T}_\mathrm{R}(\boldsymbol{p}_{t,m})\, \Sigma_t^{-1}\, \Delta\boldsymbol{I}_\mathrm{R}(\boldsymbol{p}_{t,m}) \tag{6.30}$$

with covariance matrix $\Sigma_t \in \mathbb{R}^{3\times3}$ of the initial triangle $t$. If the L1-norm distance of the stacked squared Mahalanobis distances $d_\mathrm{MHD}(\boldsymbol{p}_{t,m})$ exceeds a certain threshold $\boldsymbol{d}_\mathrm{MHD}(t) > \beta$, triangle $t$ is considered to be occluded and thus rejected as outlier. Compared to the global Mahalanobis distance, presented measure does not rely solely on the observation's mean and covariance but also takes spatial dependence in the presence of piecewise affine deformation into account.

### 6.1.3 Epipolar Constraint-Based Mesh Refinement

In this section, mesh refinement (MR) taking texture information into account is described. In particular, the epipolar constraint outlined above is incorporated into the deformable Lucas-Kanade framework (DLK) [ZLH09]. As in Equation 6.18, the refinement step considers a regularization term $\varepsilon_{\mathrm{D}}$. The total energy for the stereo view is defined as follows

$$\varepsilon_{\mathrm{MR}}(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L,R}\}} \varepsilon_{\mathrm{A}}(\boldsymbol{S}_i(\boldsymbol{q})) + \lambda_{\mathrm{D}} \sum_{i\in\{\mathrm{L,R}\}} \varepsilon_{\mathrm{D}}(\boldsymbol{S}_i(\boldsymbol{q}))\,. \tag{6.31}$$

The inverse compositional framework is adopted to minimize the residual between the current image $I_i$ and the warped template $T_i$ by the data term $\varepsilon_{\mathrm{A},i} = \varepsilon_{\mathrm{A}}(\boldsymbol{S}_i(\boldsymbol{q}))$

$$\varepsilon_{\mathrm{A},i} = \sum_{\boldsymbol{p}_m\in\Omega_{\mathrm{S}}} \rho\bigg(\Big[T_i(\boldsymbol{W}_i(\boldsymbol{p}_m,\Delta\boldsymbol{q})) - I_i(\boldsymbol{W}_i(\boldsymbol{p}_m,\boldsymbol{q}))\Big]^2\bigg), \tag{6.32}$$

where the warping function in Equation 6.17 of the pixel $\boldsymbol{p}_m$ with $m \in \{1,\ldots,M\}$ is used in the image region $\Omega_{\mathrm{S}}$ represented by the mesh [BM04]. In order to provide increased robustness against tracking outliers, Equation 6.32 is formulated in an iteratively re-weighted least squares framework based on the function

$$\rho(u) = \begin{cases} u & \text{if } u \le \sigma_H^2 \\ (2\sqrt{u} - \sigma_H)\sigma_H & \text{if } u > \sigma_H^2, \end{cases} \tag{6.33}$$

that is the norm-like Huber loss [Hub64]. For small residuals smaller than a threshold $\sigma_H^2$, $\rho(u)$ behaves as the standard unweighted least squares estimator. However, challenging conditions, such as occlusions or specular highlights on glossy tissue, require a robust cost function, as that deployed in Equation 6.33, to reduce the weight of outliers. In this case, the Huber function switches to linear behavior for large residuals. This has been proven to perform well in image-based structure and motion estimation [HB98, ZGH09, CSDE13]. The stereo approach will take adaptive cost re-weighting into account in order to increase tracking robustness for laser ablation on soft tissue.

Assuming an initial value for $\boldsymbol{q}$, the energy term in Equation 6.31 is solved iteratively considering an increment $\Delta\boldsymbol{q}$. As a result, the following expression of the mesh coordinates

$$\boldsymbol{S}_i(\boldsymbol{q}) \rightarrow \boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q}) = \boldsymbol{S}_i(\boldsymbol{q}) + \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{q}}\Delta\boldsymbol{q} \tag{6.34}$$

is obtained. This allows for reformulating the deformation energy in Equation 6.20, yielding

$$\varepsilon_{\mathrm{D},i} \approx (\boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q}))^{\mathrm{T}}\, \boldsymbol{\mathcal{K}}_i\, (\boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q})), \tag{6.35}$$

where

$$\boldsymbol{\mathcal{K}}_i = \begin{pmatrix} \boldsymbol{K}_i & 0 \\ 0 & \boldsymbol{K}_i \end{pmatrix}. \tag{6.36}$$

Analogously to Equation 6.35, the appearance-based energy in Equation 6.32 can be linearized to

$$\varepsilon_{\mathrm{A},i} \approx \sum_{\boldsymbol{p}_m \in \Omega} \rho\Big(\Big[\,\Delta I_{i,m} \,+\, \boldsymbol{J}_{i,m}\,\Delta\boldsymbol{q}\Big]^2\Big), \tag{6.37}$$

with residual

$$\Delta I_{i,m} = T_i(\boldsymbol{W}_i(\boldsymbol{p}_m,\boldsymbol{q}_0)) \,-\, I_i(\boldsymbol{W}_i(\boldsymbol{p}_m,\boldsymbol{q})) \tag{6.38}$$

describing the photometric error. The identity warp $\boldsymbol{W}_i(\boldsymbol{p}_m,\boldsymbol{q}_0)$ is evaluated at the initial parameter set $\boldsymbol{q}_0$. The Jacobian $\boldsymbol{J}_{i,m}$ at the point $\boldsymbol{p}_m$ is defined by steepest descent image

$$\boldsymbol{J}_{i,m} = \left.\frac{\partial T_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} = \nabla T_i \left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} \in \mathbb{R}^{1\times 3N}, \tag{6.39}$$

where $\nabla T_i = \big(\frac{\partial T_i}{\partial u}, \frac{\partial T_i}{\partial v}\big)$ denotes the image gradient at $\boldsymbol{p}_m$. The Jacobian of the warp is obtained by the product of the derivative of Equation 6.17

$$\left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{S}_i}\right|_{\boldsymbol{p}_m} = \begin{pmatrix} \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{2\times 2N} \tag{6.40}$$

and the derivative of the mesh coordinates in Equation 6.23 yielding

$$\left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} = \frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{S}_i}\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{q}} = \begin{pmatrix} \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N & \star_i \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N \end{pmatrix} \in \mathbb{R}^{2\times 3N}, \tag{6.41}$$

where the placeholder

$$\star_i = \begin{cases} \boldsymbol{0}_N & \text{if } i = \mathrm{L} \\ \boldsymbol{\xi}_m^{\mathrm{T}} & \text{if } i = \mathrm{R} \end{cases} \in \mathbb{R}^{1\times N} \tag{6.42}$$

is either the zero vector $\boldsymbol{0}_N$ or the barycentric coordinates $\boldsymbol{\xi}_m^{\mathrm{T}}$ of the point $\boldsymbol{p}_m$, as described in Equation 6.17. Due to the inverse compositional algorithm and the linearity of $\boldsymbol{S}_i(\boldsymbol{q})$, the Jacobian matrix in Equation 6.39 is constant and can be computed offline.

As the gradient of the linearized energy function vanishes for optimality, the parameter update $\Delta\boldsymbol{q}$ is attained by Gauss-Newton optimization as follows

$$\Delta\boldsymbol{q} = -\boldsymbol{H}_{\mathrm{MR}}^{-1} \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \Big( \boldsymbol{J}_{\mathrm{S},i}^{\mathrm{T}} \boldsymbol{D}_i\, \Delta\boldsymbol{I}_{\mathrm{S},i} \,+\, \lambda_{\mathrm{D}}\, \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}}}{\partial \boldsymbol{q}}\, \boldsymbol{\mathcal{K}}_i\, \boldsymbol{S}_i(\boldsymbol{q}) \Big) \tag{6.43}$$

with stereo-based Hessian matrix

$$\boldsymbol{H}_{\mathrm{MR}} = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \Big( \boldsymbol{J}_{\mathrm{S},i}^{\mathrm{T}} \boldsymbol{D}_i \boldsymbol{J}_{\mathrm{S},i} \,+\, \lambda_{\mathrm{D}}\, \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}}}{\partial \boldsymbol{q}}\, \boldsymbol{\mathcal{K}}_i\, \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}} \Big) \tag{6.44}$$

and diagonal matrix of the weights

$$\boldsymbol{D}_i = \operatorname{diag}\Big(\rho'\big(\Delta I_{i,1}^2\big), \dots, \rho'\big(\Delta I_{i,M}^2\big)\Big) \ . \tag{6.45}$$

In Equation 6.43 and 6.44, the pointwise residuals and the Jacobians form the matrices

$$\Delta \boldsymbol{I}_{\mathrm{S},i} = \left(\Delta I_{i,1}, \dots, \Delta I_{i,M}\right)^{\mathrm{T}} \tag{6.46}$$

and

$$\boldsymbol{J}_{\mathrm{S},i} = \big(\boldsymbol{J}_{i,1}^{\mathrm{T}}, \dots, \boldsymbol{J}_{i,M}^{\mathrm{T}}\big)^{\mathrm{T}} \ , \tag{6.47}$$

respectively. As outlined in the previous section, Mahalanobis distance-based (MHD) outlier detection is deployed taking the spatial distribution of texture information into account. Specifically, image pixels are classified as occluded if the MHD between the initial template and the current image exceed a predefined threshold $\beta$. In case of occlusion, the indicator function

$$\delta_{\mathrm{MHD}} = \begin{cases} 1 & \text{if } \boldsymbol{d}_{\mathrm{MHD}}(t) > \beta \\ 0 & \text{otherwise} \end{cases} \tag{6.48}$$

causes the modified weight

$$\rho'(u) = \big(1 - \delta_{\mathrm{MHD}}\big) \frac{\partial \rho(u)}{\partial u} \tag{6.49}$$

to be set to zero. Consequently, the associated pixel does not contribute to the refinement.

To reduce the computational load of the iteratively re-weighted least squares method, the H-algorithm is implemented [DH81]. Due to the inverse compositional approach, the unweighted Hessian matrix can then be computed offline. To avoid slow convergence or even divergence, the Huber weights are normalized to compensate for influences on the step size of the iterative optimization [BGMI03]. In addition, parameter estimation is implemented in a pyramidal coarse-to-fine scheme, reducing the computational load for robustly tracking large displacements caused by rapid motion. Algorithm 6.3 summarizes the stereo-based refinement.

**Algorithm 6.3:** Stereoscopic mesh refinement.

```
pre-compute:
(1) Initialize gradient $\nabla T_{\{\mathrm{L,R}\}}$ of template image $T_{\{\mathrm{L,R}\}}$
(2) Initialize Jacobian $\boldsymbol{J}_{\mathrm{S},\{\mathrm{L,R}\}}$ according to Equation 6.47
(3) Initialize unweighted Hessian $\boldsymbol{H}_{\mathrm{MR}}$ according to Equation 6.44
for each stereo image pair do
    input: Parameter vector $\boldsymbol{q}$ from tracking (Algorithm 6.2)
    repeat
        (4) Warp image $I_{\{\mathrm{L,R}\}}$ according to Equation 6.17
        (5) Compute residuals $\Delta\boldsymbol{I}_{\mathrm{S},\{\mathrm{L,R}\}}$ according to Equation 6.38 and 6.46
        (6) Compute weights matrix $\boldsymbol{D}_{\{\mathrm{L,R}\}}$ according to Equation 6.45 and 6.49
        (7) $\boldsymbol{q} \leftarrow \boldsymbol{q} + \Delta\boldsymbol{q}$ according to Equation 6.43
        (8) Update $\boldsymbol{S}_{\{\mathrm{L,R}\}}(\boldsymbol{q})$
    until $\|\Delta q\| \leq \epsilon$
    output: Refined stereo mesh $\boldsymbol{S}_{\{\mathrm{L,R}\}}(\boldsymbol{q})$
end
```

### 6.1.4 Considerations for Real-Time Implementation

Most of the model-based, non-rigid tracking methods cannot operate at image-acquisition rates of 30 Hz and higher. Thus, tracking might fail if the tissue undergoes rapid motion or significant deformation. This section addresses a technique for online application of the tracking method.

#### Concurrent Tracking and Mesh Refinement

Stereoscopic tracking without mesh refinement (noMR) according to Algorithm 6.2 provides a high update rate; however, it may lead to drift over time, since appearance is not considered (see Figure 6.5a). To compensate for drift, the refinement step according to Algorithm 6.3 can be invoked, once Algorithm 6.2 has finished. This is called sequential mesh refinement (sMR) (see Figure 6.5b).

When online assistance for laser surgery is intended, the processing rate of the image pipeline, including image undistortion, tracking, and mesh refinement as well as the surgical tool control loop (e.g. of the ablation laser), should be at least in the order of the image acquisition rate. To accelerate the mesh refinement, heterogeneous programming on general purpose graphics processing units (GPGPU) is deployed. However, it provides only a limited solution due to the nonlinear convergence rate. If mesh misalignment is large, subsequent camera images cannot be processed on time when sMR is used; thus, they need to be discarded until the mesh refinement of the prior frame has converged (see Figure 6.5b). This may lead not only to significantly delayed measurements but also to tracking failure if the scene concurrently undergoes rapid motion or

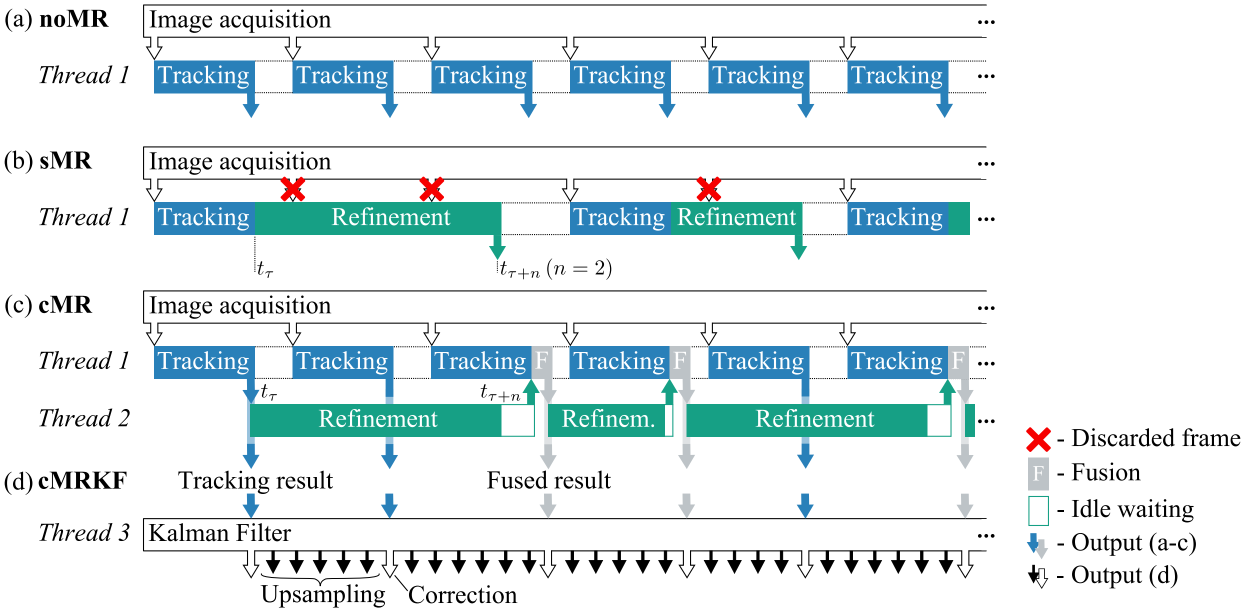


Figure 6.5: Computational pipeline considering (a) tracking by Algorithm 6.2 without mesh refinement (noMR), (b) sequential (sMR), and (c) concurrent mesh refinement (cMR) both deploying Algorithms 6.2 and 6.3. In (c) tracking runs in Thread 1 (CPU) and the refinement in Thread 2 (GPU). Subsequently, proposed fusion method (F) is called by Thread 1. For online laser control, a Kalman filter for motion upsampling (cMRKF) is running in Thread 3 (CPU) as shown in (d).

large local deformation. Thus, concurrent tracking and mesh refinement (cMR) with subsequent affine-invariant fusion, as illustrated in Figure 6.5c, is discussed (see next section). This method is compared with noMR and sMR. The latter method corresponds to the monoscopic DLK-algorithm [ZLH09, DCA$^+$15] that, in this study, has been extended to stereo vision by incorporating the epipolar constraint. Additionally, upsampling of the motion measurements using Kalman filtering (cMRKF) is considered, as shown in Figure 6.5d.

### Affine-Invariant Fusion of Tracking and Mesh Refinement

Regarding sMR, motion is initially estimated according to Algorithm 6.2. Subsequently, Algorithm 6.3 is initiated at time $t_\tau$ to compensate for mesh misalignment caused, for instance, by drift (see Figure 6.5b). Since the refinement processes dense texture information in a gradient-based optimization scheme, convergence cannot be ensured until acquisition of the next camera frame. Assuming the refinement result to be available at time $t_{\tau+n}$ with a delay of $n$ frames, concurrent computation of tracking and mesh refinement (cMR), as shown in Figure 6.5c, is required to achieve online capability for vision-guided interventions.

The fusion of both steps required for cMR is presented in the following. Let us exemplarily consider motion tracking from time $t_\tau$ to $t_{\tau+n}$ with mesh vertex $\boldsymbol{s}_{j,\tau}$ being subject to drift (see Figure 6.6a). Due to computational delay, fusion of the refinement result $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ referring to $t_\tau$

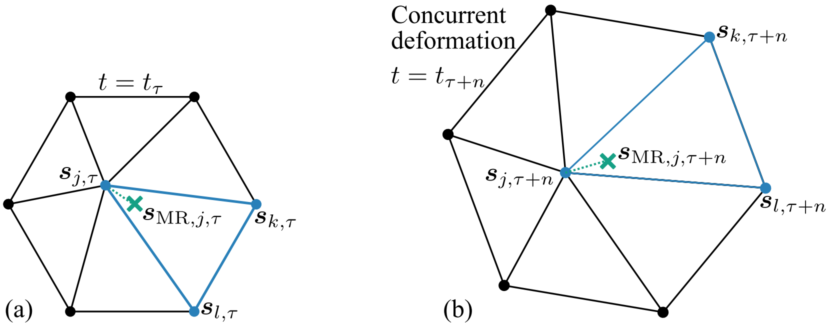

$t = t_\tau$
$\boldsymbol{s}_{j,\tau}$
$\boldsymbol{s}_{\mathrm{MR},j,\tau}$
$\boldsymbol{s}_{k,\tau}$
$\boldsymbol{s}_{l,\tau}$
(a)
Concurrent deformation
$t = t_{\tau+n}$
$\boldsymbol{s}_{k,\tau+n}$
$\boldsymbol{s}_{\mathrm{MR},j,\tau+n}$
$\boldsymbol{s}_{j,\tau+n}$
$\boldsymbol{s}_{l,\tau+n}$
(b)

Figure 6.6: Mesh in (a) initial configuration at time $t_\tau$ and (b) subsequently tracked position at time $t_{\tau+n}$. If drift occurs for vertex $\boldsymbol{s}_{j,\tau}$, the proposed mesh refinement yields the corrected position $\boldsymbol{s}_{\mathrm{MR},j,\tau}$. Fusion of delayed mesh refinement with respect to $t_\tau$ is achieved at subsequent time $t_{\tau+n}$ even if the mesh concurrently undergoes deformation from (a) to (b).

cannot be achieved until $t_{\tau+n}$. To compensate for simultaneous deformation (see Figure 6.6b), affine-invariant fusion of the corrected mesh vertex position $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ is performed by

$$\begin{aligned}\boldsymbol{s}_{\mathrm{MR},j,\tau+n} &= \boldsymbol{W}_i(\boldsymbol{s}_{\mathrm{MR},j,\tau},\boldsymbol{q}_{\tau+n}) \\ &= \begin{pmatrix} \boldsymbol{s}_{j,\tau+n} & \boldsymbol{s}_{k,\tau+n} & \boldsymbol{s}_{l,\tau+n} \end{pmatrix} \begin{pmatrix} \xi_{\mathrm{MR},j,\tau} \\ \xi_{\mathrm{MR},k,\tau} \\ \xi_{\mathrm{MR},l,\tau} \end{pmatrix} ,\end{aligned} \tag{6.50}$$

where the parametrization $\boldsymbol{q}_{\tau+n}$ is obtained from stereo-based tracking with Algorithm 6.2 running in Thread 1. After Thread 2 finishes Algorithm 6.3 at time $t_{\tau+n}$ yielding $\boldsymbol{s}_{\mathrm{MR},j,\tau}$, a triangle inlier test determines adjacent mesh vertices $(\boldsymbol{s}_j,\boldsymbol{s}_k,\boldsymbol{s}_l)^\mathrm{T}$. Specifically, $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ is considered to be inlier if its barycentric coordinates $(\xi_{\mathrm{MR},j,\tau},\xi_{\mathrm{MR},k,\tau},\xi_{\mathrm{MR},l,\tau})^\mathrm{T}$ satisfy the condition

$$0 \leq \xi_{\mathrm{MR},\{j,k,l\},\tau} \leq 1 \, . \tag{6.51}$$

If the refined vertex is located outside mesh boundaries, the triangle with the shortest distance to $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ is selected. Using barycentric coordinates in Equation 6.50, instead of Cartesian vertex positions, allows affine-invariant fusion of tracking and refinement at $t_{\tau+n}$, independent of concurrent deformation. The mesh coordinates in the two views can be corrected in one step by computing

$$\boldsymbol{S}_{\mathrm{MR},i,\tau+n} = \begin{pmatrix} \boldsymbol{\xi}^\mathrm{T}_{\mathrm{MR},1,\tau} & \boldsymbol{0}_N \\ \vdots & \vdots \\ \boldsymbol{\xi}^\mathrm{T}_{\mathrm{MR},N,\tau} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}^\mathrm{T}_{\mathrm{MR},1,\tau} \\ \vdots & \vdots \\ \boldsymbol{0}_N & \boldsymbol{\xi}^\mathrm{T}_{\mathrm{MR},N,\tau} \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}_{\tau+n}) \tag{6.52}$$

considering the pixelwise warp function in Equation 6.50 in a stacked formulation of Equation 6.17. Since the epipolar constraint is satisfied for both tracking and mesh refinements, it implicitly applies to Equation 6.52.

While Algorithm 6.2 runs on the CPU in Thread 1 (Core i7–3770, Intel Corporation, Santa Clara, CA, USA), Algorithm 6.3 is computed asynchronously in Thread 2 deploying the CUDA framework and a GeForce GTX Titan GPU (Nvidia Corporation, Santa Clara, CA, USA). Once Thread 2 finishes, fusion according to Equation 6.52 is performed. In contrast to Algorithm 6.2, inlier check in Equation 6.51 and the subsequent matrix multiplication in Equation 6.52 are computed at negligible costs.

### 6.1.5 Upsampling of the Motion Measurements

To further reduce the latency-dependent tracking misalignment, the epipolar constraint-based parameter set in Equation 6.14 is incorporated into filter-based motion upsampling. This is denoted by cMRKF. Adopting the idea of an iconic (pixelwise) representation of the Kalman filter for predicting changes in depth [VBG08], each vertex $\boldsymbol{s}_j$ is tracked individually with state vector

$$\boldsymbol{x}_{j,\tau} = \left(u_{j,\tau}, v_{j,\tau}, d_{j,\tau}, \dot{u}_{j,\tau}, \dot{v}_{j,\tau}, \dot{d}_{j,\tau}\right)^{\mathrm{T}}, \tag{6.53}$$

where the first subscript $j$ indicates the vertex index and the second subscript $\tau$ time step $t_\tau$. Instead of modeling the entire mesh within a single state space model, Equation 6.53 minimizes the size of the associated filter matrices; thus, it increases computational efficiency. The state vector is defined in disparity space, taking spatial and temporal information of the mesh vertex $\boldsymbol{s}_j$, i.e., its position $(u_j, v_j)^{\mathrm{T}}$ and motion vector $(\dot{u}_j, \dot{v}_j)^{\mathrm{T}}$, into account. Changes in depth are considered by the inversely related disparity $d_j$ and the associated disparity rate $\dot{d}_j$ between the left and right camera view. Kalman filtering enables estimating the current state, taking process and measurement noise into account. The three motion directions $\{u, v, d\}$ can be considered as independent; thus, the state representation in Equation 6.53 can be separated into the following three state vectors

$$\begin{aligned} \boldsymbol{x}_{\mathrm{u},j,\tau} &= \left(u_{j,\tau}, \dot{u}_{j,\tau}\right)^{\mathrm{T}} \\ \boldsymbol{x}_{\mathrm{v},j,\tau} &= \left(v_{j,\tau}, \dot{v}_{j,\tau}\right)^{\mathrm{T}} \\ \boldsymbol{x}_{\mathrm{d},j,\tau} &= \left(d_{j,\tau}, \dot{d}_{j,\tau}\right)^{\mathrm{T}} \end{aligned} \tag{6.54}$$

further reducing the computational complexity of the motion upsampling algorithm. Each vector deploys the analogous process model, which is exemplarily explained for $\boldsymbol{x}_{\mathrm{d},j,\tau}$ in the remainder of this section.

For implementation as part of the digital image processing, a dynamic system is modeled by the discretized state and measurement equation

$$\begin{aligned} \boldsymbol{x}_{\mathrm{d},j,\tau} &= \boldsymbol{F}_{\tau-1}\boldsymbol{x}_{\mathrm{d},j,\tau-1} + \boldsymbol{w}_{\tau-1} \\ z_{\mathrm{d},j,\tau} &= \boldsymbol{H}_{\tau}\boldsymbol{x}_{\mathrm{d},j,\tau} + v_{\tau}, \end{aligned} \tag{6.55}$$

where the process and measurement noise are represented by normal probability distributions $\boldsymbol{w}_{\tau-1} \propto \mathcal{N}(0,\boldsymbol{Q}_{\tau-1})$ and $v_{\tau} \propto \mathcal{N}(0,R_{\tau})$ with covariance matrix $\boldsymbol{Q}_{\tau-1} \in \mathbb{R}^{2\times 2}$ and measurement variance $R_{\tau}$, respectively. The state transition and measurement matrix are denoted by $\boldsymbol{F}_{\tau-1} \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{H}_{\tau} \in \mathbb{R}^{1\times 2}$, respectively. The scalar disparity measurement $z_{\mathrm{d},j,\tau} = d_{j,\tau}$ is obtained from the stereoscopic tracking method cMR. For the system model deployed in this work, the state transition matrix is given as follows

$$\boldsymbol{F}_{\tau-1} = \begin{pmatrix} 1 & \Delta T \\ 0 & 1 \end{pmatrix}, \tag{6.56}$$

where $\Delta T = t_{\tau} - t_{\tau-1}$ is the sample time. Since the disparity rate is not measured directly, the measurement matrix $\boldsymbol{H}_{\tau} = (1\ 0)$ is constant. Consequently, the disparity process update of the Kalman filter is defined by the state and covariance prediction

$$\begin{aligned} \boldsymbol{x}^{-}_{\mathrm{d},j,\tau} &= \boldsymbol{F}_{\tau-1}\boldsymbol{x}_{\mathrm{d},j,\tau-1} \\ \boldsymbol{P}^{-}_{\tau} &= \boldsymbol{F}_{\tau-1}\boldsymbol{P}_{\tau-1}\boldsymbol{F}^{\mathrm{T}}_{\tau-1} + \boldsymbol{Q}_{\tau-1} \,. \end{aligned} \tag{6.57}$$

The measurement update equations taking the disparity observation into account are as follows

$$\begin{aligned} \boldsymbol{x}_{\mathrm{d},j,\tau} &= \boldsymbol{x}^{-}_{\mathrm{d},j,\tau} + \boldsymbol{K}_{\tau}\big(z_{\mathrm{d},j,\tau} - \boldsymbol{H}_{\tau}\boldsymbol{x}^{-}_{\mathrm{d},j,\tau}\big) \\ \boldsymbol{K}_{\tau} &= \boldsymbol{P}^{-}_{\tau}\boldsymbol{H}^{\mathrm{T}}_{\tau}\big(\boldsymbol{H}_{\tau}\boldsymbol{P}^{-}_{\tau}\boldsymbol{H}^{\mathrm{T}}_{\tau} + R_{\tau}\big)^{-1} \\ \boldsymbol{P}_{\tau} &= \big(\boldsymbol{I} - \boldsymbol{K}_{\tau}\boldsymbol{H}_{\tau}\big)\boldsymbol{P}^{-}_{\tau}, \end{aligned} \tag{6.58}$$

where $\boldsymbol{P}_{\tau} \in \mathbb{R}^{2\times 2}$ denotes the estimated state covariance and $\boldsymbol{K}_{\tau} \in \mathbb{R}^{2\times 1}$ is the Kalman gain. The process noise is assumed to have zero mean and covariance matrix

$$\boldsymbol{Q}_{\tau} = \begin{pmatrix} \frac{\Delta T^4}{4} & \frac{\Delta T^3}{2} \\ \frac{\Delta T^3}{2} & \Delta T^2 \end{pmatrix} \sigma_Q^2 \tag{6.59}$$

depending on the uncertainty $\sigma_Q$. Since only position is measured, the related noise variance is defined by the constant $R_{\tau} = \sigma_R^2$. For all three motion directions, process and measurement uncertainties are empirically set to $(\sigma_Q,\sigma_R) = (1.0,0.001)$.

The Kalman filtering scheme is applied according to Figure 6.5d. If no measurement is available, only state prediction by Equation 6.57 is performed to reduce the latency-dependent misalignment. Once an image-based tracking result is available, Equation 6.58 corrects the motion estimate.

### 6.1.6 Performance Assessment on *In Vivo* Sequences

Tracking performance is assessed on laparoscopic, beating heart, and laryngeal *in vivo* tissue (IVT) datasets. Five reference points $\boldsymbol{P}_{m,\mathrm{GT}}$ with $m \in \{1, \dots, 5\}$, mostly located on distinctive blood vessels, representing the ground truth (GT) were manually selected beforehand by an experienced observer. For each sequence, eleven frames with GT were defined (equally distributed along the sequence). The error of the back-projected points with reference to frame $(\mathrm{CF})_\mathrm{L}$ is then given by

$$e_\mathrm{Track}\left(\boldsymbol{P}_m\right) = \left\| {}_{(\mathrm{L})}\boldsymbol{P}_{m,\mathrm{Track}} - {}_{(\mathrm{L})}\boldsymbol{P}_{m,\mathrm{GT}} \right\|_2 \,. \tag{6.60}$$

In total, eight stereo sequences are considered (see Table 6.1). Sequences SEQ1–5 are obtained from the laparoscopic Hamlyn dataset [HCL17]. The first three videos, denoted by SEQ1–3, are adopted from a laparoscopic porcine procedure including a scale change, a simulated occlusion, and significant deformation. Datasets SEQ4–5 describe challenging beating heart scenarios [SMD$^+$05, RPL10]. Videos SEQ6–8 were captured with a stereo endoscope (VSii, Visionsense, Petach-Tikva, Israel) in an *in vivo* laryngeal intervention, conducted by Prof. Giorgio Peretti from the Department of Otorhinolaryngology, University of Genoa, Italy. The data collection was part of the µRALP project [uRA15]. Regarding the Hamlyn dataset, the tracked region is located at a distance of 170 mm (camera baseline ~5 mm). For the beating heart and the µRALP sequences, the distance amounts to 40 mm (baseline ~5 mm) and 20 mm (baseline ~1 mm) on average, respectively.

Table 6.1: Scenarios for tracking on *in vivo* tissue (IVT).

| No. | Frames | Description |
|---|---|---|
| SEQ1 | 350 | Hamlyn-sequence with scale change |
| SEQ2 | 650 | Hamlyn sequence with simulated occlusion |
| SEQ3 | 140 | Hamlyn sequence with large deformation |
| SEQ4 | 338 | Hamlyn sequence of beating heart #1 |
| SEQ5 | 630 | Hamlyn sequence of beating heart #2 |
| SEQ6 | 600 | µRALP sequence with deformation |
| SEQ7 | 368 | µRALP sequence with partial occlusion |
| SEQ8 | 448 | µRALP sequence with large deformation |

For the evaluation study, Algorithm 6.2 is parametrized with $\lambda_\mathrm{D} = 0.01$, $\beta = 0.03$, and a triangle width of 35 pixels. In Algorithm 6.3, the Huber threshold is set to $\sigma_H = 10$. Three pyramidal levels are considered with a maximum of 20 iterations per level and a stop criterion of $\epsilon = 0.03$.

Results of the tracking cMRKF are compared with not only those of noMR and sMR but also those of state-of-the-art algorithms, namely, a non-rigid TPS-based tracking with $3 \times 3$ control points [RPL10], and the hierarchical multi-affine (HMA) feature-matching toolbox [PSM13]. The reimplemented TPS method includes specular highlight filtering and CUDA optimization. Regarding the HMA algorithm, features are matched either with respect to the initial frame (HMAi) or between consecutive frames (HMAc). Surface reconstruction (see Section 3.1) is used establish left-right correspondence for the monoscopic HMA algorithm and to initialize the TPS.

### 6.1.7 Results

Figure 6.7 illustrates the error plots for each sequence. The associated mean and standard deviation (SD) as well as the root mean square error (RMSE) are listed in Table 6.2. The results demonstrate

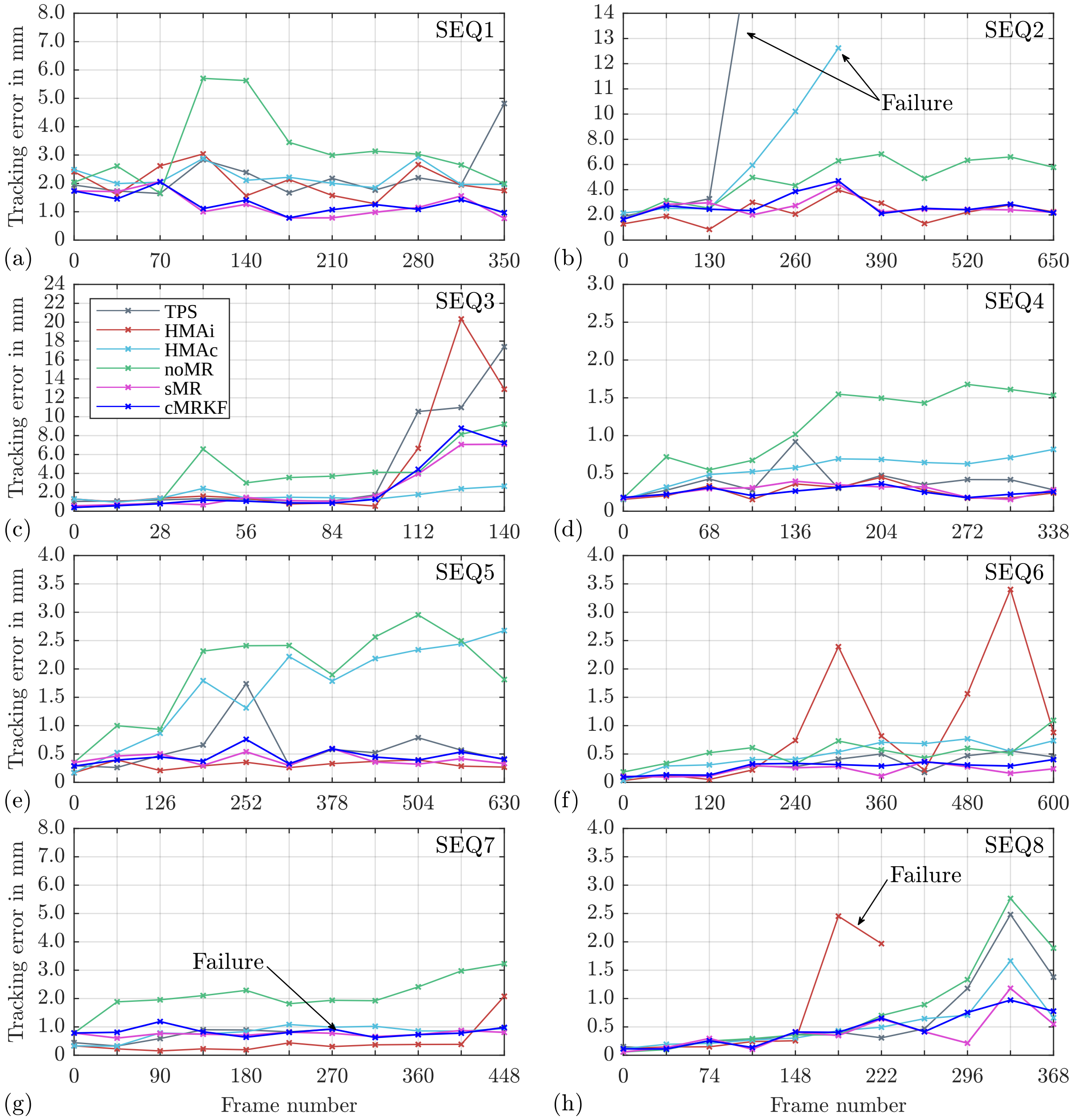


Figure 6.7: Tracking error in mm for the IVT sequences SEQ1–8. Accuracy is evaluated with respect to ground truth measured for eleven frames per sequence (distributed equally along each sequence).

Table 6.2: IVT tracking error in mm. Accuracy values are valid until tracking failure (F). **Blue** numbers represent the best performance, whereas magenta numbers denote the second best performance.

| | TPS | HMAi | HMAc | noMR | sMR | cMRKF |
|---|---|---|---|---|---|---|
| | Mean ± SD<br>RMSE | Mean ± SD<br>RMSE | Mean ± SD<br>RMSE | Mean ± SD<br>RMSE | Mean ± SD<br>RMSE | Mean ± SD<br>RMSE |
| SEQ1 | 1.91 ± 1.54 | 1.82 ± 1.10 | 2.07 ± 0.89 | 3.00 ± 1.67 | **1.07 ± 0.79** | 1.14 ± 0.73 |
| | 2.44 | 2.12 | 2.25 | 3.42 | **1.33** | 1.35 |
| SEQ2 | 5.61 ± 11.4 (F) | **1.98 ± 1.36** | 5.42 ± 6.05 (F) | 4.40 ± 2.66 | 2.20 ± 1.53 | 2.48 ± 1.38 |
| | 12.5 (F) | **2.39** | 8.05 (F) | 5.13 | 2.67 | 2.83 |
| SEQ3 | 3.66 ± 6.08 | 4.19 ± 6.40 | **1.43 ± 1.03** | 3.25 ± 3.71 | 2.13 ± 2.63 | 2.21 ± 3.07 |
| | 7.05 | 7.60 | **1.75** | 4.91 | 3.37 | 3.77 |
| SEQ4 | 0.36 ± 0.25 | 0.22 ± 0.17 | 0.50 ± 0.33 | 1.04 ± 0.67 | 0.24 ± 0.16 | **0.23 ± 0.12** |
| | 0.43 | 0.27 | 0.60 | 1.24 | 0.28 | **0.26** |
| SEQ5 | 0.48 ± 0.54 | **0.26 ± 0.17** | 1.50 ± 1.09 | 1.65 ± 1.27 | 0.36 ± 0.21 | 0.40 ± 0.24 |
| | 0.72 | **0.31** | 1.85 | 2.08 | 0.42 | 0.47 |
| SEQ6 | 0.25 ± 0.24 | 0.65 ± 1.26 | 0.42 ± 0.33 | 0.48 ± 0.33 | **0.17 ± 0.16** | 0.22 ± 0.18 |
| | 0.35 | 1.41 | 0.54 | 0.59 | **0.23** | 0.29 |
| SEQ7 | 0.59 ± 0.44 (F) | **0.38 ± 0.58** | 0.64 ± 0.54 | 1.97 ± 1.00 | 0.66 ± 0.36 | 0.75 ± 0.38 |
| | 0.73 (F) | **0.69** | 0.83 | 2.20 | 0.75 | 0.84 |
| SEQ8 | 0.49 ± 0.84 | 0.66 ± 1.02 (F) | 0.43 ± 0.51 | 0.72 ± 0.91 | **0.33 ± 0.38** | 0.38 ± 0.38 |
| | 0.96 | 1.20 (F) | 0.66 | 1.16 | **0.50** | 0.54 |

superior performance of the tracking with mesh refinement when either sMR or cMRKF is applied. Since no differences between cMR and cMRKF during the IVT validation are observed, presenting the results of cMR is skipped in this section. A comparison of the two methods is provided in the next section, where the GT hexapod motion is taken into account.

According to the results listed in Table 6.2, the TPS method is able to adequately track tissue deformation in most cases; however, it fails in sequences SEQ2 and SEQ7 due to partial occlusion and rapid motion, respectively. The feature matching strategy HMAi provides high accuracy in scenes with smooth deformation, as in the beating heart sequences SEQ4–5, and under partial occlusions, as illustrated by SEQ2 and SEQ7. However, HMAi-based tracking of large tissue deformation, as in SEQ3 as well as SEQ8, shows poor performance and even tracking failure (see Figure 6.7). In addition, there is no temporal consistency when HMAi (flickering, e.g., in SEQ6) is used. These limitations can be successfully addressed by matching features on subsequent frames employing HMAc alternatively; however, this method fails at partial occlusions and suffers from drift, as illustrated by SEQ2 and SEQ4, respectively.

By contrast, tracking with sMR or cMRKF performs accurately in all scenarios, without tracking failure. The norm-like Huber function penalizes partial occlusions to some extent, as shown in Figure 6.8a, where the instrument tip enters the tracked region. When the MHD-based detection scheme (see Figure 6.8b) is incorporated into the reweighting process (see Equation 6.49), robustness to partial occlusions, such as those caused by instruments or laser ablation with carbonization, can be improved. In comparison with noMR (i.e. SEQ3–5), drift is successfully eliminated.

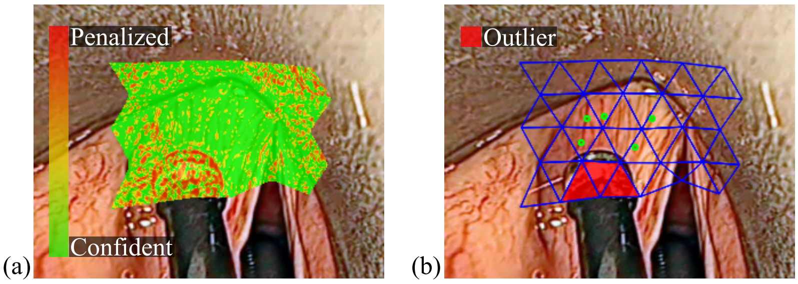


Figure 6.8: Instrument-induced partial occlusion in SEQ7 penalized by implementing (a) the Huber loss function and (b) the MHD-based detection scheme.

Decomposing the RMSE for the cMRKF method reveals that the error in $z$-direction predominates. For sequence SEQ1, the RMSE amounts to $(e_x,e_y,e_z) = (0.22,0.26,1.31)\,\mathrm{mm}$, revealing that the $z$-error is approximately five times higher compared with that in the other two spatial directions. Since the depth resolution rises with an increasing baseline-to-distance ratio, a reduced predominance in the $z$-direction is revealed for SEQ4 and SEQ6, yielding $(e_x,e_y,e_z) = (0.10,0.10,0.22)\,\mathrm{mm}$ and $(e_x,e_y,e_z) = (0.08,0.12,0.25)\,\mathrm{mm}$, respectively.

Computation time is discussed for SEQ8, which exhibits significant deformation, and is depicted in Figure 6.9 and Table 6.3. Initially, the TPS method runs at nearly constant $13\,\mathrm{ms}$ per frame, since only slight motion occurs, that is estimated in a few iterations. Since tracking of deformation requires a higher number of iterations to converge, the runtime drastically increases up to $70\,\mathrm{ms}$. Even though it is more accurate, the sMR method converges as non-deterministic as the TPS approach. This observation is substantiated by the high standard deviation of $14.5\,\mathrm{ms}$ (see Table 6.3).

Even though no restriction on the computation time was imposed to ensure convergence, the

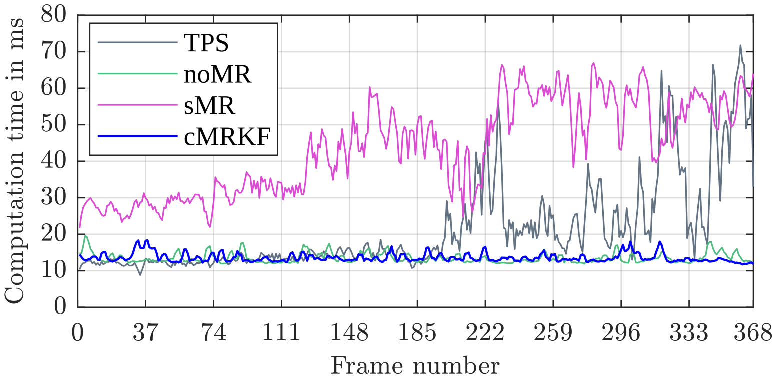


Figure 6.9: Runtime in milliseconds for SEQ8 (8 × 4 triangle mesh).

Table 6.3: Runtime in milliseconds for SEQ8 (mean ± standard deviation (SD)).

| | TPS | noMR | sMR | cMRKF |
|---|---|---|---|---|
| Mean ± SD | 21.5 ± 15.6 | 13.5 ± 1.8 | 43.7 ± 14.5 | 13.6 ± 1.8 |

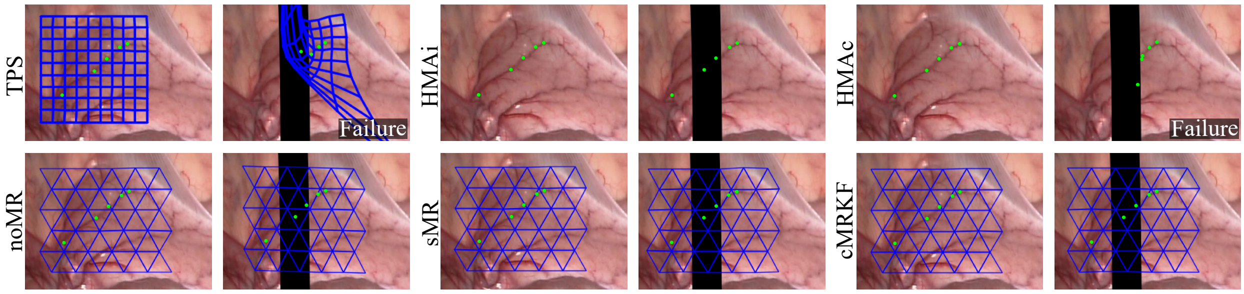


Figure 6.10: Comparison of tracking results in laparoscopic sequence SEQ2 whereas frame 2 (left) and frame 254 (right) are shown. Five landmarks (green dots) are tracked. A partial occlusion is simulated by the bar moving from left to right. Tracking failure is detected for methods TPS and HMAc.

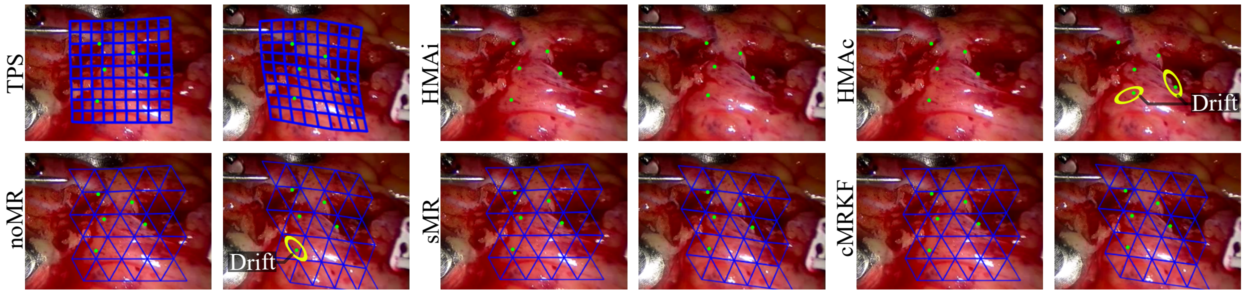


Figure 6.11: Comparison of tracking results in beating heart sequence SEQ4 whereas frame 2 (left) and frame 323 (right) are shown. Five landmarks (green dots) are tracked with respect to ground truth. Drift of certain landmarks was observed for methods HMAc and noMR.

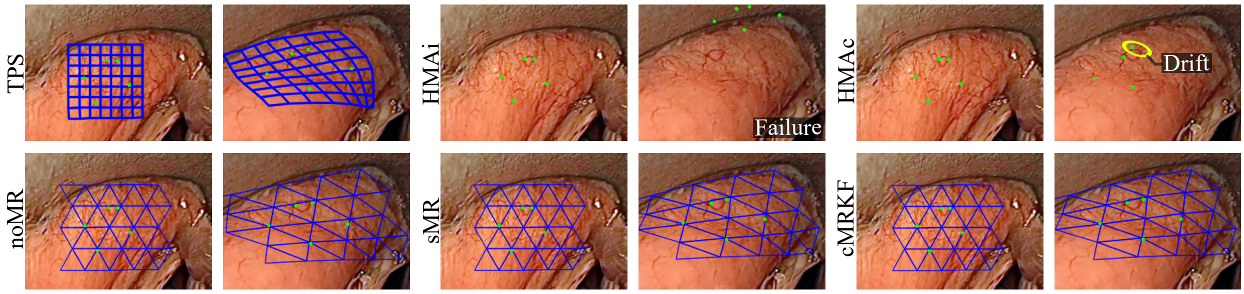


Figure 6.12: Comparison of tracking results in laryngeal sequence SEQ8 whereas frame 2 (left) and frame 216 (right) are shown. Five landmarks (green dots) are tracked with respect to ground truth. Drift was observed for method HMAc. Tracking failure was detected for method HMAi.

available runtime of the cMRKF method was limited to 50 ms (framerate of 20 Hz) in order to demonstrate the fusion of tracking and delayed mesh refinement. On this condition, cMRKF shows a constant run-time of $(13.6 \pm 1.8)$ ms for the entire sequence SEQ8. Consequently, cMRKF combines the computational efficiency of noMR with the drift-free tracking accuracy of sMR.

Unfortunately, reproducing the HMA runtime was not possible; thus, an average time of $(50 \pm 20)$ ms per frame was assumed, as presented in the original work [PSM13]. Considering the additional time of 35 ms required to establish stereo correspondence, an overall matching time of at least $85$ ms reveals that real-time capability similar to TPS and sMR cannot be achieved. Thus, online laser control can only be addressed by cMRKF.

The results of the deformation tracking are exemplarily illustrated in Figure 6.10–6.12. Due to the complexity and limited resolution of manually acquiring ground truth data, the IVT results do not provide quantitative evidence for the entire sequence, especially between consecutive frames. Thus, the next section provides a more detailed analysis on the delay-dependent tracking error in order to assess the real-time capability of the presented mesh refinement strategies.

To conclude, non-rigid tracking based on a linear, easy-to-understand parametrization enforcing left-right consistency for stereo vision has been presented in this section. In contrast to computationally expensive, direct methods discussed in literature, dense texture information is processed concurrently to correct tracking misalignment. Thus, highly accurate, online-capable motion estimation as a prerequisite for intraoperative assistance such as vision-guided ablation control in laser microsurgery is enabled. Tracking robustness is enhanced by incorporating efficient outlier rejection into the robust estimator-based mesh refinement step. The experimental outcome on *in vivo* data demonstrates enhanced accuracy compared to state-of-the-art methods.

Even though the parameter set is optimized in disparity space, solely back-projection of the tracked mesh is required in order to map the motion estimate to task space and thus to enable laser positioning and focusing on the target surface. Regardless of application in laser surgery as discussed in the following sections, vision-guided control of further surgical or even robotic tools is conceivable.

## 6.2 Application to Image Stabilization during Incision Planning

Prospective interfaces for laser surgery require an optimal setting of visualization and interactive planning. In addition to accurate ablation path definition as discussed in Section 5.3, a further technical challenge affecting the surgical performance arises from soft tissue motion. To address this, the following section describes a novel method that incorporates online motion compensation into tablet-based incision planning. The non-rigid tracking scheme outlined in Section 6.1 is used to compute an inverse mapping for image stabilization. Figure 6.13a–b exemplarily demonstrate this technique for a laparoscopic and a laryngeal sequence. In comparison to the non-compensated (NC) view, motion compensation (MC) aims at providing the surgeon with a stabilized live image at time $t = t_\tau$ that highly correlates with the initial view at $t = 0$, e.g., when initiating path planning. A user study is conducted and its results are discussed with respect to path tracing accuracy and usability when performing free-hand drawing with a stylus. Improved performance, as demonstrated with image stabilization, highlights the potential of the methodology in laser-assisted surgery.

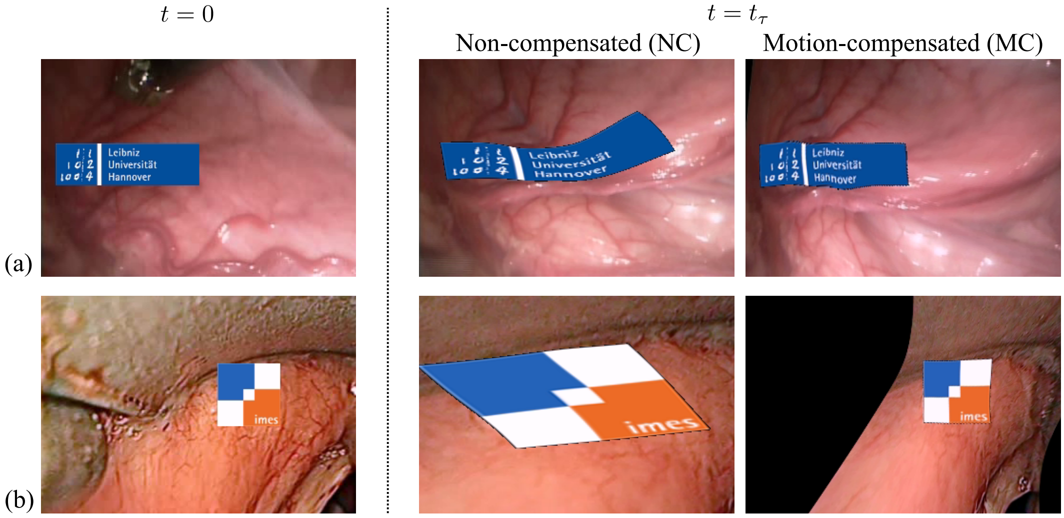


Figure 6.13: Image stabilization concept shown for (a) a laparoscopic sequence with significant deformation [HCL17], and (b) a laryngeal dataset with additional scale changes [Per14]. The initial frame at $t = 0$ is shown on the left whereas its deformed state at $t = t_\tau$ is depicted on the right illustrating both the non-compensated (NC) and the motion-compensated (MC) image view of the logo.

### 6.2.1 User Study Design

The developed tracking scheme is integrated into a tablet-based interface for laser path definition, as described in Section 5.3. The framework aims at short-term image stabilization to increase the planning accuracy. Modern stylus-tablet configurations as the one used in this section (Intuos CTH-480S, Wacom Co. Ltd., Japan) provide a proximity range of a certain height above the active area of the tablet in which the stylus is detected. In combination with online motion estimation, the proximity mode can trigger the image stabilization as a prerequisite to relieve hand-eye coordination during incision planning once the optimal field of view is established.

Deploying the setup shown in Figure 6.14a, the user study addresses free-hand tracing of a blood vessel being subject to tissue motion (see Figure 6.14b). Ground truth of has been acquired in

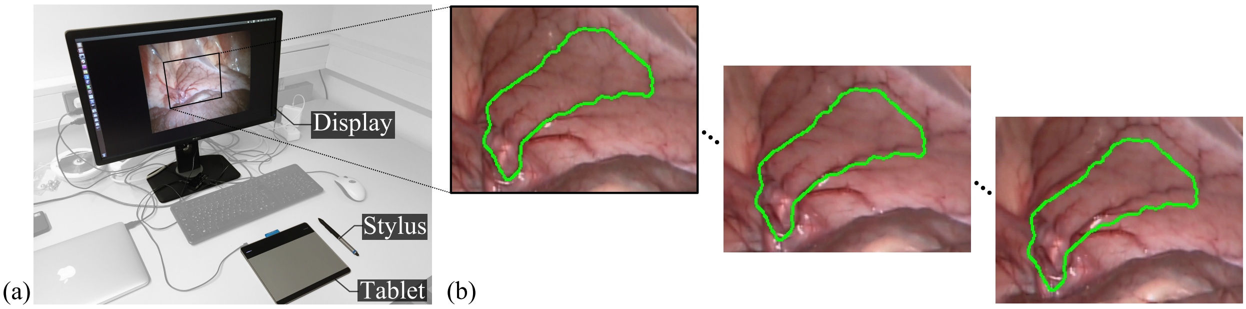


Figure 6.14: User performance is assessed with the setup shown in (a) consisting of a display, a stylus, and a tablet. The green line (b) marks the blood vessel with respect to the initial frame of the sequence.

the initial frame. It is then tracked on the deforming scene yielding an approximate of the true position over time. Despite that the tracking provides consistent stereo information, monoscopic visualization is chosen to avoid visual discomfort when using a 3D monitor, as found in Section 5.3.

For the user performance assessment, 22 non-expert subjects with background in medical engineering have been asked to trace the blood vessel with the stylus on the non-compensated (NC) and on the motion compensated (MC) scene. In the latter case, the tracked deformation is transformed back to its initial pose stabilizing the view onto the target structure. The distance error between the traced path and ground truth as well as required task completion time are analyzed. Usability is assessed by carrying out the After Scenario Questionnaire (ASQ) subsequently to each task. The ASQ consists of hypotheses for assessing user satisfaction in terms of ease of use and task completion time [Lew91]. Scoring is defined by seven-point Likert scale (1 - strongly disagree, 7 - strongly agree). High rating correlates with increased user satisfaction. The evaluated ASQ hypotheses are

Hyp–1: “Overall, I am satisfied with the ease of completing the tasks in this scenario”, and
Hyp–2: “Overall, I am satisfied with the amount of time it took to complete the tasks”.

### 6.2.2 Results

In the following, the experimental outcome of the user study is discussed demonstrating the potential of motion compensation in a tablet-based planning interface for laser surgery. Figure 6.15a–c summarize the distance errors and the task completion time for both the non-compensated (NC) and the motion compensated (MC) sequence. The RMSE median of path tracing reduces from 2.73 mm (NC) to 2.14 mm (MC). Applying the non-parametric Wilcoxon signed-rank test (significance

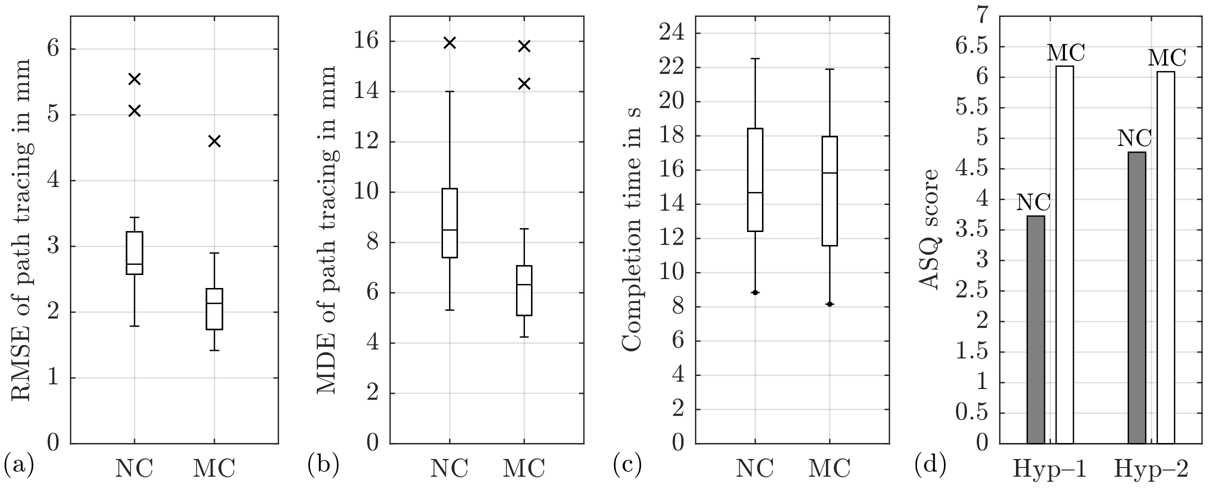


Figure 6.15: User study results on accuracy (a–b), completion time (c), and ASQ (d) measured for the non-compensated (NC) and motion compensated (MC) scene.

level $p = 0.05$) reveals superior performance for the motion compensated planning scenario ($p = 4 \times 10^{-5}$). Same applies for the maximum distance error (MDE) of 8.5 mm (NC) and 6.34 mm (MC) ($p = 1.8 \times 10^{-4}$). In addition, there is a significant difference between measured time medians of $14.68\,\mathrm{s}$ (NC) and $15.83\,\mathrm{s}$ (MC) ($p = 0.036$). One hypothesis is that the subjects spent more time on completing the task in the stabilized view (MC) as they realized that the desired path can be traced more precisely. By contrast, the stylus has to be moved quickly for the scenario NC in order to compensate for scene motion and thereby to reduce erroneous free-hand drawing.

The results of the conducted ASQ are shown in Figure 6.15d. On the one hand, the rating of Hyp–$1$ indicates superior ease of use for MC compared to NC. On the other hand, an increase in the user satisfaction in terms of task completion time has also been measured for MC; even though, it took longer to trace the path compared with NC (see Hyp–2). This can be associated with the positive feedback on achieving higher path tracing accuracy. To summarize, the gain in the user performance demonstrates the potential of the motion tracking framework integrated into a stylus-tablet-based interface for laser surgery. Since usability is discussed for monoscopic visualization only, future studies have to address planning in the stereo view without compromising the visual comfort of the 3D display.

## 6.3 Application to Motion Compensation during Laser Ablation

The accuracy evaluation on *in vivo* data, as described in Section 6.1, demonstrates superior performance of the tracking scheme compared to existing state-of-the-art algorithms. However, integration of the computational method into a clinical setting demands for real-time capability. In particular when aiming at online assistance for laser surgery, the entire processing routine including image acquisition and rectification, tracking and mesh refinement as well as surgical tool control (e.g. of the ablation laser) should not cause latencies that compromise required ablation accuracy.

This section is dedicated to embedding the developed tracking strategies into a laser setup for soft tissue ablation. Initially, the latency-dependent tracking error is assessed. Therefor, a high-precision parallel-kinematic platform is employed to generate ground truth trajectories taking two motion patterns into account. A comparative study considering tracking with and without mesh refinement as well as concurrent processing and motion upsampling is presented. Finally, motion-compensated laser ablation on moving tissue samples is discussed.

### 6.3.1 Experimental Design

In the following, the system design for online performance evaluation of the tracking is presented. Based on integration into a surgical framework according to Figure 6.1, laser ablation trials conducted on tissue substitute and porcine *ex vivo* tissue (EVT) are described.

### Assessment of the Latency-Dependent Tracking Error

In addition to the *in vivo* validation (see Section 6.1), the latency-dependent tracking error is assessed to demonstrate the superior performance of cMRKF compared with noMR and sMR. To provide motion ground truth (GT), a tissue sample is translated with a high-precision, parallel-kinematic platform (Hexapod H-824.G11, Physik Instrumente (PI), Karlsruhe, Germany) that has a repeatability of $\pm 0.5\,\mu\mathrm{m}$. Stereo images are acquired with two cameras (UI-3370-CP-C-HQ, IDS Imaging Development Systems GmbH, Obersulm, Germany) equipped with C-mount lenses (FL-HC0614-2M, Ricoh Company, Ltd., Tokyo, Japan). The two cameras are mounted with a baseline of $37\,\mathrm{mm}$ at a sample distance of $60\,\mathrm{mm}$. A schematic overview is shown in Figure 6.16a.

For simplicity, tissue deformation is not considered in this part of the evaluation, since acquiring online ground truth is complex. Instead, rigid movements of the sample are performed while GT is measured from the hexapod encoder. An incision line defined by points $\boldsymbol{P}_{m,\mathrm{GT}}$ is planned in sample frame $(\mathrm{CF})_{\mathrm{S}}$. The position with respect to the hexapod home frame $(\mathrm{CF})_{\mathrm{H,0}}$ is calculated by

$$_{(\mathrm{H,0})}\tilde{\boldsymbol{P}}_{m,\mathrm{GT}} = {}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H,0}}^{-1}\, {}_{(\mathrm{S})}\tilde{\boldsymbol{P}}_{m,\mathrm{GT}}, \tag{6.61}$$

where position $\tilde{\boldsymbol{P}}_{m,\mathrm{GT}}$ is represented in homogeneous coordinates. Transform ${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H,0}}$ maps the incision line from frame $(\mathrm{CF})_{\mathrm{S}}$ to frame $(\mathrm{CF})_{\mathrm{H,0}}$ and is given by

$${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H,0}} = {}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H}}\, {}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{H,0}}, \tag{6.62}$$

where ${}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{H,0}}$ is measured from the hexapod encoders. The unknown but constant transform ${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H}}$ between the sample and hexapod frame is obtained by

$${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H}} = {}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{S,init}}^{-1}\, {}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{H,0}}\, {}^{\mathrm{H,init}}\boldsymbol{T}_{\mathrm{H,0}}^{-1} \tag{6.63}$$

assuming an arbitrary initial pose ${}^{\mathrm{H,init}}\boldsymbol{T}_{\mathrm{H,0}}$ that may differ from the hexapod home pose. Transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{S,init}}$ is assumed to have its origin at the first point of the planned incision, whereas its orientation is set equal to the initial hexapod rotation with respect to $(\mathrm{CF})_{\mathrm{L}}$. The image-based tracking result is finally mapped by

$$_{(\mathrm{H,0})}\tilde{\boldsymbol{P}}_{m,\mathrm{track}} = {}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{H,0}}^{-1}\, {}_{(\mathrm{L})}\tilde{\boldsymbol{P}}_{m,\mathrm{track}} \tag{6.64}$$

to hexapod frame $(\mathrm{CF})_{\mathrm{H,0}}$, whereas the camera-to-hexapod transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{H,0}}$ is computed offline by hand-eye calibration [TL89]. The latency-dependent misalignment (LD) caused by the image acquisition and processing is then assessed with respect to GT by the error function

$$e_{\mathrm{LD}}\left(\boldsymbol{P}_m\right) = \left\| {}_{(\mathrm{H,0})}\boldsymbol{P}_{m,\mathrm{track}} - {}_{(\mathrm{H,0})}\boldsymbol{P}_{m,\mathrm{GT}} \right\|_2 . \tag{6.65}$$

During the experiments, two motion patterns were considered in order to assess (1) the drift when noMR is used, (2) the online performance of sMR as well as cMR in compensating for the

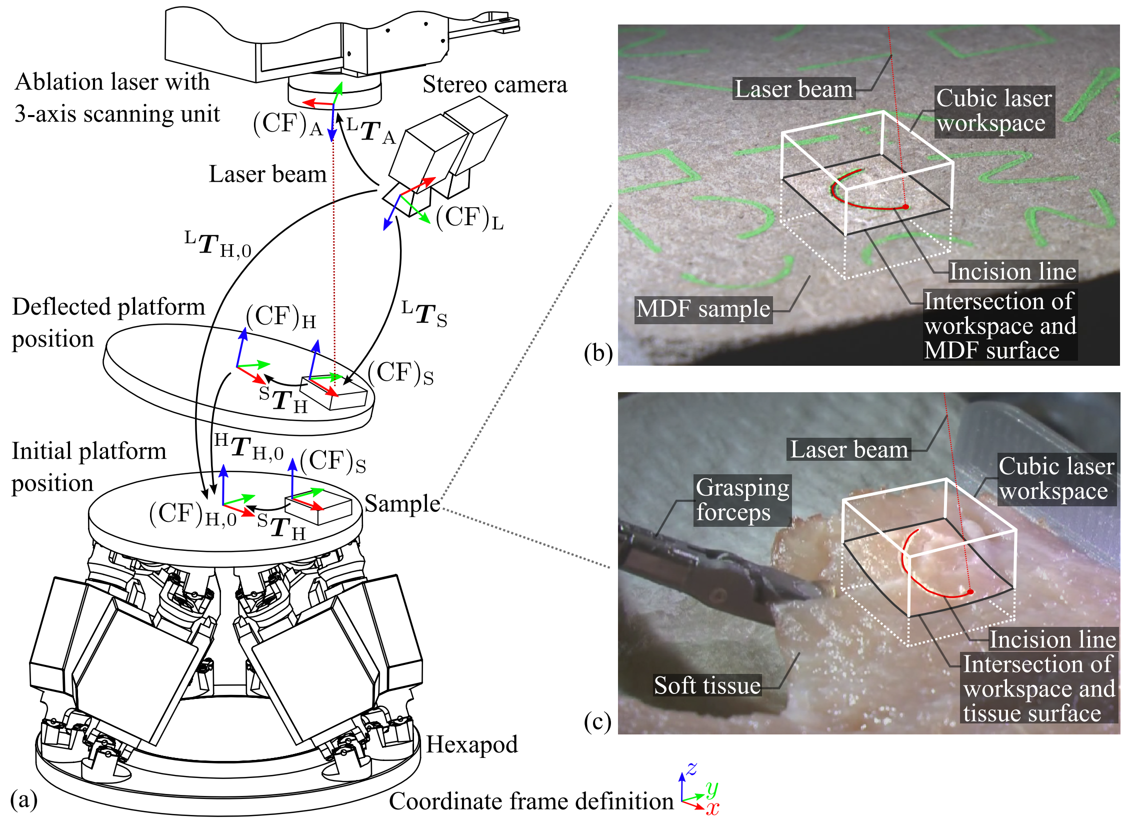

Ablation laser with
3-axis scanning unit
Stereo camera
$(\mathrm{CF})_\mathrm{A}$
${}^{\mathrm{L}}\boldsymbol{T}_\mathrm{A}$
Laser beam
$(\mathrm{CF})_\mathrm{L}$
${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{H},0}$
Deflected platform
position
$(\mathrm{CF})_\mathrm{H}$
${}^{\mathrm{L}}\boldsymbol{T}_\mathrm{S}$
$(\mathrm{CF})_\mathrm{S}$
${}^{\mathrm{S}}\boldsymbol{T}_\mathrm{H}$
Initial platform
position
${}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{H},0}$
$(\mathrm{CF})_\mathrm{S}$
$(\mathrm{CF})_{\mathrm{H},0}$
${}^{\mathrm{S}}\boldsymbol{T}_\mathrm{H}$
Sample
Hexapod
(a)
Coordinate frame definition
$z$
$y$
$x$
Laser beam
Cubic laser
workspace
Incision line
MDF sample
Intersection of
workspace and
MDF surface
(b)
Laser beam
Grasping
forceps
Cubic laser
workspace
Incision line
Soft tissue
Intersection of
workspace and
tissue surface
(c)

Figure 6.16: Experimental design is shown in (a) with a rigid setup deploying a stereo camera, a surgical laser, and a parallel robot for positioning tasks to assess tracking performance. Motion estimation and laser ablation trials conducted on tissue substitute (MDF) and *ex vivo* tissue (EVT) samples are shown in (b,c). For both specimens, the surface has to be positioned in the cubic laser workspace.

aforementioned drift, and (3) the capability of the proposed motion upsampling cMRKF to further reduce the latency-dependent tracking misalignment. The first scenario, which is called lateral, considers movements in the lateral direction (along the $y-$axis of $(\mathrm{CF})_{\mathrm{H},0}$), i.e., perpendicular to the optical axis of the the laser. In a clinical scenario, such a motion can be induced by camera motion or tissue manipulation with grasping forceps to expose the tissue during ablation. To point out performance differences when tracking with mesh refinement, concurrent processing, and motion upsampling, the trajectory is repeated $10$ times with an amplitude of $3\,\mathrm{mm}$ and a maximum velocity of $2.1\,\mathrm{mm/s}$. The second scenario, which is called axial, is defined by movements with an amplitude of $4\,\mathrm{mm}$ at $1\,\mathrm{mm/s}$ in the depth direction (along the $z-$axis of $(\mathrm{CF})_{\mathrm{H},0}$), which is perpendicular to the optical axis. Hereby, a clinical scenario with changing distance between the tissue surface and the laser is simulated. Tracking such a motion enables continuous adjustment of the laser focus for optimal ablation characteristics.

For each motion pattern, two types of tissue are considered in the experimental study. As illustrated in Figure 6.16b, tracking is initially performed on a highly textured, non-reflective medium density fiberboard (MDF) to demonstrate tracking under ideal conditions. To mimic clinical conditions, porcine *ex vivo* tissue (EVT) is tracked in an additional scenario to assess the performance on glossy and weakly textured environment (see Figure 6.16c). The active sensor area of the camera was cropped to $400 \times 400$ pixels enabling an image acquisition frame rate of $80\,\mathrm{Hz}$. The stereo camera system was calibrated with a re-projection error of $0.1$ pixel. To achieve online-capability of tracking, a mesh of $6 \times 4$ triangles with an edge length of $75$ pixels was chosen. Motion upsampling rate was set to $200\,\mathrm{Hz}$ in accordance with the hexapod encoder sampling rate.

### Laser Ablation Framework

To demonstrate vision-guided laser control, ablation trials were conducted on MDF and EVT utilizing the laser setup described in Section 2.5. The cameras' field of view is optimized with respect to the area of intersection between tissue surface and laser scanning range. Prior to the ablation trials, laser-to-camera transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{A}}$, as shown in Figure 6.16a, is estimated by the registration method outlined in Section 4.1.

The results of ablating a straight and a curved line when cMRKF is used are discussed. Such scan patterns are commonly employed in transoral laser microsurgery and are provided by commercial systems. During the experiments, the laser settings were set to constant pulse duration $\tau_{\mathrm{P}} = 150\,\mu\mathrm{s}$, diode current $I_{\mathrm{D}} = 150\,\mathrm{A}$, and pulse frequency $f_{\mathrm{P}} = 220\,\mathrm{Hz}$. Multiple passes with a scanning velocity of $v_{\mathrm{S}} = 200\,\mathrm{mm/s}$ were performed, minimizing the risk of local thermal damage of the tissue. The entire image processing and control software was implemented on a nodelet-based, high-level control layer deploying C++ and the Robot Operating System (ROS) [QGC$^{+}$09, ROS].

### Laser Ablation Trials on MDF

Straight and curved lines were stamped with green ink onto the MDF sample that was positioned in the laser focal range using the hexapod (see Figure 6.16b). Tracking and ablation were simultaneously performed considering the lateral and axial motion patterns. The root mean square error (RMSE) was computed between the initial and the ablated shape, both segmented by thresholding. Ablation on a static sample was also performed to quantify the impact of the laser-to-camera registration.

### Laser Ablation Trials on EVT

Path tracing on the EVT was conducted to demonstrate online laser control on biological tissue. In contrast to the MDF sample, the straight and curved incision lines were manually defined and segmented after ablation using a stylus-based tablet interface (see Section 5.3). Ablation accuracy was assessed for comparing (1) the two strategies sMR and cMRKF under lateral motion, and (2) non-focused and focused ablation while the sample is moved in the axial direction. Finally, the paths were analyzed under microscopic imaging regarding ablation quality, shape, and carbonization.

Qualitative validation of motion compensation is provided for ablation on tissue manipulated with a surgical forceps (Serpent Articulating Grasping Forceps 3 mm, Smith & Nephew plc, London, UK). As shown in Figure 6.16c, a tissue sample mimicking a vocal fold was prepared and push-pull movements were induced to expose the tissue in the laser workspace. Moreover, the trials included deformation in the axial direction to simulate respiratory motion artifacts. Due to the limited laser workspace, only small movements were feasible.

### 6.3.2 Results

In the following, results of the performance assessment addressing the latency-dependent tracking error and laser ablation accuracy are presented.

#### Assessment of the Latency-Dependent Tracking Error

Two cyclic motion trajectories were carried out on the MDF and EVT sample to assess the latency-dependent tracking error. The results of tracking the lateral MDF motion are shown in Figure 6.17, including the position over time, the associated ground truth and the tracking error. Due to the very small movement, error effects such as from hand-eye calibration can be neglected. Even though this scenario can be regarded as tracking under ideal conditions (significant texture, no specular highlights, or occlusions), a remaining misalignment due to drift is observed at the end of the trajectory when noMR is used (see Figure 6.17a–b). By contrast, tracking with subsequent mesh refinement (sMR) compensates for drift; however, it drastically increases the tracking error (see Figure 6.17c–d). Since the motion estimate is computed with significant delay, the error grows to $0.78$ mm when the sample is moved in the lateral direction at the maximum velocity of $2.1$ mm/s. If the mesh refinement is processed concurrently (cMR), real-time performance with simultaneous compensation for drift is achieved (see Figure 6.17e–f). The maximum deviation of the lateral position does not exceed $0.09$ mm. Further reduction of the latency-dependent misalignment is attained with filter-based motion upsampling (cMRKF), as shown in Figure 6.17g–h.

The error curves of Figure 6.17 are summarized in the form of box plots shown in Figure 6.18a. For each strategy, two box plots are shown. The left and right plot represent the lateral and axial tracking result, respectively. The red-colored circle defines the remaining misalignment at the end of the motion pattern. The related error values are listed in Table 6.4.

Regarding motion estimation on the EVT sample, the influence of drift is significantly more distinct when noMR is used. The maximum error is $0.505$ mm in lateral and $1.29$ mm in axial direction, respectively. As in the trials outlined above, concurrent mesh refinement cMR drastically reduces the temporal misalignment compared with the sequential method sMR. cMRKF outperforms all other methods, providing a RMSE of below $0.05$ mm for lateral and axial movements. Since tracking on the glossy tissue sample is affected by specular highlights and reduced texture, the related error values, as listed in Table 6.4, are higher than those for the MDF specimen.

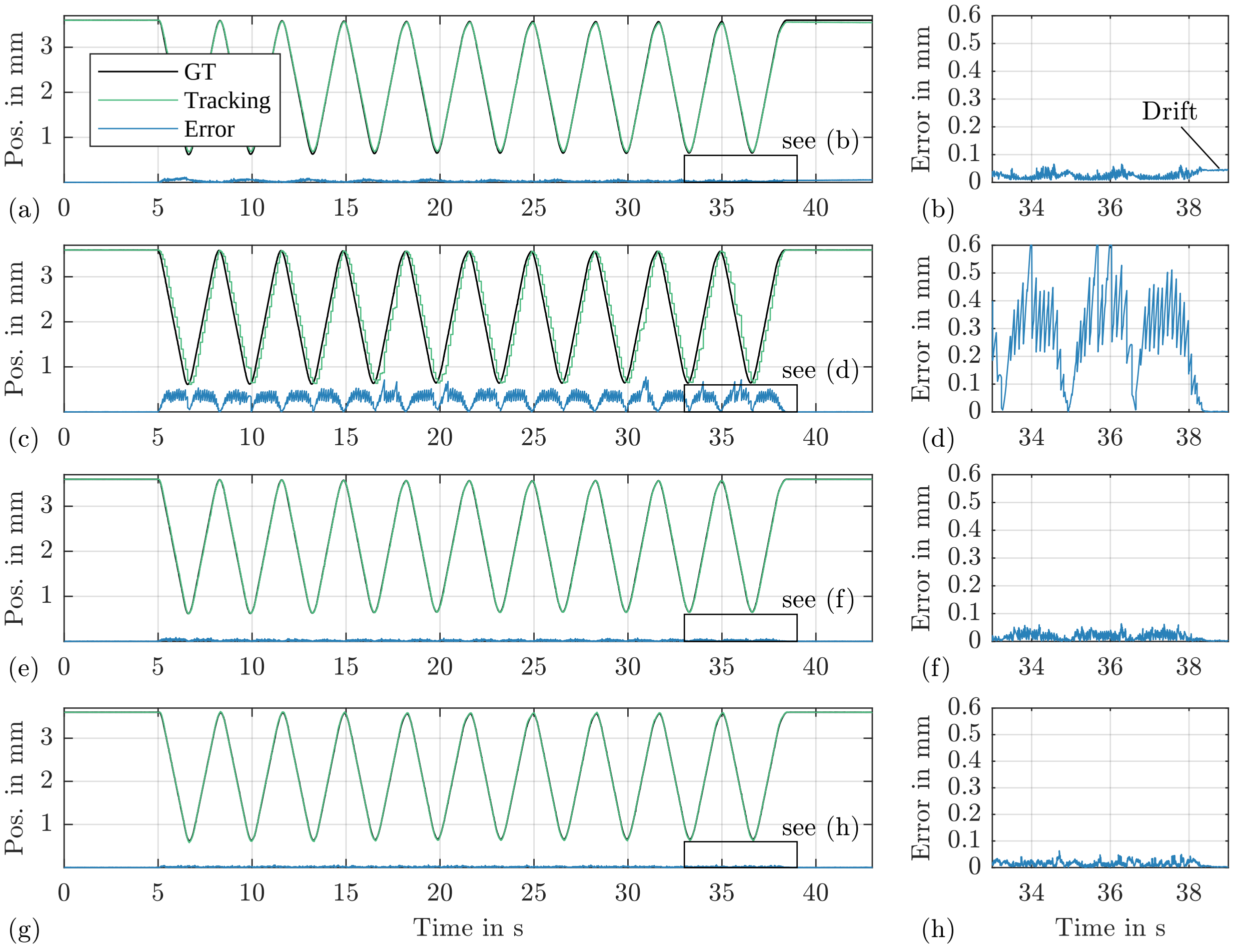


Figure 6.17: Results of tracking the MDF sample for the lateral motion pattern. The position and associated ground truth (GT) trajectory plotted over time (including magnified view) for methods (a,b) noMR, (c,d) sMR, (e,f) cMR, and (g,h) cMRKF.

Compared with the IVT validation, the IDS cameras enable highly accurate tracking due to not only the low-noise CMOS sensor but also the higher baseline-to-distance ratio. For instance, compared with the µRALP setting, which has a ratio of $1/20 = 0.05$, the IDS stereo setup provides a much higher depth resolution at a ratio of $37/60 = 0.62$, resulting in highly accurate tracking.

The computational load mainly depends on the number of mesh vertices and the size of the tracked region. Given an image area of $200 \times 200$ pixels, the associated computation time for tracking is listed in Table 6.5. In particular, for the PFN optimization scheme (Algorithm 6.2), iteration time drastically increases with the number of model parameters. The overhead of the affine-invariant fusion and the additional motion upsampling with less than a millisecond can be neglected. During the laser ablation trials discussed in the next section, a mesh with $6 \times 4$ triangles with an edge length of 75 pixels was chosen to estimate tissue motion. Consequently, the entire processing pipeline, including image rectification, cMRKF-based tracking, and laser ablation control, runs at a chosen image acquisition rate of 80 Hz.

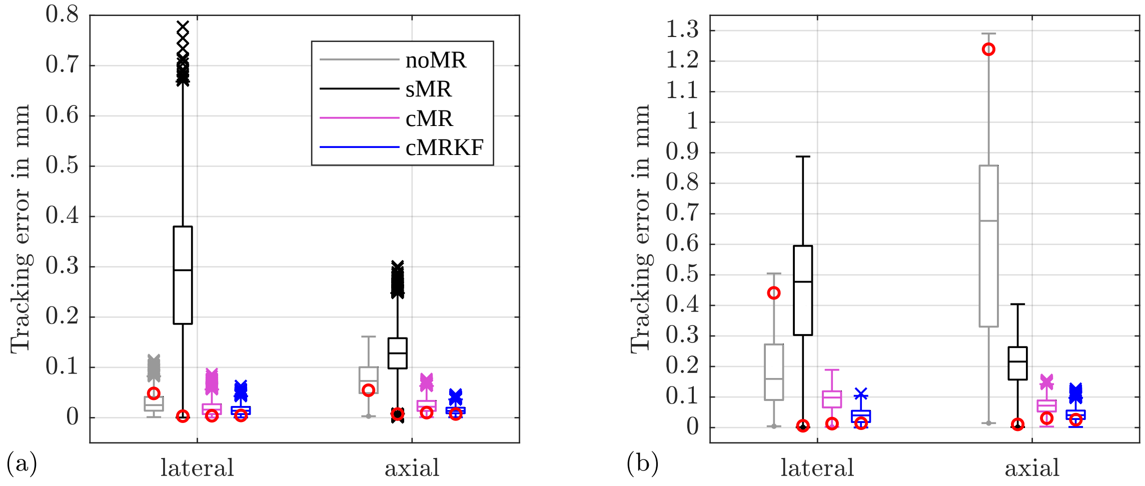


Figure 6.18: Box plot illustrating the tracking error measured for (a) the MDF specimen and (b) the porcine EVT sample. For each method, the results of the lateral and axial motion pattern are represented by the gray and black-colored boxplot, respectively. The final misalignment (drift) after returning to the hexapod home position is indicated by a red circle. Outliers are marked by the symbol ×.

Table 6.4: Tracking error in mm measured for the MDF and EVT samples moved by the hexapod robot. Lateral movements were performed at 2.1 mm/s and axial movements at 1 mm/s, respectively. **Blue** numbers represent the best performance while magenta numbers denote the second best.

| | | | noMR | sMR | cMR | cMRKF |
|---|---|---|---|---|---|---|
| MDF | Lateral | Mean ± SD | 0.029 ± 0.018 | 0.279 ± 0.143 | 0.019 ± 0.014 | **0.015 ± 0.010** |
| | | RMSE | 0.034 | 0.313 | 0.023 | **0.018** |
| | | Max. | 0.115 | 0.777 | 0.087 | **0.063** |
| | Axial | Mean ± SD | 0.075 ± 0.034 | 0.125 ± 0.050 | 0.024 ± 0.013 | **0.015 ± 0.008** |
| | | RMSE | 0.082 | 0.135 | 0.028 | **0.017** |
| | | Max. | 0.161 | 0.301 | 0.077 | **0.046** |
| EVT | Lateral | Mean ± SD | 0.189 ± 0.122 | 0.438 ± 0.202 | 0.091 ± 0.038 | **0.038 ± 0.021** |
| | | RMSE | 0.225 | 0.483 | 0.098 | **0.044** |
| | | Max. | 0.505 | 0.888 | 0.189 | **0.111** |
| | Axial | Mean ± SD | 0.626 ± 0.347 | 0.203 ± 0.082 | 0.071 ± 0.026 | **0.043 ± 0.020** |
| | | RMSE | 0.716 | 0.219 | 0.076 | **0.047** |
| | | Max. | 1.290 | 0.404 | 0.156 | **0.128** |

Table 6.5: Computation time in ms as a function of the horizontal triangle edge length when tracking an image region of 200 × 200 pixels. The Kalman filter prediction and update time is negligible.

| Triangle edge length | 25 | 50 | 75 | 100 |
|---|---|---|---|---|
| Mesh dimension | 16 × 12 | 8 × 6 | 6 × 4 | 4 × 3 |
| Tracking (Alg. 6.2) | 751.1 | 28.6 | 9.3 | 6.3 |
| Refinement (Alg. 6.3) | 112.9 | 38.7 | 20.5 | 16.7 |
| Tracking with sMR | 864.0 | 67.3 | 29.8 | 23.0 |
| Tracking with cMR(KF) | 751.8 | 28.9 | 9.6 | 6.5 |

### Laser Ablation Trials on MDF

The results of path tracing on the MDF specimen are listed in Table 6.6. Regarding the static scenario, the ablation misalignment is below $0.07$ mm, which correlates with the laser-to-camera registration error discussed in Section 4.1. In accordance to the cMRKF-based tracking error presented in Table 6.4, a slightly increased ablation error of $0.089$ mm and $0.084$mm is observed when the sample is moved in the lateral and axial direction, respectively, whereas the difference between the motion patterns is not significant. Regarding ablation of the curved incision line, as shown in Figure 6.19a, three snapshot images of acquired video sequence are depicted in Figure 6.19b, clearly illustrating the progressively ablated incision. Microscopic images of the straight and curved line, demonstrating precise path tracing, are provided in Figure 6.20a.

Table 6.6: Ablation accuracy (RMSE) in millimeters.

| Specimen | MDF | | | EVT | |
|---|---|---|---|---|---|
| Motion pattern | static | lateral | axial | lateral | axial |
| Straight line | 0.067 | 0.080 | 0.077 | 0.129 | 0.117 |
| Curved line | 0.068 | 0.089 | 0.084 | 0.206 | 0.173 |

### Laser Ablation Trials on EVT

The results of path tracing on porcine EVT are listed in Table 6.6. The associated snapshots of the lateral motion sequence are shown in Figure 6.19c–d. Compared with the MDF trials, the increase in the ablation error correlates with the larger tracking deviation, as listed in Table 6.4. In particular, for the lateral scenario, the path tracing error of $0.206$ mm is slightly higher than that for the axial motion ($0.173$ mm). Due to the inhomogeneous structure of soft tissue, heat exposure causes inevitable, anisotropic shrinking effects; thus, it can lead to distorted path tracing measurements. Therefore, this effect is assumed to be more distinct for the curved incision line. Nevertheless, microscopic examination of both shapes reveals high incision quality when cMRKF-based tracking is used for online laser control (see Figure 6.20b–c). To summarize, the ablation error is kept below $0.21$ mm regardless of its source, such as laser-to-camera registration, camera calibration, image-based tracking, scanning latency, and tissue shrinking effects. In comparison, deploying sMR leads to poor incision quality, as highlighted in Figure 6.20d. Due to significant delay of the motion estimate, the desired incision path is clearly fanned out; hence, it is only superficially ablated.

The benefit of vision-guided laser control is further demonstrated by comparing the tracking-based results in Figure 6.20c with laser ablation without focus adjustment when moving in the axial direction. As illustrated in Figure 6.20e, carbonization at the incision edges can be observed as a result of non-optimal energy exposure to the tissue. This may influence the desired incision

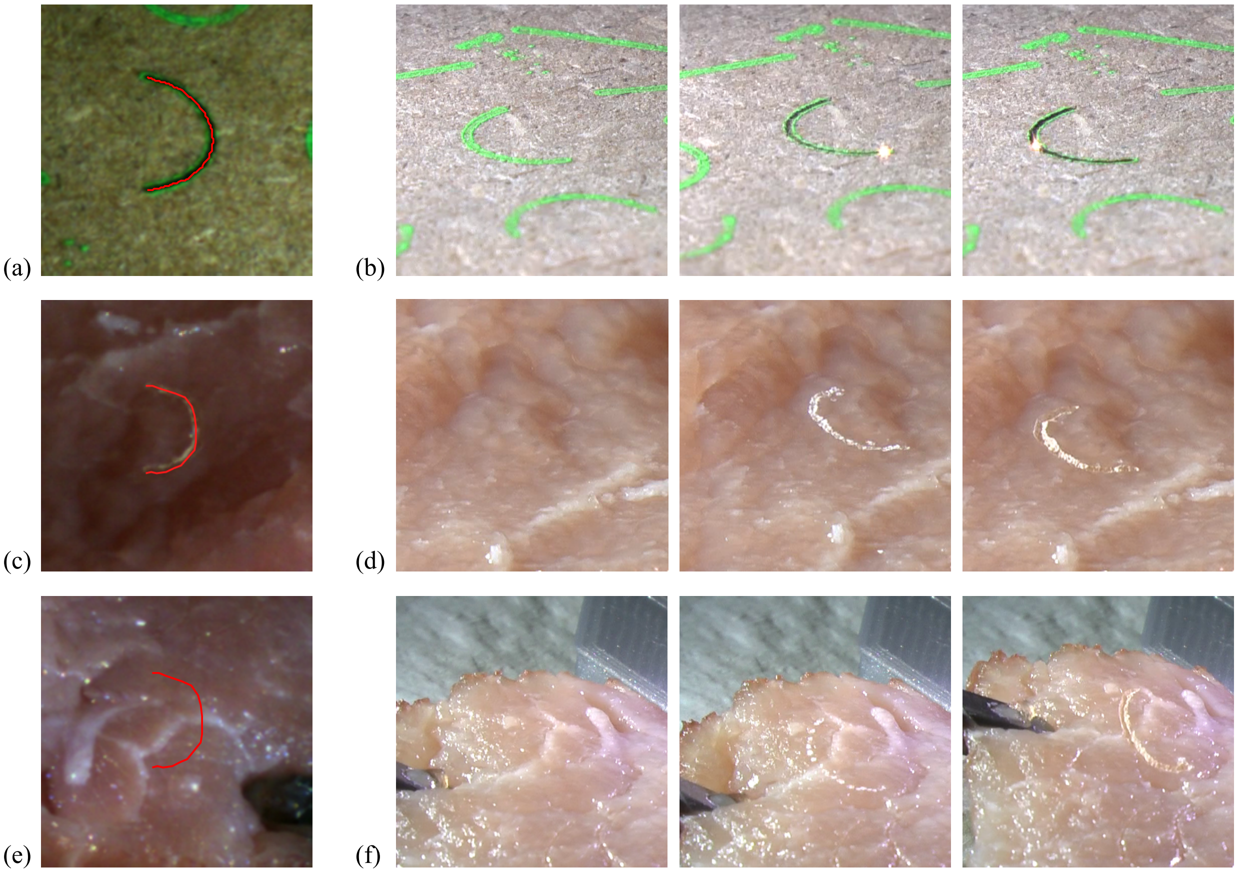


Figure 6.19: Laser ablation of the curved scan pattern. In the first row, results of ablation onto the MDF specimen while moving into lateral direction are shown in (a) the left camera view including segmented incision (red line) and (b) three snapshot images of the motion sequence acquired with an external, high-resolution video camera. In comparison, ablation on porcine EVT considering the same motion pattern is illustrated in the second row (c,d). Online laser path adaption on tissue manipulated with surgical grasping forceps is shown in the third row (e,f).

quality and lead to increased trauma and healing duration of the tissue. Thus, proper focusing by maintaining constant distance is mandatory for laser surgery in a soft tissue environment.

Finally, to demonstrate tracking in a clinically motivated scenario, tissue manipulation with surgical grasping forceps is performed simulating both motion in the depth direction and push-pull movements in the lateral direction (see Figure 6.19e–f). The microscopic examination clearly indicates that improved incision quality can be reproduced even on tissue undergoing deformation.

## 6.4 Conclusion

In this chapter, non-rigid tracking based on a linear, straightforward parametrization enabling left-right consistency for stereo vision has been presented. In contrast to computationally expensive, direct methods discussed in the literature, dense texture information is processed concurrently to

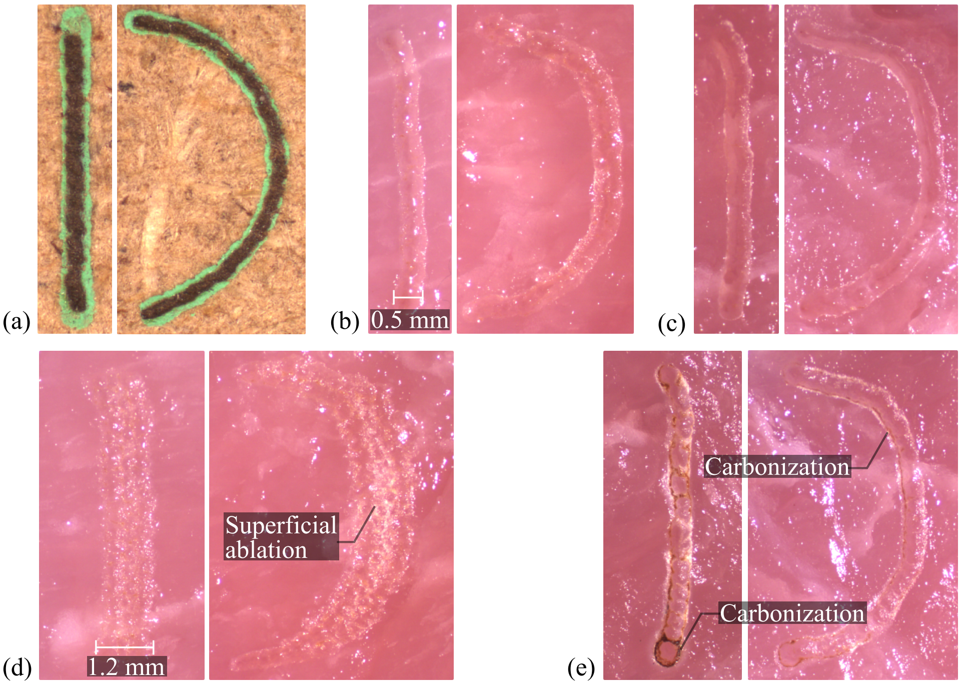


Figure 6.20: Qualitative results of the laser ablation trials on the MDF and the EVT sample under lateral (a,b) and axial movements (c) deploying concurrent tracking scheme cMRKF. In comparison with (b), sequential mesh refinement (sMR), as depicted in (d), leads to poor incision quality characterized by widened, superficial ablation due to the tracking latency. In particular, for axial movements, as shown in (c), slight carbonization occurs at the incision edges, as illustrated in (e), if the tracking-based adaptation of the laser focus is disabled.

correct tracking misalignment. Thus, highly accurate, online-capable motion estimation, which is a prerequisite for intraoperative assistance such as vision-guided ablation control in laser microsurgery, is enabled. Tracking robustness is enhanced by incorporating efficient outlier rejection in the robust estimator-based mesh refinement step.

Experimental results on *in vivo* data demonstrate enhanced accuracy compared with state-of-the-art approaches. Among the strategies discussed in this work, highest accuracy is achieved by concurrent tracking and mesh refinement as well as upsampling of the motion measurements. The entire image processing pipeline has been integrated into a control framework for laser microsurgery. The results reveal that tissue motion estimation is successfully integrated into the visual feedback loop, facilitating image stabilization during path definition and online adjustment of the ablation path.

Even though the parameter set is optimized in disparity space, only back-projection of the tracked mesh is required in order to map the motion estimate to task space and to enable laser positioning and focusing on the target surface. In general, control of other surgical or even robotic tools is conceivable. Future work might focus on the investigation of different online-capable feature detection and matching techniques extending this method and allowing for global retargeting of the tracked region after total occlusion or re-entering into the field of view.

# 7 Conclusion and Perspectives

Laser microsurgery is considered as the most advanced technique for contact-less and atraumatic removal of delicate pathological structures. In practice, function preservation of the organ usually conflicts with radicality of the dissection. Finding adequate resection margins requires imaging and visualization of the submucosal tumor extent as well as accurate laser focusing and positioning, i.e., in the presence of soft tissue motion. Addressing these challenges, this dissertation proposes a novel approach for stereo vision-guided laser microsurgery with focus on laryngeal application.

Primarily, a GPGPU-based algorithm for online and robust stereo matching is implemented. Compared with state-of-the-art methods, superior performance is achieved in particular for reconstruction accuracy and runtime. Special emphasis is given to the assessment of the developed method on different stereo imaging devices. Reconstruction errors below a millimeter for both endoscopic and microscopic vision clearly reveal the potential for application in microsurgery. In this regard, a methodology for laser-to-camera registration and laser view synthesis based on a trifocal model is derived for an Er:YAG laser. Practical evidence is given for online laser workspace highlighting on arbitrarily shaped surfaces. Quantitative results are discussed for distance-based laser focus adjustment yielding uniform ablation geometries in a depth range of several millimeters.

However, optimal ablation is exclusively achieved if the laser focal range is aligned to the vocal folds. This is addressed by color-encoding the laser depth of field within the surgical view for interactive focus positioning. The effectiveness of the overlay is proven by a user study demonstrating positioning errors in submillimeter range. Furthermore, surface information and laser-to-camera registration are integrated into an user interface enabling incision planning in the live camera view. Several path input strategies are evaluated. The point-based and the pen display approach perform most accurate and thus provide powerful strategies for prospective interfaces in laser surgery.

A further contribution aims at the fusion of OCT and stereo vision. A registration and a segmentation framework enable the detection and visualization of submucosal changes in laryngeal tissue. Practical demonstration is provided with a phantom mimicking the optical properties of laryngeal tissue, i.e., of the epithelium layers. Compared to conventional tomographic imaging, the use of OCT allows for intraoperative visualization of submucosal structures. Integrating OCT into the tablet-based planning framework, a color gradient ranging from red (function preservation) to green (maximum radicality) is proposed for decision support regarding resection margins.

Finally, laser ablation in a dynamic soft tissue environment is addressed by a non-rigid tracking scheme. The experimental results on *in vivo* data reveal superior performance compared with existing approaches. Among the strategies developed in this thesis, the highest accuracy is achieved for concurrent tracking and mesh refinement. An image-based closed loop control has been successfully implemented within the laser control enabling live view stabilization and motion compensation during laser ablation of soft tissue.

The achievements of this dissertation demonstrate the high potential of stereo vision-guided laser microsurgery even though further research is required. Future perspectives, as described in the following, should not solely aim at the improvement of the presented multimodal acquisition and processing framework, but in particular should take a profound clinical validation into account.

Regarding the stereo-based surface reconstruction, accuracy and density can be increased when considering a matching confidence that is high for correct disparities and is low for mismatches [HM12]. On the one hand, this allows refining erroneous or missing depth estimates. On the other hand, risk maps, that inversely correlate with the confidence measure, can be visualized to highlight areas where distance-based laser focus adjustment might perform less accurate. Another relevant aspect is automatic parametrization of the matching algorithm, i.e., the disparity range affecting the memory requirement and real-time performance. An estimate of the disparity range could be determined by establishing feature correspondences in the stereo view [SSPY10].

The developed motion estimation scheme performs accurately and robustly in a local proximity. However, further research is required for introducing global consistency of the tracking result allowing for reinitialization after tracking failure or loss, i.e., if the target region leaves and reenters the field of view. This will enable to recover and to track laryngeal pathologies segmented from a prior volumetric OCT scan. In this context, long term motion estimation is required that, e.g., could be addressed by dense tracking with quadrifocal constraints [CHDS14], or by online retargeting with a feature-based tracking-by-detection scheme [YGMY16]. Even though the latter approach would require an extension to stereo vision and online tracking of multiple targets, both methods could provide reinitialization of the deformation model presented in this thesis.

Beside an algorithmic optimization, translation of the presented framework to clinical scenarios requires further validation. Next to the integration of stereo vision, laser, and OCT into an endoscopic or microscopic prototype, the interaction of all three modalities has to be assessed thoroughly. If *in vivo* or *ex vivo* trials are not easily possible, at least a laryngeal tissue phantom is required that provides (1) a realistic texture information for stereo matching, (2) a subepithelial layer structure with embedded pathologies detectable via OCT, and (3) the possibility to perform laser ablation. Instead of using a silicone bulk structure, agar-based gel specimens (high water content of 98 %) could be incorporated into the phantom design process, as presented in this thesis, in order to attain the desired appearance and ablation characteristics [PFCM15].

Moreover, greater attention has to be given to laser ablation quality. In especially, finding the optimum of the pulse overlap that is defined by a ratio of scanning velocity, beam waist diameter, and pulse frequency which has been neglected in the presented studies. Consequently, determining a set of well-correlated parameters for soft tissue ablation needs further attention. In this context, histological examination indicating damage at the cellular level will provide clinical evidence.

# Bibliography

[ACBP$^+$13] M. Alicandri-Ciufelli, M. Bonali, A. Piccinini, L. Marra, A. Ghidini, E. M. Cunsolo, A. Maiorana, L. Presutti, and P. F. Conte. Surgical margins in head and neck squamous cell carcinoma: what is close? *European Archives of Oto-Rhino-Laryngology*, 270(10):2603–2609, 2013.

[ARV$^+$06] W. B. Armstrong, J. M. Ridgway, D. E. Vokes, S. Guo, J. Perez, R. P. Jackson, M. Gu, J. Su, R. L. Crumley, and T. Y. Shibuya. Optical coherence tomography of laryngeal cancer. *The Laryngoscope*, 116(7):1107–1113, 2006. `doi:10.1097/01.mlg.0000217539.27432.5a`.

[AT15] N. Andreff and B. Tamadazte. Laser steering using virtual trifocal visual servoing. *The International Journal of Robotics Research*, 35(6):672–694, 2015. `doi:10.1177/0278364915585585`.

[ATDH13] N. Andreff, B. Tamadazte, S. Dembélé, and Z. E. Hussnain. Preliminary variation on multiview geometry for vision-guided laser surgery. *Proc. Workshop on Multi-View Geometry in Robotics*, 2013.

[Ber14] J. Bergmeier. OCT-based augmented reality for depth measurements in laser ablation of biological tissue, Diploma Thesis, Institute of Mechatronic Systems, Leibniz Universität Hannover, Supervisor: A. Schoob. 2014.

[BF74] M. B. Brown and A. B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974. `doi:10.1080/01621459.1974.10482955`.

[BGBB$^+$11] F. Brunet, V. Gay-Bellile, A. Bartoli, N. Navab, and R. Malgouyres. Feature-driven direct non-rigid image registration. *International Journal of Computer Vision*, 93(1):33–52, 2011. `doi:10.1007/s11263-010-0407-x`.

[BGMI03] S. Baker, R. Gross, I. Matthews, and T. Ishikawa. Lucas-Kanade 20 years on: A unifying framework: Part 2. Technical Report CMU-RI-TR-03-01, Robotics Institute, Pittsburgh, PA, 2003.

[BJK$^+$13] A. Böttcher, N. Jowett, S. Kucher, R. Reimer, U. Schumacher, R. Knecht, W. Wöllmer, A. Münscher, and C. V. Dalchow. Use of a microsecond Er:YAG laser in laryngeal surgery reduces collateral thermal injury in comparison to superpulsed $CO_2$ laser. *European Archives of Oto-Rhino-Laryngology*, 271(5):1121–1128, 2013. `doi:10.1007/s00405-013-2761-0`.

[BKS$^+$15] J. Bergmeier, D. Kundrat, A. Schoob, L. A. Kahrs, and T. Ortmaier. Methods for a fusion of optical coherence tomography and stereo camera image data. *Proc. SPIE Medical Imaging*, 9415, 2015. `doi:10.1117/12.2082511`.

[BM92] 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.

[BM04] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[BM16] C. Barbalata and L. S. Mattos. Laryngeal tumor detection and classification in endoscopic video. *IEEE Journal of Biomedical and Health Informatics*, 20(1):322–332, 2016. `doi:10.1109/JBHI.2014.2374975`.

[Bou86] J.-L. Boulnois. Photophysical processes in recent medical laser developments: a review. *Lasers in Medical Science*, 1(1):47–66, 1986.

[Bou00] J.-Y. Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker. Technical report, Microprocessor Research Labs, Intel Corporation, 2000.

[Bou04] J. Y. Bouguet. The camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/, 2004.

[BPH11] E. Bogatyrenko, P. Pompey, and U. D. Hanebeck. Efficient physics-based tracking of heart surface motion for beating heart surgery robotic systems. *International Journal of Computer Assisted Radiology and Surgery*, 6(3):387–399, 2011.

[Bra00] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Bro96] J. Brooke. SUS - a quick and dirty usability scale. *Usability Evaluation in Industry*, 189:194, 1996.

[BRR11] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. *Proc. British Machine Vision Conference*, 11:1–11, 2011.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[BWD15] R. A. Buckmire, Y.-T. Wong, and A. M. Deal. The application of robotics to microlaryngeal laser surgery. *The Laryngoscope*, 125(6):1393–1400, 2015.

[BWRH07] T. Bader, A. Wiedemann, K. Roberts, and U. D. Hanebeck. Model-based motion estimation of elastic surfaces for minimally invasive cardiac surgery. *Proc. IEEE International Conference on Robotics and Automation*, pages 2261–2266, 2007.

[CAS17] Open-CAS collection of datasets for validating and benchmark computer-assisted surgery systems, http://opencas.webarchiv.kit.edu/, 2017. accessed 20 Apr 2017 (online).

[Cat74] E. E. Catmull. *A Subdivision Algorithm for Computer Display of Curved Surfaces.* PhD thesis, The University of Utah, 1974.

[CB12] T. Collins and A. Bartoli. Towards live monocular 3d laparoscopy using shading and specularity information. *Proc. International Conference on Information Processing in Computer-Assisted Interventions*, pages 11–21, 2012.

[CBBC16] T. Collins, A. Bartoli, N. Bourdel, and M. Canis. Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos. *Proc. International Conference on Medical Image Computing and Computer-Assisted Interventions*, pages 404–412, 2016.

[CCC$^+$08] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. Meshlab: an open-source mesh processing tool. *Proc. Eurographics Italian Chapter Conference*, pages 129–136, 2008.

[CGS$^+$15] D. P. Coelho, M. G. Garcia, A. Schoob, D. Kundrat, L. A. Kahrs, and T. Ortmaier. Design, evaluation and augmented reality visualization of a vocal fold tissue phantom for optical coherence tomography. *Proc. 14th Annual Conference of the German Society for Computer and Robot-Assisted Surgery*, (ISBN 978-3-00-050359-7), 2015.

[CHDS14] P.-L. Chang, A. Handa, A. J. Davison, and D. Stoyanov. Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. *Proc. International Conference on Information Processing in Computer-Assisted Interventions*, 2014.

[CIM$^+$13] M. Canis, F. Ihler, A. Martin, H. A. Wolff, C. Matthias, and W. Steiner. Organ preservation in t4a laryngeal cancer: is transoral laser microsurgery an option? *European Archives of Oto-Rhino-Laryngology*, 270(10):2719–2727, 2013.

[CKI$^+$04] M. Csanády, J. G. Kiss, L. Iván, J. Jóri, and J. Czigner. Ala (5-aminolevulinic acid)-induced protoporphyrin ix fluorescence in the endoscopic diagnostic and control of pharyngo-laryngeal cancer. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 261(5):262–266, 2004.

[CKS11] A. Curatolo, B. F. Kennedy, and D. D. Sampson. Structured three-dimensional optical phantom for optical coherence tomography. *Optics Express*, 19(20):19480–19485, 2011.

[CLZQ03] M. Chan, W. Lin, C. Zhou, and J. Y. Qu. Miniaturized three-dimensional endoscopic imaging system based on active stereovision. *Applied Optics*, 42(10):1888–1898, 2003.

[CMI$^+$14] M. Canis, A. Martin, F. Ihler, H. A Wolff, M. Kron, C. Matthias, and W. Steiner. Transoral laser microsurgery in treatment of pt2 and pt3 glottic laryngeal squamous cell carcinoma–results of 391 patients. *Head & Neck*, 36(6):859–866, 2014.

[COOS10] L. A. Crause, D. E. O'Donoghue, J. E. O'Connor, and F. Strümpfer. Use of a faro arm for optical alignment. *Proc. SPIE Modern Technologies in Space- and Ground-based Telescopes and Instrumentation*, 7739, 2010. `doi:10.1117/12.856810`.

[CSDE13] P.-L. Chang, D. Stoyanov, A. J. Davison, and P. Edwards. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. *Proc. International Conference on Medical Image Computing and Computer-Assisted Interventions*, 8149:42–49, 2013.

[CSG$^+$13] P. P. Caffier, B. Schmidt, M. Gross, K. Karnetzky, T. Nawka, A. Rotter, M. Seipelt, and B. Sedlmaier. A comparison of white light laryngostroboscopy versus autofluorescence endoscopy in the evaluation of vocal fold pathology. *The Laryngoscope*, 123(7):1729–1734, 2013.

[CSMH$^+$11] N. T. Clancy, D. Stoyanov, L. Maier-Hein, A. Groch, G.-Z. Yang, and D. S. Elson. Spectrally encoded fiber-based structured lighting probe for intraoperative 3d imaging. *Biomedical Optics Express*, 2(11):3119–3128, 2011.

[Dan99] G. Danuser. Photogrammetric calibration of a stereo light microscope. *Journal of Microscopy*, 193(1):62–83, 1999. `doi:10.1046/j.1365-2818.1999.00425.x`.

[DBR$^+$15] S. Donner, S. Bleeker, T. Ripken, M. Ptok, M. Jungheim, and A. Krueger. Automated working distance adjustment enables optical coherence tomography of the human larynx in awake patients. *Journal of Medical Imaging*, 2(2):026003–026003, 2015.

[DBW11] C. Dick, R. Burgkart, and R. Westermann. Distance visualization for interactive 3d implant planning. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2173–2182, 2011. `doi:10.1109/TVCG.2011.189`.

[DCA$^+$15] X. Du, N. Clancy, S. Arya, G. B. Hanna, J. Kelly, D. S. Elson, and D. Stoyanov. Robust surface tracking combining features, intensity and illumination compensation. *International Journal of Computer Assisted Radiology and Surgery*, 10(12):1915–1926, 2015. `doi:10.1007/s11548-015-1243-9`.

[DH81] R. Dutter and P. J. Huber. Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation*, 13(2):79–113, 1981.

[DKG$^+$13] J. Díaz Díaz, D. Kundrat, K.-F. Goh, O. Majdani, and T. Ortmaier. Towards intraoperative oct guidance for automatic head surgery: first experimental results. *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–354, 2013.

[DMC15a] G Dagnino, L. S. Mattos, and D. G. Caldwell. A vision-based system for fast and accurate laser scanning in robot-assisted phonomicrosurgery. *International*

*Journal of Computer Assisted Radiology and Surgery*, 10(2):217–229, 2015. `doi:10.1007/s11548-014-1078-9`.

[DMC15b] N. Deshpande, L. S. Mattos, and D. G. Caldwell. New motorized micromanipulator for robot-assisted laser phonomicrosurgery. *Proc. IEEE International Conference on Robotics and Automation*, pages 4755–4760, 2015.

[DMCM01] F. Devernay, F. Mourgues, and È. Coste-Manière. Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. *Proc. IEEE International Workshop on Medical Imaging and Augmented Reality*, pages 16–20, 2001.

[DOCM14] N. Deshpande, J. Ortiz, D. Caldwell, and L.S. Mattos. Enhanced computer-assisted laser microsurgeries with a virtual microscope based surgical system. *Proc. IEEE International Conference on Robotics and Automation*, pages 4194–4199, 2014.

[DSJG08] S. C. Desai, C.-K. Sung, D. W. Jang, and E. M. Genden. Transoral robotic surgery using a carbon dioxide flexible laser for tumors of the upper aerodigestive tract. *The Laryngoscope*, 118(12):2187–2189, 2008.

[DSN$^+$16] A. T. Day, P. Sinha, B. Nussenbaum, D. Kallogjeri, and B. H. Haughey. Management of primary T1–T4 glottic squamous cell carcinoma by transoral laser microsurgery. *The Laryngoscope*, 2016.

[ED04] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics*, 23(3):673–678, August 2004. `doi:10.1145/1015706.1015778`.

[FB81] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[FHM$^+$93] O. Faugeras, B. Hotz, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, and J. Vuillemin. Real time correlation-based stereo: algorithm, implementations and applications. Technical report, Inria, 1993.

[FMS$^+$16] R. Furukawa, H. Morinaga, Y. Sanomura, S. Tanaka, S. Yoshida, and H. Kawasaki. Shape acquisition and registration for 3d endoscope based on grid pattern projection. *Proc. European Conference on Computer Vision*, pages 399–415, 2016.

[FPB$^+$15] A. Fuchs, S. Pengel, J. Bergmeier, L. A. Kahrs, and T. Ortmaier. Fast and automatic depth control of iterative bone ablation based on optical coherence tomography data. *Proc. SPIE European Conferences on Biomedical Optics*, 2015.

[FPBB00] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski. Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy. *Neoplasia*, 2(1):9–25, 2000.

[FRGTA13] P. Filzmoser, A. Ruiz-Gazen, and C. Thomas-Agnan. Identification of local multivariate outliers. *Statistical Papers*, 55(1):29–47, 2013.

[FSD$^+$15] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[FSE$^+$13] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, 2013. http://globocan.iarc.fr (accessed on 04/01/2017).

[FSFLT$^+$13] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374–1403, 2013.

[FSK$^+$12] A. Fuchs, M. Schultz, A. Krüger, D. Kundrat, J. Diaz Diaz, and T. Ortmaier. Online measurement and evaluation of the er:yag laser ablation process using an integrated OCT system. In *Proc. Annual Conference of the German Society for Biomedical Engineering*, pages 434–437, January 2012. `doi:10.1515/bmt-2012-4231`.

[FTV00] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.

[FYSBY14] O. Ferhanoglu, M. Yildirim, K. Subramanian, and A. Ben-Yakar. A 5-mm piezo-scanning fiber device for high speed ultrafast laser microsurgery. *Biomedical Optics Express*, 5(7):2023–2036, 2014.

[Gia08] J. F. Giallo. *A medical robotic system for laser phonomicrosurgery*. PhD thesis, North Carolina State University, 2008.

[GLU12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[GRU10] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. *Proc. Asian Conference on Computer Vision*, pages 25–38, 2010.

[GVSY13] S. Giannarou, M. Visentini-Scarzanella, and G.-Z. Yang. Probabilistic tracking of affine-invariant anisotropic regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):130–143, 2013. `doi:10.1109/TPAMI.2012.81`.

[HB98] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[HBC$^+$09] M. Hashibe, P. Brennan, S. Chuang, S. Boccia, X. Castellsague, C. Chen, M. P. Curado, L. Dal Maso, A. W. Daudt, and E. Fabianova. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the inhance consortium. *Cancer epidemiology, Biomarkers & Prevention*, 18(2):541, 2009.

[HCHF93] R. C. Herdman, A. Charlton, A. E. Hinton, and A. J. Freemont. An in vitro comparison of the erbium:yag laser and the carbon dioxide laser in laryngeal surgery. *The Journal of Laryngology & Otology*, 107:908–911, 1993. `doi:10.1017/S0022215100124764`.

[HCL17] Hamlyn Centre Laparoscopic / Endoscopic Video Datasets, http://hamlyn.doc.ic.ac.uk/vision/, 2017. accessed 20th Apr 2017 (online).

[HCP$^+$15] N. Haouchine, S. Cotin, I. Peterlik, J. Dequidt, M. S. Lopez, E. Kerrien, and M.-O. Berger. Impact of soft tissue heterogeneity on augmented reality for liver surgery. *IEEE Transactions on Visualization and Computer Graphics*, 21(5):584–597, 2015.

[HFBG$^+$13] M. L. Hinni, A. Ferlito, M. S. Brandwein-Gensler, R. P. Takes, C. E. Silver, W. H. Westra, R. R. Seethala, J. P. Rodrigo, J. Corry, and C. R. Bradford. Surgical margins in head and neck cancer: a contemporary review. *Head & Neck*, 35(9):1362–1370, 2013.

[HHW14] R. J. Hendrick, S. D. Herrell, and R. J. Webster. A multi-arm hand-held robotic system for transurethral laser prostate surgery. *Proc. IEEE International Conference on Robotics and Automation*, pages 2850–2855, 2014.

[Hir05] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2:807–814, 2005.

[HKJK13] P. Heise, S. Klose, B. Jensen, and A. Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proc. IEEE International Conference on Computer Vision*, pages 2360–2367, 2013.

[HM12] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2121–2133, 2012.

[HPF$^+$12] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes. Reconstruction of a 3d surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. *Medical Image Analysis*, 16(3):597–611, 2012.

[HQ73] G. M. Hale and M. R. Querry. Optical constants of water in the 200-nm to 200-$\mu$m wavelength region. *Applied Optics*, 12(3):555–563, 1973.

[HRB+13] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2013.

[HS97] R. I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997.

[HS09] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009. `doi:10.1109/TPAMI.2008.221`.

[HSG+07] M. L. Hinni, J. R. Salassa, D. G. Grant, B. W. Pearson, R. E. Hayden, A. Martin, H. Christiansen, B. H. Haughey, B. Nussenbaum, and W. Steiner. Transoral laser microsurgery for advanced laryngeal cancer. *Archives of Otolaryngology-Head and Neck Surgery*, 133(12):1198–1204, 2007. `arXiv:/data/Journals/OTOL/11935/ooa70125_1198_1204.pdf`, `doi:10.1001/archotol.133.12.1198`.

[HSK+10] O. R. Hughes, N. Stone, M. Kraft, C. Arens, and M. A. Birchall. Optical and molecular techniques to identify tumor margins within the larynx. *Head & Neck*, 32(11):1544–1553, 2010.

[HSO07] M. Harris, S. Sengupta, and J. D. Owens. Parallel prefix sum (scan) with cuda. *GPU Gems*, 3(39):851–876, 2007.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[HZ04] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[IK88] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87 – 116, 1988. `doi:10.1016/S0734-189X(88)80033-1`.

[ITK17] Insight Segmentation and Registration Toolkit (ITK), https://itk.org/, 2017. accessed 7th May 2017 (online).

[Jak72] G. J. Jako. Laser surgery of the vocal cordsan experimental study with carbon dioxide lasers on dogs. *The Laryngoscope*, 82(12):2204–2216, 1972. `doi:10.1288/00005537-197212000-00009`.

[KBLA11] M. Kraft, C. S. Betz, A. Leunig, and C. Arens. Value of fluorescence endoscopy for the early diagnosis of laryngeal cancer and its precursor lesions. *Head & Neck*, 33(7):941–948, 2011.

[KDDF+14] J.-P. Kobler, J. Diaz Diaz, M. J. Fitzpatrick, J. G. Lexow, O. Majdani, and T. Ortmaier. Localization accuracy of sphere fiducials in computed tomography images. *Proc. SPIE Medical Imaging*, 9036, 2014. `doi:10.1117/12.2043472`.

[KFG+16] M. Kraft, K. Fostiropoulos, N. Gürtler, A. Arnoux, N. Davaris, and C. Arens. Value of narrow band imaging in the early diagnosis of laryngeal cancer. *Head & Neck*, 38(1):15–20, 2016.

[KFS+16] D. Kundrat, A. Fuchs, A. Schoob, L. A. Kahrs, and T. Ortmaier. Endoluminal non-contact soft tissue ablation using fiber-based Er: YAG laser delivery. *Proc. SPIE Optical Fibers and Sensors for Medical Diagnostics and Treatment Applications*, 9702, 2016.

[KGvG+08] M. Kraft, H. Glanz, S. von Gerlach, H. Wisweh, H. Lubatschowski, and C. Arens. Clinical value of optical coherence tomography in laryngology. *Head & Neck*, 30(12):1628–1635, 2008. `doi:10.1002/hed.20914`.

[Kle13] L. Kleingrothe. Stereo camera-based focus adaptation of a surgical laser system. Master's thesis, Institute of Mechatronic Systems, Leibniz Universität Hannover, Supervisors: A. Schoob and D. Kundrat, 2013.

[KMP+15] A. N. Kumar, M. I. Miga, T. S. Pheiffer, L. B. Chambless, R. C. Thompson, and B. M. Dawant. Persistent and automatic intraoperative 3d digitization of surfaces under dynamic magnifications of an operating microscope. *Medical Image Analysis*, 19(1):30–45, 2015.

[KOCD+12] M. Kersten-Oertel, S. J. Chen, S. Drouin, D. S. Sinclair, and D. L. Collins. Augmented reality visualization for guidance in neurovascular surgery. *Proc. Medicine Meets Virtual Reality*, pages 225–229, 2012.

[KRKS11] O. Kalentev, A. Rai, S. Kemnitz, and R. Schneider. Connected component labeling on a 2D grid using CUDA. *Journal of Parallel and Distributed Computing*, 71(4):615–620, 2011. `doi:10.1016/j.jpdc.2010.10.012`.

[KSKO15] D. Kundrat, A. Schoob, L. A. Kahrs, and T. Ortmaier. Flexible robot for laser phonomicrosurgery. *Soft Robotics*, pages 265–271, 2015. `doi:10.1007/978-3-662-44506-8_22`.

[KSMO13] D. Kundrat, A. Schoob, B. Munske, and T. Ortmaier. Towards an endoscopic device for laser-assisted phonomicrosurgery. *Proc. Hamlyn Symposium on Medical Robotics*, pages 55–56, 2013.

[Lav15] M.-H. Laves. Three-dimensional soft tissue motion tracking in laser surgery. Master's thesis, Institute of Mechatronic Systems, Leibniz Universität Hannover, Supervisor: A. Schoob, 2015.

[LC87] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987.

[Lek14] S. Lekon. Experimental studies and optimization of visualization and input concepts for surgical laser ablation. Master's thesis, Institute of Mechatronic Systems, Leibniz Universität Hannover, Supervisor: A. Schoob, 2014.

[Lew91] J. R. Lewis. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *ACM SIGCHI Bulletin*, pages 78–81, 1991. `doi:10.1145/122672.122692`.

[LKK$^+$12] G. Lamouche, B. F. Kennedy, K. M. Kennedy, C.-E. Bisaillon, A. Curatolo, G. Campbell, V. Pazos, and D. D. Sampson. Review of tissue simulating phantoms with controllable optical, mechanical and structural properties for use in optical coherence tomography. *Biomedical Optics Express*, 3(6):1381–1398, 2012.

[LKO$^+$13] E. Lankenau, M. Krug, S. Oelckers, N. Schrage, T. Just, and G. Hüttmann. iOCT with surgical microscopes: a new imaging during microsurgery. *Advanced Optical Technologies*, 2(3):233–239, 2013.

[LLP07] K. Lüerssen, H. Lubatschowski, and M. Ptok. Erbium:YAG-Laserchirurgie an Stimmlippengewebe. *HNO*, 55(6):443–446, 2007. `doi:10.1007/s00106-006-1479-3`.

[LMH$^+$16] S. Lang, S. Mattheis, P. Hasskamp, G. Lawson, C. Güldner, M. Mandapathil, P. Schuler, T. Hoffmann, M. Scheithauer, and M. Remacle. A european multicenter study evaluating the flex robotic system in transoral robotic surgery. *The Laryngoscope*, 127(2):391–395, 2016. `doi:10.1002/lary.26358`.

[LPK$^+$09] J. H. Lubin, M. Purdue, K. Kelsey, Z.-F. Zhang, D. Winn, Q. Wei, R. Talamini, N. Szeszenia-Dabrowska, E. M. Sturgis, and E. Smith. Total exposure and exposure rate effects for alcohol and smoking and risk of head and neck cancer: a pooled analysis of case-control studies. *American Journal of Epidemiology*, 170(8):937–947, 2009.

[LRC$^+$04] W. W. Lau, N. A. Ramey, J. J. Corso, N. V. Thakor, and G. D. Hager. Stereo-based endoscopic tracking of cardiac surface deformation. *Proc. International Conference on Medical Image Computing and Computer-Assisted Interventions*, 3217:494–501, 2004. `doi:10.1007/978-3-540-30136-3_61`.

[LRVdM92] Z.-Z. Li, L. Reinisch, and W. P. Van de Merwe. Bone ablation with Er:YAG and $CO_2$ laser: Study of thermal and acoustic effects. *Lasers in Surgery and Medicine*, 12(1):79–85, 1992. `doi:10.1002/lsm.1900120112`.

[LWFY12] B. Y. Leung, P. J. Webster, J. M. Fraser, and V. Yang. Real-time guidance of thermal and ultrashort pulsed laser ablation in hard tissue using inline coherent imaging. *Lasers in Surgery and Medicine*, 44(3):249–256, 2012.

[MADdM12] X. Maurice, C. Albitar, C. Doignon, and M. de Mathelin. A structured light-based laparoscope with real-time organs' surface reconstruction for minimally invasive surgery. *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5769–5772, 2012.

[MBC11] A. Malti, A. Bartoli, and T. Collins. Template-based conformal shape-from-motion from registered laparoscopic images. *Proc. Medical Image Understanding and Analysis*, 1(2):6, 2011.

[MDB$^+$14] L. S. Mattos, N. Deshpande, G. Barresi, L. Guastini, and G. Peretti. A novel computerized surgeon-machine interface for robot-assisted laser phonomicrosurgery. *The Laryngoscope*, 124(8):1887–1894, 2014. `doi:10.1002/lary.24566`.

[MDC11] L. S. Mattos, M. Dellepiane, and D. G. Caldwell. Next-generation micromanipulator for computer-assisted laser phonomicrosurgery. *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4555–4559, 2011.

[MDGA02] K. Malzahn, T. Dreyer, H. Glanz, and C. Arens. Autofluorescence endoscopy in the diagnosis of early laryngeal cancer and its precursor lesions. *The Laryngoscope*, 112(3):488–493, 2002.

[MDWF10] S. Marur, G. D'Souza, W. H. Westra, and Arlene A Forastiere. HPV-associated head and neck cancer: a virus-related cancer epidemic. *The Lancet Oncology*, 11(8):781–789, 2010.

[MGLvE$^+$13] T. Meyer, O. Guntinas-Lichius, F. von Eggeling, G. Ernst, D. Akimov, M. Schmitt, B. Dietzek, and J. Popp. Multimodal nonlinear microscopic investigations on head and neck squamous cell carcinoma: Toward intraoperative imaging. *Head & Neck*, 35(9):E280–E287, 2013.

[MHGB$^+$14] L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N.T. Clancy, D.S. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Muller-Stich, D. Oladokun, S. Rohl, T.R. dos Santos, H.-P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, and D. Stoyanov. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE Transactions on Medical Imaging*, 33(10):1913–1930, 2014. `doi:10.1109/TMI.2014.2325607`.

[MIH11] D. J. Mirota, M. Ishii, and G. D. Hager. Vision-based navigation in image-guided interventions. *Annual Review of Biomedical Engineering*, 13:297–319, 2011.

[MMS$^+$11] S. Mersmann, M. Müller, A. Seitel, F. Arnegger, R. Tetzlaff, J. Dinkel, M. Baumhauer, B. Schmied, H.-P. Meinzer, and L. Maier-Hein. Time-of-flight camera technique for augmented reality in computer-assisted interventions. *Proc. SPIE Medical Imaging*, 7964, 2011.

[MNSU13] K. Mishima, A. Nakano, R. Shiraishi, and Y. Ueyama. Range image of the velopharynx produced using a 3-d endoscope with pattern projection. *The Laryngoscope*, 123(12):E122–E126, 2013.

[MTL78] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[MY08] P. Mountney and G.-Z. Yang. Soft tissue tracking for minimally invasive surgery: Learning local deformation online. *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 5242:364–372, 2008.

[Nat13] United Nations. World population prospects the 2012 revision – highlights and advance tables. *United Nations – Department of Economic and Social Affairs – Population Division*, ESA/P/WP.228, 2013.

[NHH$^+$16] M. Neitsch, I.-S. Horn, M. Hofer, A. Dietz, and M. Fischer. Integrated multipoint-laser endoscopic airway measurements by transoral approach. *BioMed Research International*, 2016, 2016.

[NHX$^+$11] X. G. Ni, S. He, Z. G. Xu, L. Gao, N. Lu, Z. Yuan, S. Q. Lai, Y. M. Zhang, J. L. Yi, and X. L. Wang. Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. *The Journal of Laryngology & Otology*, 125(03):288–296, 2011.

[Nie13] M. H. Niemz. *Laser-tissue interactions: fundamentals and applications*. Springer Science & Business Media, 2013.

[NTP07] A. Noce, J. Triboulet, and P. Poignet. Efficient tracking of the heart using texture. *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4480–4483, 2007.

[Nvi17] Nvidia. Cuda toolkit documentation v8.0.61. *Santa Clara (CA, USA): Nvidia Corporation*, 2017.

[OEM$^+$12] M. A. Ororke, M. V. Ellison, L. J. Murray, M. Moran, J. James, and L. A. Anderson. Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis. *Oral Oncology*, 48(12):1191–1201, 2012.

[OGB$^+$05] T. Ortmaier, M. Gröger, D. H. Boehm, V. F., and G. Hirzinger. Motion estimation in beating heart surgery. *IEEE Transactions on Biomedical Engineering*, 52(10):1729–1740, 2005. `doi:10.1109/TBME.2005.855716`.

[OK85] Y. Ohta and T. Kanade. Stereo by intra-and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):139–154, 1985.

[ORJ$^+$14] V. Oswal, M. Remacle, S. Jovanvic, S. M. Zeitels, J. P. Krespi, and C. Hopper. *Principles and practice of lasers in otorhinolaryngology and head and neck surgery*. Kugler Publications, 2014.

[PB16] K. Prokopetc and A. Bartoli. A comparative study of transformation models for the sequential mosaicing of long retinal sequences of slit-lamp images obtained in a closed-loop motion. *International Journal of Computer Assisted Radiology and Surgery*, 11(12):2163–2172, 2016.

[PCDB$^+$10] C. Piazza, D. Cocco, L. De Benedetto, F. Del Bon, P. Nicolai, and G. Peretti. Narrow band imaging and high definition television in the assessment of laryngeal cancer: a prospective study on 279 patients. *European Archives of Oto-rhino-laryngology*, 267(3):409–414, 2010.

[PDLA$^+$14] F. Preiswerk, V. De Luca, P. Arnold, Z. Celicanin, L. Petrusca, C. Tanner, O. Bieri, R. Salomir, and P. C. Cattin. Model-guided respiratory organ motion prediction of the liver from 2d ultrasound. *Medical Image Analysis*, 18(5):740–751, 2014.

[Per14] G. Peretti. Laryngeal Dataset of the µRALP Project, Department of Otorhinolaryngology, University of Genoa, Italy, 2014.

[PFCM15] D. Pardo, L. Fichera, D. Caldwell, and L. S. Mattos. Learning temperature dynamics on agar-based phantom tissue surface during single point $CO_2$. *Neural Processing Letters*, 42(1):55–70, 2015.

[PFJ05] P. Paul, O. Fleig, and P. Jannin. Augmented virtuality based on stereoscopic reconstruction in multimodal image-guided neurosurgery: Methods and performance evaluation. *IEEE Transactions on Medical Imaging*, 24(11):1500–1511, 2005.

[PHS$^+$09] J. Penne, K. Höller, M. Stürmer, T. Schrauder, A. Schneider, R. Engelbrecht, H. Feußner, B. Schmauss, and J. Hornegger. Time-of-flight 3-d endoscopy. *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–474, 2009.

[PKW$^+$16] M. Peller, A. Katalinic, B. Wollenberg, I. U. Teudt, and J.-E. Meyer. Epidemiology of laryngeal carcinoma in germany, 1998–2011. *European Archives of Oto-Rhino-Laryngology*, 273(6):1481–1487, 2016.

[PLF08] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *International Journal of Computer Vision*, 76(2):109–122, 2008. `doi:10.1007/s11263-006-0017-9`.

[PN12] C.D. Pantilie and S. Nedevschi. Optimizing the census transform on CUDA enabled GPUs. *Proc. IEEE International Conference on Intelligent Computer Communication and Processing*, pages 201–207, 2012. `doi:10.1109/ICCP.2012.6356186`.

[Pod07] V. Podlozhnyuk. Image convolution with CUDA. *Nvidia Corporation White Paper*, 2097(3), 2007.

[Pod13] F. Podszus. Implementation of a real-time capable and robust disparity map computation for medical stereo-imaging, Master Thesis, Institute of Mechatronic Systems, Leibniz Universität Hannover, Supervisor: A. Schoob. 2013.

[POM$^+$16] V. Penza, J. Ortiz, L. S. Mattos, A. Forgione, and E. De Momi. Dense soft tissue 3d reconstruction refined with super-pixel segmentation for robotic abdominal surgery. *International Journal of Computer Assisted Radiology and Surgery*, 11(2):197–206, 2016.

[PPP$^+$16] G. Peretti, C. Piazza, S. Penco, G. Santori, F. Del Bon, S. Garofolo, A. Paderno, L. Guastini, and P. Nicolai. Transoral laser microsurgery as primary treatment for selected T3 glottic and supraglottic cancers. *Head & Neck*, 38(7):1107–1112, 2016.

[PRK$^+$12] S. Patel, M. Rajadhyaksha, S. Kirov, Y. Li, and R. Toledo-Crow. Endoscopic laser scalpel for head and neck cancer surgery. *Proc. SPIE Photonic Therapeutics and Diagnostics*, 8207, 2012. `doi:10.1117/12.909172`.

[PRRA16] E. Pengwang, K. Rabenorosoa, M. Rakotondrabe, and N. Andreff. Scanning micromirror platform based on mems technology for medical application. *Micromachines*, 7(2):24, 2016. URL: `http://www.mdpi.com/2072-666X/7/2/24`, `doi:10.3390/mi7020024`.

[PSA$^+$04] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics*, 23(3):664–672, 2004. `doi:10.1145/1015706.1015777`.

[PSM13] G.A. Puerto-Souza and G.-L. Mariottini. A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images. *IEEE Transactions on Medical Imaging*, 32(7):1201–1214, 2013. `doi:10.1109/TMI.2013.2239306`.

[QGC$^+$09] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. ROS: an open-source robot operating system. *IEEE International Conference on Robotics and Automation - Workshop on Open Source Robotics*, 2009.

[RA11] M. Rubinstein and W. B. Armstrong. Transoral laser microsurgery for laryngeal cancer: A primer and review of laser dosimetry. *Lasers in Medical Science*, 26(1):113–124, 2011. `doi:10.1007/s10103-010-0834-5`.

[RBB09] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. *Proc. IEEE Conference on Robotics and Automation*, pages 3212–3217, 2009.

[RBMB08] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3384–3391, 2008.

[RBS$^+$12] S. Röhl, S. Bodenstedt, S. Suwelack, H. Kenngott, B. P. Muller-Stich, R. Dillmann, and S. Speidel. Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Medical Physics*, 39:1632, 2012. `doi:10.1118/1.3681017`.

[RC11] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). *Proc. IEEE International Conference on Robotics and Automation*, pages 1–4, 2011.

[RCL$^+$04] N. A. Ramey, J. J. Corso, W. W. Lau, D. Burschka, and G. D. Hager. Real-time 3d surface tracking and its applications. *Workshop IEEE International Conference on Computer Vision and Pattern Recognition*, pages 34–34, 2004.

[RDM15] S. Russo, P. Dario, and A. Menciassi. A novel robotic platform for laser-assisted transurethral surgery of the prostate. *IEEE Transactions on Biomedical Engineering*, 62(2):489–500, 2015.

[RHC$^+$05] M. Remacle, F. Hassan, D. Cohen, G. Lawson, and M. Delos. New computer-guided scanner for improving $CO_2$ laser-assisted microincision. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 262(2):113–119, 2005. URL: `http://dx.doi.org/10.1007/s00405-004-0746-8`, `doi:10.1007/s00405-004-0746-8`.

[RKD$^+$17] L. Royer, A. Krupa, G. Dardenne, A. Le Bras, É. Marchand, and M. Marchal. Real-time target tracking of soft tissues in 3d ultrasound images based on robust visual information and mechanical simulation. *Medical Image Analysis*, 35:582–598, 2017.

[RLND08] M. Remacle, G. Lawson, M.-C. Nollevaux, and M. Delos. Current state of scanning micromanipulator applications with the carbon dioxide laser. *Annals of Otology, Rhinology & Laryngology*, 117(4):239–244, 2008.

[ROS] The Robot Operating System (ROS), http://www.ros.org/. accessed 9th May 2017 (online).

[RPL10] R. Richa, P. Poignet, and C. Liu. Three-dimensional motion tracking for beating heart surgery using a thin-plate spline deformable model. *The International Journal of Robotics Research*, 29(2-3):218–230, 2010. `doi:10.1177/0278364909356600`.

[RS08] C. A. Rosen and B. Simpson. *Operative techniques in laryngology*. Springer Science & Business Media, 2008.

[RSH06] T. Ropinski, F. Steinicke, and K. Hinrichs. Visually supporting depth perception in angiography imaging. *Smart Graphics*, 4073:93–104, 2006. `doi:10.1007/11795018_9`.

[RSJZ$^+$12] C. M. Rivera-Serrano, P. Johnson, B. Zubiate, R. Kuenzler, H. Choset, M. Zenati, S. Tully, and U. Duvvuri. A transoral highly flexible robot. *The Laryngoscope*, 122(5):1067–1071, 2012.

[RTR$^+$16] R. Renevier, B. Tamadazte, K. Rabenorosoa, L. Tavernier, and N. Andreff. Endoscopic laser surgery: Design, modeling and control. *IEEE/ASME Transactions on Mechatronics*, 22(1):99–106, 2016. `doi:10.1109/TMECH.2016.2595625`.

[SA00] W. Steiner and P. Ambrosch. *Endoscopic laser surgery of the upper aerodigestive tract: with special emphasis on cancer surgery*. Thieme, 2000.

[Sam81] H. Samet. Connected components labeling using quadtrees. *Journal of the ACM*, 28(3):487–501, 1981.

[SCJ$^{+}$11] R. T. Sataloff, F. Chowdhury, S. Joglekar, M. Hawkshaw, and J. Portnoy. Atlas of endoscopic laryngeal surgery. *Cricoarytenoid and Cricothyroid Joint Injury: Evaluation and Treatment*, pages 249–57, 2011.

[SDP13] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013. `doi:10.1109/TMI.2013.2265603`.

[SDY05] D. Stoyanov, A. Darzi, and G.-Z. Yang. A practical approach towards accurate dense 3d depth recovery for robotic laparoscopic surgery. *Computer Aided Surgery*, 10(4):199–208, 2005. `arXiv:http://www.tandfonline.com/doi/pdf/10.3109/10929080500230379`, `doi:10.3109/10929080500230379`.

[SFSA12] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou. An endoscopic 3d scanner based on structured light. *Medical Image Analysis*, 16(5):1063–1072, 2012. `doi:10.1016/j.media.2012.04.001`.

[SGS04] H.W. Schreier, D. Garcia, and M.A. Sutton. Advances in light microscope stereo vision. *Experimental Mechanics*, 44(3):278–288, 2004. `doi:10.1007/BF02427894`.

[Shi94] J. Shi. Good features to track. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[SHK$^{+}$14] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. *Proc. German Conference on Pattern Recognition*, pages 31–42, 2014.

[SKG$^{+}$14] T. Sandner, S. Kimme, T. Grasshoff, U. Todt, A. Graf, C. Tulea, A. Lenenbach, and H. Schenk. Micro-scanning mirrors for high-power laser applications in laser surgery. *Proc. SPIE MOEMS and Miniaturized Systems*, 8977, 2014.

[SKK$^{+}$15] A. Schoob, D. Kundrat, L. Kleingrothe, L. A. Kahrs, N. Andreff, and T. Ortmaier. Tissue surface information for intraoperative incision planning and focus adjustment in laser surgery. *International Journal of Computer Assisted Radiology and Surgery*, 10(2):171–181, 2015. `doi:10.1007/s11548-014-1077-x`.

[SKKO16] A. Schoob, D. Kundrat, L. A. Kahrs, and T. Ortmaier. Comparative study on surface reconstruction accuracy of stereo imaging devices for microsurgery. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):145–156, 2016. `doi:10.1007/s11548-015-1240-z`.

[SKKO17] A. Schoob, D. Kundrat, L. A. Kahrs, and T. Ortmaier. Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery. *Medical Image Analysis*, 40:80–95, 2017. `doi:10.1016/j.media.2017.06.004`.

[SKL$^+$16] A. Schoob, D. Kundrat, S. Lekon, L. A. Kahrs, and T. Ortmaier. Color-encoded distance for interactive focus positioning in laser microsurgery. *Optics and Lasers in Engineering*, 83:71–79, 2016. `doi:10.1016/j.optlaseng.2016.03.002`.

[SLK$^+$14] A. Schoob, S. Lekon, D. Kundrat, L. A. Kahrs, and T. Ortmaier. Interactive focus positioning in laser surgery - a preliminary study. *Proc. 13th Annual Conference of the German Society for Computer and Robotic Assisted Surgery*, (ISBN 978-3-00-047154-4), 2014.

[SLK$^+$15] A. Schoob, S. Lekon, D. Kundrat, L. A. Kahrs, L. S. Mattos, and T. Ortmaier. Comparison of tablet-based strategies for incision planning in laser microsurgery. *Proc. SPIE Medical Imaging*, 9415, 2015. `doi:10.1117/12.2081032`.

[SLKO16] A. Schoob, M.-H. Laves, L. A. Kahrs, and T. Ortmaier. Soft tissue motion tracking with application to tablet-based incision planning in laser surgery. *International Journal of Computer Assisted Radiology and Surgery*, 11(12):2325–2337, 2016. `doi:10.1007/s11548-016-1420-5`.

[SMD$^+$05] D. Stoyanov, G. P. Mylonas, F. Deligianni, A. Darzi, and G.-Z. Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 139–146, 2005. `doi:10.1007/11566489_18`.

[SPK$^+$13] A. Schoob, F. Podszus, D. Kundrat, L. A. Kahrs, and T. Ortmaier. Stereoscopic surface reconstruction in minimally invasive surgery using efficient non-parametric image transforms. *Proc. 3rd Joint Workshop on New Technologies for Computer/Robot Assisted Surgery*, pages 26–29, 2013.

[SPP$^+$10] J. Süßmuth, W.-D. Protogerakis, A. Piazza, F. Enders, R. Naraghi, G. Greiner, and P. Hastreiter. Color-encoded distance visualization of cranial nerve-vessel contacts. *International Journal of Computer Assisted Radiology and Surgery*, 5(6):647–654, 2010. `doi:10.1007/s11548-010-0410-2`.

[SPT$^+$06] M. Sauvée, P. Poignet, J. Triboulet, E. Dombre, E. Malis, and R. Demaria. 3d heart motion estimation using endoscopic monocular vision system. *Modeling and Control in Biomedical Systems*, 6(1):141–146, 2006. `doi:10.3182/20060920-3-FR-2912.00029`.

[SRB$^+$14] S. Suwelack, S. Röhl, S. Bodenstedt, D. Reichard, R. Dillmann, T. dos Santos, L. Maier-Hein, M. Wagner, J. Wünscher, and H. Kenngott. Physics-based shape matching for intraoperative image guidance. *Medical Physics*, 41(11), 2014.

[SRH12] D. Stoyanov, A. Rayshubskiy, and E. Hillman. Robust registration of multispectral images of the cortical surface in neurosurgery. *Proc. IEEE 9th International Symposium on Biomedical Imaging*, pages 1643–1646, 2012.

[SRS$^+$12] C. Suárez, J. P. Rodrigo, C. E. Silver, D. M. Hartl, R. P. Takes, A. Rinaldo, P. Strojan, and A. Ferlito. Laser surgery for early to moderately advanced glottic, supraglottic, and hypopharyngeal cancers. *Head & Neck*, 34(7):1028–1035, 2012.

[SS02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002. `doi:10.1023/A:1014573219977`.

[SS07] C. A. Solares and M. Strome. Transoral robot-assisted $CO_2$ laser supraglottic laryngectomy: Experimental and clinical data. *The Laryngoscope*, 117(5):817–820, 2007.

[SSPY10] D. Stoyanov, M. Scarzanella, P. Pratt, and G.-Z. Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. *Proc. Medical Image Computing and Computer-Assisted Interventions*, 6361:275–282, 2010. `doi:10.1007/978-3-642-15705-9{\_}34`.

[STA15] J.-A. Seon, B. Tamadazte, and N. Andreff. Decoupling path following and velocity profile in vision-guided laser steering. *IEEE Transactions on Robotics*, 31(2):280–289, 2015.

[Ste87] R. A. Steenblik. The chromostereoscopic process: A novel single image stereoscopic process. *Proc. SPIE True Three-Dimensional Imaging Techniques & Display Technologies*, 0761:27–34, 1987. `doi:10.1117/12.940117`.

[STL15] B. Su, J. Tang, and H. Liao. Automatic laser ablation control algorithm for an novel endoscopic laser ablation end effector for precision neurosurgery. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4362–4367, 2015.

[Str75] M. S. Strong. Laser excision of carcinoma of the larynx. *The Laryngoscope*, 85(8):1286–1289, 1975. `doi:10.1288/00005537-197508000-00003`.

[SW14] B. Stewart and C. P. Wild. World cancer report 2014. *International Agency for Research on Cancer, World Health Organization*, 2014.

[SY07] D. Stoyanov and G.-Z. Yang. Stabilization of image motion for robotic assisted beating heart surgery. *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 4791:417–424, 2007. `doi:10.1007/978-3-540-75757-3_51`.

[SY09] D. Stoyanov and G.-Z. Yang. Soft tissue deformation tracking for robotic assisted minimally invasive surgery. *Proc. IEEE International Conference Engineering in Medicine and Biology Society*, pages 254–257, 2009. `doi:10.1109/IEMBS.2009.5334010`.

[SZS03] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003.

[TBV$^+$06] H.-W Tang, Hendrik Brussel, Van, J. V. Sloten, D. Reynaerts, G. De Win, B. V. Cleynenbreugel, and P. R. Koninckx. Evaluation of an intuitive writing interface in robot-aided laser laparoscopic surgery. *Computer Aided Surgery*, 11(1):21–30, 2006. `arXiv:http://informahealthcare.com/doi/pdf/10.3109/10929080500450886`, `doi:10.3109/10929080500450886`.

[THNI14] D. J. Tan, S. Holzer, N. Navab, and S. Ilic. Deformable template tracking in 1ms. *Proc. British Machine Vision Conference*, 2014.

[TL89] 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.

[TPT$^+$13] T.-H. Tsai, B. Potsaid, Y. K. Tao, V. Jayaraman, J. Jiang, P. J. Heim, M. F. Kraus, C. Zhou, J. Hornegger, and H. Mashimo. Ultrahigh speed endoscopic optical coherence tomography using micromotor imaging catheter and vcsel technology. *Biomedical Optics Express*, 4(7):1119–1132, 2013.

[Trä12] F. Träger. *Springer handbook of lasers and optics*. Springer Science & Business Media, 2012.

[TVBR$^+$03] H.-W. Tang, H. Van Brussel, D. Reynaerts, J. Vander Sloten, and P. R. Koninckx. A laparoscopic robot with intuitive interface for gynecological laser laparoscopy. *Proc. IEEE International Conference on Robotics and Automation*, 2:2646–2650, 2003.

[uRA15] The µRALP Project, http://www.microralp.eu, 2012–2015. accessed 20th Apr 2017 (online).

[VBB$^+$16] I. Vilaseca, J. L. Blanch, J. Berenguer, J. J. Grau, E. Verger, Á. Muxí, and M. Bernal-Sprekelsen. Transoral laser microsurgery for locally advanced (T3–T4a) supraglottic squamous cell carcinoma: Sixteen years of experience. *Head & Neck*, 38(7):1050–1057, 2016.

[VBG08] T. Vaudrey, H. Badino, and S. Gehrig. Integrating disparity images by incorporating disparity rate. *International Workshop on Robot Vision*, pages 29–42, 2008.

[Vek03] O. Veksler. Fast variable window for stereo correspondence using integral images. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1:556–561, 2003.

[VSJ78] Charles W. Vaughan, M.Stuart Strong, and Geza J. Jako. Laryngeal carcinoma: Transoral treatment utilizing the $CO_2$ laser. *The American Journal of Surgery*, 136(4):490–493, 1978. `doi:10.1016/0002-9610(78)90267-2`.

[VSSY12] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang. Metric depth recovery from monocular images using shape-from-shading and specularities. *Proc. IEEE International Conference on Image Processing*, pages 25–28, 2012.

[VTK17] The Visualization Toolkit (VTK), http://www.vtk.org/, 2017. accessed 7th May 2017 (online).

[WDK+14] W. Wieser, W. Draxinger, T. Klein, S. Karpf, T. Pfeiffer, and R. Huber. High definition live 3D-OCT in vivo: design and evaluation of a 4D OCT engine with 1 GVoxel/s. *Biomedical Optics Express*, 5(9):2963–2977, 2014.

[Weg60] P. Wegner. A technique for counting ones in a binary computer. *Communications of the ACM*, 3(5):322, 1960. URL: `http://doi.acm.org/10.1145/367236.367286`, `doi:10.1145/367236.367286`.

[WFD89] J. T. Walsh, T. J. Flotte, and T. F. Deutsch. Er:YAG laser ablation of tissue: Effect of pulse duration and tissue type on thermal damage. *Lasers in Surgery and Medicine*, 9(4):314–326, 1989. `doi:10.1002/lsm.1900090403`.

[WFW+12] J. Wang, P. Fallavollita, L. Wang, M. Kreiser, and N. Navab. Augmented reality during angiography: Integration of a virtual mirror for improved 2d/3d visualization. *Proc. IEEE International Symposium on Mixed and Augmented Reality*, pages 257–264, 2012. `doi:10.1109/ISMAR.2012.6402565`.

[WJG+05] B. J. Wong, R. P. Jackson, S. Guo, J. M. Ridgway, U. Mahmood, J. Su, T. Y. Shibuya, R. L. Crumley, M. Gu, and W. B. Armstrong. In vivo optical coherence tomography of the human larynx: normative and benign pathology in 82 patients. *The Laryngoscope*, 115(11):1904–1911, 2005.

[WNJ10] C. Wu, S. G. Narasimhan, and B. Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86(2-3):211–228, 2010.

[WOM+12] Gregory S Weinstein, Bert W O'Malley, J Scott Magnuson, William R Carroll, Kerry D Olsen, Lixia Daio, Eric J Moore, and F Christopher Holsinger. Transoral robotic surgery: a multicenter study to assess feasibility, safety, and surgical margins. *The Laryngoscope*, 122(8):1701–1707, 2012.

[WYLP13] W.-K. Wong, B. Yang, C. Liu, and P. Poignet. A quasi-spherical triangle-based approach for efficient 3-d soft-tissue motion tracking. *IEEE/ASME Transactions on Mechatronics*, 18(5):1472–1484, 2013. `doi:10.1109/TMECH.2012.2203919`.

[XP98] C. Xu and J. L. Prince. Generalized gradient vector flow external forces for active contours. *Signal Processing*, 71(2):131–139, 1998. `doi:10.1016/S0165-1684(98)00140-6`.

[YGMY16] M. Ye, S. Giannarou, A. Meining, and G.-Z. Yang. Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. *Medical Image Analysis*, 30:144–157, 2016.

[YJL$^+$14] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang. Fast stereo matching using adaptive guided filtering. *Image and Vision Computing*, 32(3):202–211, 2014.

[YK06] K.-J. Yoon and I. S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006.

[YLMR15] S. Yang, L. A. Lobes, J. N. Martel, and C. N. Riviere. Handheld-automated microsurgical instrumentation for intraocular laser surgery. *Lasers in Surgery and Medicine*, 47(8):658–668, 2015. `doi:10.1002/lsm.22383`.

[YLS$^+$12] M.C. Yip, D.G. Lowe, S.E. Salcudean, R.N. Rohling, and C.Y. Nguan. Tissue tracking and registration for image-guided surgery. *IEEE Transactions on Medical Imaging*, 31(11):2169–2182, 2012. `doi:10.1109/TMI.2012.2212718`.

[YWLP14] B. Yang, W.-K. Wong, C. Liu, and P. Poignet. 3d soft-tissue tracking using spatial-color joint probability distribution and thin-plate spline model. *Pattern Recognition*, 47(9):2962–2973, 2014. `doi:10.1016/j.patcog.2014.03.020`.

[YYM$^+$10] N. Yamanaka, H. Yamashita, K. Masamune, T. Chiba, and T. Dohi. An endoscope with 2 DOFs steering of coaxial Nd:YAG laser beam for fetal surgery. *IEEE/ASME Transactions on Mechatronics*, 15(6):898–905, 2010. `doi:10.1109/TMECH.2010.2078828`.

[ZGH09] J. Zhu, L. Van Gool, and S. C. H. Hoi. Unsupervised face alignment by robust nonrigid mapping. *Proc. IEEE International Conference on Computer Vision*, pages 1265–1272, 2009.

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330 – 1334, 2000. `doi:10.1109/34.888718`.

[ZHAK08] C. Zinner, M. Humenberger, K. Ambrosch, and W. Kubinger. An optimized software-based implementation of a census-based stereo matching algorithm. *Advances in Visual Computing*, 5358:216–227, 2008. `doi:10.1007/978-3-540-89639-5{\_}21`.

[ZIN$^+$13] S. Zenbutsu, T. Igarashi, R. Nakamura, T. Nakaguchi, and T. Yamaguchi. 3d ultrasound assisted laparoscopic liver surgery by visualization of blood vessels. *Proc. IEEE International Ultrasonics Symposium*, pages 840–843, 2013. `doi:10.1109/ULTSYM.2013.0216`.

[ZK11] K. Zhang and J. U. Kang. Real-time intraoperative 4D full-range FD-OCT based on the dual graphics processing units architecture for microsurgery guidance. *Biomedical Optics Express*, 2(4):764–770, 2011.

[ZLH09] J. Zhu, M. R. Lyu, and T. S. Huang. A fast 2d shape recovery approach by fusing features and appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1210–1224, 2009. `doi:10.1109/TPAMI.2008.151`.

[ZPW$^+$14] Y. Zhang, T. Pfeiffer, M. Weller, W. Wieser, R. Huber, J. Raczkowsky, J. Schipper, H. Wörn, and T. Klenzner. Optical coherence tomography guided laser cochleostomy: Towards the accuracy on tens of micrometer scale. *BioMed Research International*, 2014, 2014.

[ZS84] T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984. `doi:10.1145/357994.358023`.

[ZW94] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. *Proc. European conference on computer vision*, 801:151–158, 1994.

# A Reprint Permissions

Table A.1 lists third-party publications where figures have been reproduced from with kind permission of associated publishers, including license numbers (via Copyright Clearance Center (CCC)) if provided.

Table A.1: Giving credit to referenced third-party publications for reprinting of figures.

| | Publication | Publisher | Remarks |
|---|---|---|---|
| Figure 1.5a | [LMH$^+$16] | John Wiley and Sons | License number: 4456140050473 |
| Figure 1.5b | [PRK$^+$12] | SPIE Society | — |
| Figure 1.5c | [RTR$^+$16] | IEEE | — |
| Figure 1.6a | [AT15] | SAGE Publications | — |
| Figure 1.6b | [YLMR15] | John Wiley and Sons | License number: 4456150051413 |
| Figure 1.7a | [SFSA12] | Elsevier | License number: 4456150966186 |
| Figure 1.7b | [RBS$^+$12] | John Wiley and Sons | License number: 4456151361389 |
| Figure 1.8a | [GVSY13] | IEEE | — |
| Figure 1.8b | [RPL10] | SAGE Publications | — |
| Figure 1.8c | [ZLH09] | IEEE | — |
| Figure 1.9a | [BM16] | IEEE | — |
| Figure 1.9b / 5.9d | [ARV$^+$06] | John Wiley and Sons | License number: 4457001093399 |
| Figure 1.9c / 5.9f | [KGvG$^+$08] | John Wiley and Sons | License number: 4457010611176 |

Table A.2 lists publications originated from the research conducted in this dissertation. The content of chapters, including figures and tables, has mainly been reproduced from the listed publications with kind permission of associated publishers, including license numbers (via CCC) if provided.

Table A.2: Giving credit to publications originated from the research of this dissertation.

| | Publication | Publisher | Remarks |
|---|---|---|---|
| Chapter 3 | [SPK$^+$13] | CRAS Society | — |
| Chapter 3 | [SKKO16] | Springer Nature | License number: 4455540483639 |
| Chapter 4 | [SKK$^+$15] | Springer Nature | License number: 4455530798533 |
| Chapter 4 | [BKS$^+$15] | SPIE Society | — |
| Chapter 5 | [SLK$^+$14] | CURAC Society | — |
| Chapter 5 | [SLK$^+$15] | SPIE Society | — |
| Chapter 5 | [CGS$^+$15] | CURAC Society | — |
| Chapter 5 | [SKL$^+$16] | Elsevier | — |
| Chapter 6 | [SLKO16] | Springer Nature | License number: 4455550353823 |
| Chapter 6 | [SKKO17] | Elsevier | Original work published under CC BY license |

# B Curriculum Vitae

### Personal

| | |
|---|---|
| Name | Andreas Schoob |
| Date of Birth | 12th December 1985 |
| Place of Birth | Dessau |
| Nationality | German |

### Work Experience

| | |
|---|---|
| since May/17 | Computer Vision Engineer at Yuanda Robotics GmbH, Hannover |
| Dec/11 – Apr/17 | Research Associate at the Institute of Mechatronic Systems, Leibniz Universität Hannover |

### Academic Education

| | |
|---|---|
| Oct/06 – Nov/11 | Mechatronics (Diplom-Ingenieur)<br>Technische Universität Dresden<br>Focus Areas: Information Processing, Micro and Medical Technology |
| Aug/08 | Mechatronics (Prediploma)<br>Technische Universität Dresden |

### Student Experiences and Interships

| | |
|---|---|
| Apr/11 – Oct/11 | Degree Candidate<br>Electronic Development, Ottobock HealthCare GmbH, Duderstadt |
| Oct/10 – Feb/11 | Internship<br>Electronic Development, Ottobock HealthCare GmbH, Duderstadt |
| Aug/09 – Sep/09 | Internship<br>Electronic Development, seleon GmbH, Dessau |
| May/05 – Jun/05 | Internship<br>Production Department, AEM Dessau GmbH, Dessau |

### School Education

| | |
|---|---|
| Aug/96 – Jul/05 | Gymnasium Philanthropinum, Dessau |
| Aug/92 – Jul/96 | Grundschule Mauerstraße, Dessau |

Hannover, November 2018