Medical Image Analysis 40 (2017) 80–95

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery

Andreas Schoob*, Dennis Kundrat, Lüder A. Kahrs, Tobias Ortmaier

Leibniz Universität Hannover, Institute of Mechatronic Systems, Appelstr. 11a, 30167 Hanover, Germany

ARTICLE INFO

Article history:
Received 8 November 2016
Revised 11 April 2017
Accepted 6 June 2017
Available online 8 June 2017

Keywords:
Non-rigid tracking
Stereo vision
Epipolar constraint
Motion compensation

ABSTRACT

Recent research has revealed that image-based methods can enhance accuracy and safety in laser microsurgery. In this study, non-rigid tracking using surgical stereo imaging and its application to laser ablation is discussed. A recently developed motion estimation framework based on piecewise affine deformation modeling is extended by a mesh refinement step and considering texture information. This compensates for tracking inaccuracies potentially caused by inconsistent feature matches or drift. To facilitate online application of the method, computational load is reduced by concurrent processing and affine-invariant fusion of tracking and refinement results. The residual latency-dependent tracking error is further minimized by Kalman filter-based upsampling, considering a motion model in disparity space. Accuracy is assessed in laparoscopic, beating heart, and laryngeal sequences with challenging conditions, such as partial occlusions and significant deformation. Performance is compared with that of state-of-the-art methods. In addition, the online capability of the method is evaluated by tracking two motion patterns performed by a high-precision parallel-kinematic platform. Related experiments are discussed for tissue substitute and porcine soft tissue in order to compare performances in an ideal scenario and in a setup mimicking clinical conditions. Regarding the soft tissue trial, the tracking error can be significantly reduced from 0.72 mm to below 0.05 mm with mesh refinement. To demonstrate online laser path adaptation during ablation, the non-rigid tracking framework is integrated into a setup consisting of a surgical Er:YAG laser, a three-axis scanning unit, and a low-noise stereo camera. Regardless of the error source, such as laser-to-camera registration, camera calibration, image-based tracking, and scanning latency, the ablation root mean square error is kept below 0.21 mm when the sample moves according to the aforementioned patterns. Final experiments regarding motion-compensated laser ablation of structurally deforming tissue highlight the potential of the method for vision-guided laser surgery.

1. Introduction

Surgery on delicate anatomical structures often demands high-resolution imaging and microinstruments for precise soft tissue manipulation. More advanced tools, such as medical lasers, facilitate contactless treatment of the pathology and minimize tissue trauma. A state-of-the-art clinical application is transoral laser microsurgery (TLM) for resection of benign or cancerous tissue on vocal cords (Rubinstein and Armstrong, 2011). Regarding the surgical treatment, a direct line-of-sight is established by inserting a laryngoscope in the throat of the patient. Precise resection of the lesion is achieved using a stereo microscope providing a magnified view of the surgical site and an ablation laser manually steered with a micromanipulator attached to the setup. Since the surgeon operates at a large distance from the patient, long and intensive training is required to master this task. Furthermore, soft tissue deformation induced by respiration artifacts and manipulation strongly affects the accuracy of laser ablation. Furthermore, misalignment of the laser path and loss of focus are evoked by the non-stiff mechanical fastening of the laser system to the patient; thus, motion externally applied to the microscope head most likely results in positional deviation of the laser spot. Deformations and camera motion are difficult to cope with, especially when the aim is function preservation with resection margins of less than 1 mm. To overcome this limitation, vision-based tracking of tissue motion and its application to motion-compensated laser ablation is addressed in this study as a continuation of our recently discussed method (Schoob et al., 2016) for image stabilization during incision planning. Moreover, the proposed method is not solely restricted to laser microsurgery. Further vision-guided, robot-assisted interventions or augmented reality concepts involving

* Corresponding author.
E-mail address: andreas.schoob@imes.uni-hannover.de (A. Schoob).

http://dx.doi.org/10.1016/j.media.2017.06.004

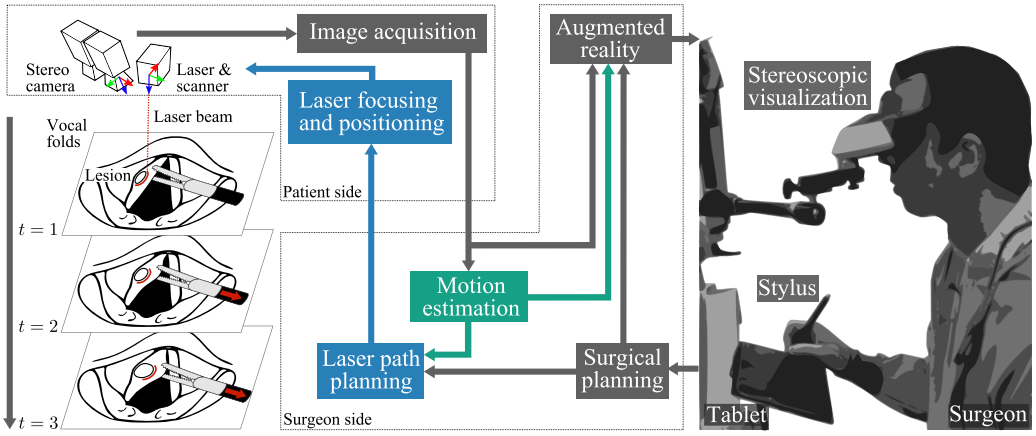

Fig. 1. Workflow for application of motion compensation in laser phonomicrosurgery. While exposing the vocal fold lesion (oval structure) by pulling with the grasping forceps, tissue motion is tracked to adapt online the ablation scan pattern (red line). Vision-guided laser control, as considered in this study, is intended to be integrated into a surgical framework with intuitive, stylus-tablet-based planning and augmented reality visualization as developed in the μRALP-project. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

superimposing tracking results directly to stereoscopic displays are conceivable.

Advances in laser surgery have been achieved regarding tablet-based planning interfaces (Tang et al., 2006; Mattos et al., 2014; Schoob et al., 2015c) and vision-guided laser control and micro-robotic scanning units for beam deflection (Dagnino et al., 2015; Renevier et al., 2016). Except for recent developments in the field of laser photocoagulation in retinal surgery, where tissue motion tracking performs adequately when rigid, affine or similarity transforms are used (Yang et al., 2015; Prokopetc and Bartoli, 2016), online estimation of larger tissue deformation during ablation has not been addressed so far. In particular, respiratory motion artifacts and tissue manipulation with surgical forceps as well as camera movements can lead to unintended injury of risk structures surrounding the lesion. To overcome this limitation, deformation tracking for online adaptation of laser spot positioning on the tissue surface is required (see Fig. 1).

Recently, vision-based tracking has been focused on minimally invasive surgery due to advances in medical imaging, augmented reality and robotics. Early studies discussed motion tracking, particularly in the field of beating heart surgery, considering matching of salient feature points (Ortmaier et al., 2005; Stoyanov et al., 2005; Sauvée et al., 2006). Intensive research has been conducted to improve robustness of feature-based tracking by including geometrical constraints for spatial consistency (Yip et al., 2012), multi-affine clustering of the target region (Puerto-Souza and Mariottini, 2013), affine-invariant feature descriptors (Giannarou et al., 2013), or online tracking-by-detection for surgical site retargeting (Ye et al., 2016).

By contrast, physical or geometric models can be incorporated into the non-rigid tracking framework. Associated optimization then aims at minimizing the shape bending energy and the matching error between the current frame and its template model. If shape priors are available or acquired preoperatively, organ deformation can be accurately estimated in real time depending on the complexity of the mechanical model and its intraoperative registration (Suwelack et al., 2014; Haouchine et al., 2015; Collins et al., 2016). If tracking of local deformations without knowledge of the anatomical shape is intended, stereo-based methods considering Free-Form Deformation (FFD) (e.g., piecewise bi-linear maps or B-splines) or Radial Basis Functions (RBF) (e.g., Thin Plate Splines (TPS)) have been shown to perform well for beating heart motion estimation (Lau et al., 2004; Stoyanov et al., 2004; Richa et al., 2010). In order to reduce the computational load when TPS is used, tracking can be split into intra-frame shape registration and inter-frame motion estimation (Yang et al., 2014). If a deformation is small, primitive models, such as quasi-spherical triangles, can perform as accurately as TPS-based methods (Wong et al., 2013). To further accelerate tracking, inverse compositional optimization (Brunet et al., 2011) or learning of non-linear template transformation provide promising solutions (Tan et al., 2014).

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 (Sotiras et al., 2013). A noteworthy method in the field of vision-based, non-rigid tracking (Pilet et al., 2008) 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. In addition to considering progressive finite Newton (PFN) optimization (Zhu et al., 2009b), application to soft tissue motion estimation for white light and multispectral imaging has been recently discussed (Stoyanov and Yang, 2009; Stoyanov et al., 2012; Du et al., 2015). Piecewise affine warps have been considered not only for endoscopic vision but also for online ultrasound image registration, due to their reduced computational complexity (Preiswerk et al., 2014; Royer et al., 2017).

Most model-based, non-rigid tracking methods cannot operate at image-acquisition rates of 30 Hz and higher. In particular, direct methods often require a non-deterministic, Gauss-Newton-like optimization scheme. Thus, convergence is not ensured until the camera acquires the next frame. In this regard, tracking with a fixed number of iterations or using general purpose graphics processing units (GPGPU) provide only a limited solution to the problem. In particular, if the image alignment error is large and the minimization process requires several frames, tracking might fail if the tissue concurrently undergoes significant motion or deformation.

This study presents a novel method for stereoscopic tracking of soft tissue motion. Extending our recent work (Schoob et al., 2016), we follow the idea of splitting the optimization into (1) robust, quasi-deterministic tracking and (2) appearance-based mesh refinement to compensate for tracking inaccuracies such as drift. Instead of sequential processing, as described in the original work (Zhu et al., 2009b), concurrent computation of both steps is proposed. Once convergence is reached for the mesh refinement, affine-invariant fusion with respect to the current tracking

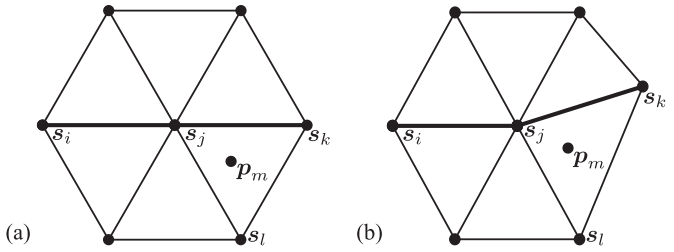

Fig. 2. Hexagonal element of the triangular mesh in (a) undeformed state and (b) deformed, penalized state.

estimate is performed at minimal computational cost. Even though concurrent processing significantly reduces the latency-dependent tracking error, further reduction is required if application in online laser ablation control is intended. To provide output at control loop frequency, Kalman filter-based upsampling of the motion measurements is used.

A major contribution of this work is that an epipolar constraint-based linear parametrization is applied throughout the entire framework of tracking, mesh refinement and motion upsampling. In contrast to the single-view approach combining tracking and refinement (Zhu et al., 2009b; Du et al., 2015), our method fully exploits stereoscopic constraints to estimate soft tissue motion, including changes in depth. Stereo-optical triangulation allows efficient computation of mesh structure and flow in task space. This enables real-time vision guidance. Moreover, stereoscopic augmented reality is feasible by directly superimposing the tracking results to the left and right camera view.

The proposed method is evaluated on *in vivo* image data. To assess the latency-dependent tracking error, an evaluation methodology with ground truth for each frame is presented. A comparative study considering tracking with and without mesh refinement as well as concurrent processing and motion upsampling is discussed. Finally, ablation trials on moving tissue samples demonstrate the potential of online tracking in laser-assisted surgery on soft tissue.

2. Material and methods

In this section, non-rigid tracking with a piecewise affine model is presented for motion compensation in laser surgery (see Fig. 1). Initially, non-rigid tracking for mono and stereo vision, as recently discussed, are briefly revisited. Subsequently, mesh refinement and related concurrent processing as well as fusion with the tracking result are described. With regard to integration into laser ablation control, filter-based motion upsampling is used.

2.1. Monoscopic non-rigid tracking

Tracking in mono view requires a triangular mesh, as illustrated in Fig. 2a. The area of interest is approximated by a triangular mesh of N vertices $\boldsymbol{s}_j = \left(u_j, v_j\right)^{\mathrm{T}}$ concatenated to form the vector

$$\boldsymbol{S} = (u_1, \ldots, u_N, v_1, \ldots, v_N)^{\mathrm{T}} \in \mathbb{R}^{2N}. \tag{1}$$

A point $\boldsymbol{p}_m$ inside this region can be described by the 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

$$\boldsymbol{W}(\boldsymbol{p}_m, \boldsymbol{S}) = \begin{pmatrix} u_j & u_k & u_l \\ v_j & v_k & v_l \end{pmatrix} \begin{pmatrix} \xi_j \\ \xi_k \\ \xi_l \end{pmatrix}. \tag{2}$$

Common feature matching techniques can be used to locally track motion. In this study, correspondence between consecutive frames

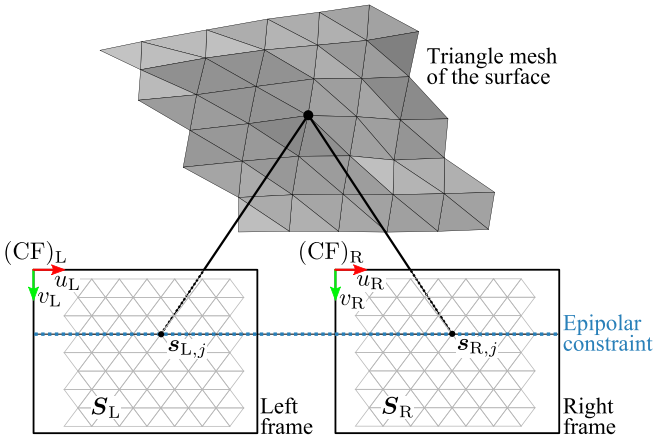

Fig. 3. Rectified stereo configuration illustrating the epipolar constraint for left-right consistency of the mesh model.

is established with the pyramidal Lucas–Kanade (LK) method (Bouguet, 2000). With the tracked features and the triangular mesh model, the non-rigid tracking problem can be formulated by the energy term

$$\varepsilon(\boldsymbol{S}) = \varepsilon_{\mathrm{C}}(\boldsymbol{S}) + \lambda_{\mathrm{D}}\varepsilon_{\mathrm{D}}(\boldsymbol{S}) \tag{3}$$

with ε_{C} defining the mesh correspondence energy and ε_{D} the mesh deformation energy weighted by λ_{D}. The correspondence energy ε_{C} is computed from local feature matches, whereas outliers are rejected in a coarse-to-fine scheme taking a penalty function with an iteratively decreasing confidence radius r into account (Zhu et al., 2009b). The deformation energy ε_{D} regularizes the deformation by considering the second-order derivatives of every collinear connected triplet $(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k)$ of each hexagon element (see Fig. 2b). Finally, the energy term (3) can be reformulated as an unconstrained quadratic optimization problem

$$\varepsilon(\boldsymbol{S}) = \boldsymbol{S}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{S} - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{S} + c, \tag{4}$$

where $\boldsymbol{U} \in \mathbb{R}^{2N\times 2N}$, $\boldsymbol{b} \in \mathbb{R}^{2N}$, and c is constant. For further description of the matrix $\boldsymbol{U}$ and the vector $\boldsymbol{b}$, the reader is referred to our previous study (Schoob et al., 2016). Finally, Eq. (4) is minimized with respect to $\boldsymbol{S}$ applying the quasi-deterministic, progressive finite Newton (PFN) scheme (Zhu et al., 2009b).

2.2. Stereoscopic non-rigid tracking

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 (Richa et al., 2010; Wong et al., 2013; Yang et al., 2014). 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 was found by formulating the problem in disparity space (Schoob et al., 2016). When the epipolar constraint for a rectified stereo configuration with coplanar image planes (see Fig. 3) is applied, a linear parametrization can be defined as follows

$$\boldsymbol{q} = (u_1, \ldots, u_N, v_1, \ldots, v_N, d_1, \ldots, d_N)^{\mathrm{T}}, \tag{5}$$

where the stacked vertex coordinates and the associated disparities $\boldsymbol{d} = (d_1, \ldots, d_N)^{\mathrm{T}}$ 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 $\left(u_{\mathrm{L},j}, v_{\mathrm{L},j}\right)^{\mathrm{T}} = \left(u_j, v_j\right)^{\mathrm{T}}$. For a rectified stereo view, as shown in Fig. 3, corresponding points are enforced to have the same vertical coordinate v. 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

$$\boldsymbol{S}_i = \boldsymbol{S}_i(\boldsymbol{q}) \quad \text{with } i \in \{\mathrm{L}, \mathrm{R}\} \tag{6}$$

with j-th mesh point

$$\boldsymbol{s}_{i,j}(\boldsymbol{q}) = \begin{cases} (u_j, v_j)^\mathrm{T} & \text{if } i = \mathrm{L} \\ (u_j + d_j, v_j)^\mathrm{T} & \text{if } i = \mathrm{R} \end{cases}. \tag{7}$$

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

$$\boldsymbol{W}_i(\boldsymbol{p}_m, \boldsymbol{q}) = \boldsymbol{W}(\boldsymbol{p}_m, \boldsymbol{S}_i(\boldsymbol{q})) = \begin{pmatrix} \boldsymbol{\xi}_m^\mathrm{T} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^\mathrm{T} \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}), \tag{8}$$

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 (Zhu et al., 2009b), 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

$$\varepsilon(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \varepsilon_\mathrm{C}(\boldsymbol{S}_i(\boldsymbol{q})) + \lambda_\mathrm{D} \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \varepsilon_\mathrm{D}(\boldsymbol{S}_i(\boldsymbol{q})) \tag{9}$$

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

$$\varepsilon_{\mathrm{C},i} = \boldsymbol{S}_i(\boldsymbol{q})^\mathrm{T} \boldsymbol{A}_i \boldsymbol{S}_i(\boldsymbol{q}) - 2\boldsymbol{b}_i^\mathrm{T} \boldsymbol{S}_i(\boldsymbol{q}) + c_i \tag{10}$$

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

$$\varepsilon_{\mathrm{D},i} = \boldsymbol{S}_i(\boldsymbol{q})^\mathrm{T} \mathcal{K}_i \boldsymbol{S}_i(\boldsymbol{q}) \tag{11}$$

of the left and right view, respectively. The matrices $\boldsymbol{A}_i \in \mathbb{R}^{2N\times 2N}$ and $\boldsymbol{b}_i \in \mathbb{R}^{2N}$ are obtained from the residuals for the inlier correspondences in the penalty function (Zhu et al., 2009b). The sparse matrix $\mathcal{K}_i \in \mathbb{R}^{2N\times 2N}$ consists of the squared second-order derivatives of the vertex coordinates. The total energy in Eq. (9) can be summarized as

$$\varepsilon(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \left[\boldsymbol{S}_i(\boldsymbol{q})^\mathrm{T} \boldsymbol{U}_i \boldsymbol{S}_i(\boldsymbol{q}) - 2\boldsymbol{b}_i^\mathrm{T} \boldsymbol{S}_i(\boldsymbol{q}) + c_i\right], \tag{12}$$

yielding an unconstrained quadratic optimization problem similar to Eq. (4). Thus, the PFN method can be used to minimize the energy term (12) by computing each Newton step

$$\Delta\boldsymbol{q} = -\boldsymbol{H}^{-1}(\boldsymbol{q})\, \nabla\varepsilon(\boldsymbol{q}) \tag{13}$$

in a coarse-to-fine scheme with a decreasing confidence radius r. Algorithm 1 summarizes the non-rigid tracking. For further details on the stereo extension and derivation of the gradient $\nabla\varepsilon(\boldsymbol{q})$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{q})$, the reader is referred to our previous study (Schoob et al., 2016).

In addition to each triangle center, salient gradient-based landmarks (usually 5 to 7 points) are selected as mesh support points (Shi, 1994). After establishing initial correspondence between consecutive frames using the LK method, the point positions are corrected by the PFN-based mesh deformation and the refinement presented in the next section. In addition, the endoscopic images are rank-encoded, providing increased robustness to nonlinear illumination changes (Zabih and Woodfill, 1994). Adopting the idea of identifying local multivariate outliers (Filzmoser et al., 2013), a cross-channel, pairwise Mahalanobis distance considering the spatial context of the image texture facilitates consistent temporal detection of occlusions (Schoob et al., 2016).

Algorithm 1: Stereoscopic non-rigid tracking.

pre-compute:
(1) Initialize $\boldsymbol{S}_{\{\mathrm{L},\mathrm{R}\}}$ according to (Schoob et al., 2016)
(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 $\nabla\varepsilon(\boldsymbol{q})$ (Schoob et al., 2016)
    (6) Compute Hessian $\boldsymbol{H}(\boldsymbol{q})$ (Schoob et al., 2016)
    (7) $\boldsymbol{q} \leftarrow \boldsymbol{q} + \Delta\boldsymbol{q}$ according to Eq. (13)
    (8) Update $\boldsymbol{S}_{\{\mathrm{L},\mathrm{R}\}}(\boldsymbol{q})$
    (9) $r \leftarrow \eta r$ with $0 < \eta < 1$
  until $r \le r_\mathrm{end}$
  output: Updated stereo mesh $\boldsymbol{S}_{\{\mathrm{L},\mathrm{R}\}}(\boldsymbol{q})$
end

2.3. 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) (Zhu et al., 2009b). As in Eq. (9), the refinement step considers a regularization term ε_D. The total energy for the stereo view is defined as follows

$$\varepsilon_\mathrm{MR}(\boldsymbol{q}) = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \varepsilon_\mathrm{A}(\boldsymbol{S}_i(\boldsymbol{q})) + \lambda_\mathrm{D} \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \varepsilon_\mathrm{D}(\boldsymbol{S}_i(\boldsymbol{q}))\ . \tag{14}$$

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 $\varepsilon_{\mathrm{A},i} = \varepsilon_\mathrm{A}(\boldsymbol{S}_i(\boldsymbol{q}))$

$$\varepsilon_{\mathrm{A},i} = \sum_{\boldsymbol{p}_m\in\Omega} \rho\left(\left[T_i(\boldsymbol{W}_i(\boldsymbol{p}_m, \Delta\boldsymbol{q})) - I_i(\boldsymbol{W}_i(\boldsymbol{p}_m, \boldsymbol{q}))\right]^2\right), \tag{15}$$

where we use the warping function (8) of the pixel $\boldsymbol{p}_m$ with $m \in \{1, \ldots, M\}$ in the image region Ω represented by the mesh. In order to provide increased robustness against tracking outliers, Eq. (15) is formulated in an iteratively re-weighted least squares framework based on the norm-like Huber function

$$\rho(u) = \begin{cases} u & \text{if } u \le \sigma_H^2 \\ (2\sqrt{u} - \sigma_H)\sigma_H & \text{if } u > \sigma_H^2 \end{cases}. \tag{16}$$

For small residuals, $\rho(u)$ 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 Eq. (16), to reduce the weight of outliers. In this case, the norm-like Huber function switches to linear behavior for large residuals. This has been proven to perform well in image-based structure and motion estimation (Hager and Belhumeur, 1998; Zhu et al., 2009a; Chang et al., 2013). Our stereo approach will take adaptive cost re-weighting into account in order to increase tracking robustness for laser ablation on soft tissue.

Applying the first order Taylor expansion, we obtain the following linearization of the mesh coordinates

$$\boldsymbol{S}_i(\boldsymbol{q}) \to \boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q}) = \boldsymbol{S}_i(\boldsymbol{q}) + \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{q}} \Delta\boldsymbol{q}. \tag{17}$$

This allows reformulating the deformation energy (11), yielding

$$\varepsilon_{\mathrm{D},i} \approx (\boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q}))^\mathrm{T} \mathcal{K}_i (\boldsymbol{S}_i(\boldsymbol{q}) + \Delta\boldsymbol{S}_i(\boldsymbol{q}))\ . \tag{18}$$

Analogously to Eq. (18), the appearance-based energy (15) can be linearized to

$$\varepsilon_{\mathrm{A},i} \approx \sum_{\boldsymbol{p}_m \in \Omega} \rho\left(\left[\, \Delta I_{i,m} + \boldsymbol{J}_{i,m}\, \Delta \boldsymbol{q} \right]^2\right) \tag{19}$$

with residual

$$\Delta I_{i,m} = T_i(\boldsymbol{W}_i(\boldsymbol{p}_m, \boldsymbol{q}_0)) - I_i(\boldsymbol{W}_i(\boldsymbol{p}_m, \boldsymbol{q})) \tag{20}$$

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

$$\boldsymbol{J}_{i,m} = \left.\frac{\partial T_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} = \nabla T_i \left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} \in \mathbb{R}^{1\times 3N}, \tag{21}$$

where $\nabla T_i = \left(\frac{\partial T_i}{\partial u}, \frac{\partial T_i}{\partial v}\right)$ denotes the image gradient at $\boldsymbol{p}_m$. The Jacobian of the warp is obtained by the product of the derivative of Eq. (8)

$$\left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{S}_i}\right|_{\boldsymbol{p}_m} = \begin{pmatrix} \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{2\times 2N} \tag{22}$$

and the derivative of the mesh coordinates

$$\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{q}} = \begin{pmatrix} \boldsymbol{I}_N & \boldsymbol{0}_N & *_i \\ \boldsymbol{0}_N & \boldsymbol{I}_N & \boldsymbol{0}_N \end{pmatrix} \in \mathbb{R}^{2N\times 3N} \tag{23}$$

with respect to stereoscopic parametrization $\boldsymbol{q}$. According to i, $*_i$ is

$$*_i = \begin{cases} \boldsymbol{0}_N & \text{if } i = \mathrm{L} \\ \boldsymbol{I}_N & \text{if } i = \mathrm{R} \end{cases} \in \mathbb{R}^{N\times N}. \tag{24}$$

Hence, $*_i$ is either the zero matrix $\boldsymbol{0}_N$ or the identity matrix $\boldsymbol{I}_N$. Consequently, the Jacobian of the warp is given by

$$\left.\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}_m} = \frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{S}_i}\frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{q}} = \begin{pmatrix} \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N & \star_i \\ \boldsymbol{0}_N & \boldsymbol{\xi}_m^{\mathrm{T}} & \boldsymbol{0}_N \end{pmatrix} \in \mathbb{R}^{2\times 3N}, \tag{25}$$

where the placeholder

$$\star_i = \begin{cases} \boldsymbol{0}_N & \text{if } i = \mathrm{L} \\ \boldsymbol{\xi}_m^{\mathrm{T}} & \text{if } i = \mathrm{R} \end{cases} \in \mathbb{R}^{1\times N} \tag{26}$$

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 Eq. (8). Due to the inverse compositional algorithm and the linearity of $\boldsymbol{S}_i(\boldsymbol{q})$, the Jacobian matrix (21) 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

$$\Delta \boldsymbol{q} = -\boldsymbol{H}_{\mathrm{MR}}^{-1} \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \left(\boldsymbol{J}_{\mathrm{S},i}^{\mathrm{T}} \boldsymbol{W}_i\, \Delta \boldsymbol{I}_{\mathrm{S},i} + \lambda_{\mathrm{D}} \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}}}{\partial \boldsymbol{q}} \mathcal{K}_i\, \boldsymbol{S}_i(\boldsymbol{q})\right) \tag{27}$$

with stereo-based Hessian matrix

$$\boldsymbol{H}_{\mathrm{MR}} = \sum_{i\in\{\mathrm{L},\mathrm{R}\}} \left(\boldsymbol{J}_{\mathrm{S},i}^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{J}_{\mathrm{S},i} + \lambda_{\mathrm{D}} \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})^{\mathrm{T}}}{\partial \boldsymbol{q}} \mathcal{K}_i \frac{\partial \boldsymbol{S}_i(\boldsymbol{q})}{\partial \boldsymbol{q}}\right) \tag{28}$$

and diagonal matrix of the weights

$$\boldsymbol{W}_i = \mathrm{diag}\left(\rho'\left(\Delta I_{i,1}^2\right), \ldots, \rho'\left(\Delta I_{i,M}^2\right)\right). \tag{29}$$

In Eqs. (27) and (28), the pointwise residuals and the Jacobians form the matrices

$$\Delta \boldsymbol{I}_{\mathrm{S},i} = (\Delta I_{i,1}, \ldots, \Delta I_{i,M})^{\mathrm{T}} \tag{30}$$

and

$$\boldsymbol{J}_{\mathrm{S},i} = \left(\boldsymbol{J}_{i,1}^{\mathrm{T}}, \ldots, \boldsymbol{J}_{i,M}^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{31}$$

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 β. Consequently, the indicator function

$$\delta_{\mathrm{MHD}} = \begin{cases} 1 & \text{if MHD} \geq \beta \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

causes the modified weight

$$\rho'(u) = \left(1 - \delta_{\mathrm{MHD}}\right) \frac{\partial \rho(u)}{\partial u} \tag{33}$$

to be set to zero in case of occlusion. Thus, the associated pixel does not contribute to the mesh refinement process.

To reduce the computational load of the iteratively re-weighted least squares method, the H-algorithm is implemented (Dutter and Huber, 1981). 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 (Baker et al., 2003). In addition, parameter estimation is implemented hierarchically in a coarse-to-fine scheme, reducing the computational load and robustly tracking large displacements caused by rapid scene and camera motion. Algorithm 2 summarizes the stereo-based mesh refinement outlined above.

Algorithm 2: 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 Eq. (31)
(3) Initialize unweighted Hessian $\boldsymbol{H}_{\mathrm{MR}}$ according to Eq. (28)
for *each stereo image pair* **do**
  input: Parameter vector $\boldsymbol{q}$ from tracking(Algorithm 1)
  repeat
    (4) Warp image $I_{\{\mathrm{L,R}\}}$ according to Eq. (8)
    (5) Compute residuals $\Delta \boldsymbol{I}_{\mathrm{S},\{\mathrm{L,R}\}}$ acc. Eq. (20),(30)
    (6) Compute weightings $\boldsymbol{W}_{\{\mathrm{L,R}\}}$ acc. Eq. (29),(33)
    (7) $\boldsymbol{q} \leftarrow \boldsymbol{q} + \Delta\boldsymbol{q}$ according to Eq. (27)
    (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

2.4. Concurrent tracking and mesh refinement

Stereoscopic tracking without mesh refinement (noMR) according to Algorithm 1 provides high update rate; however, it may lead to drift over time, since appearance is not considered (see Fig. 4a). To compensate for drift, the mesh refinement step according to Algorithm 2 can be invoked, once Algorithm 1 has finished. This is called sequential mesh refinement (sMR) strategy (see Fig. 4b).

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

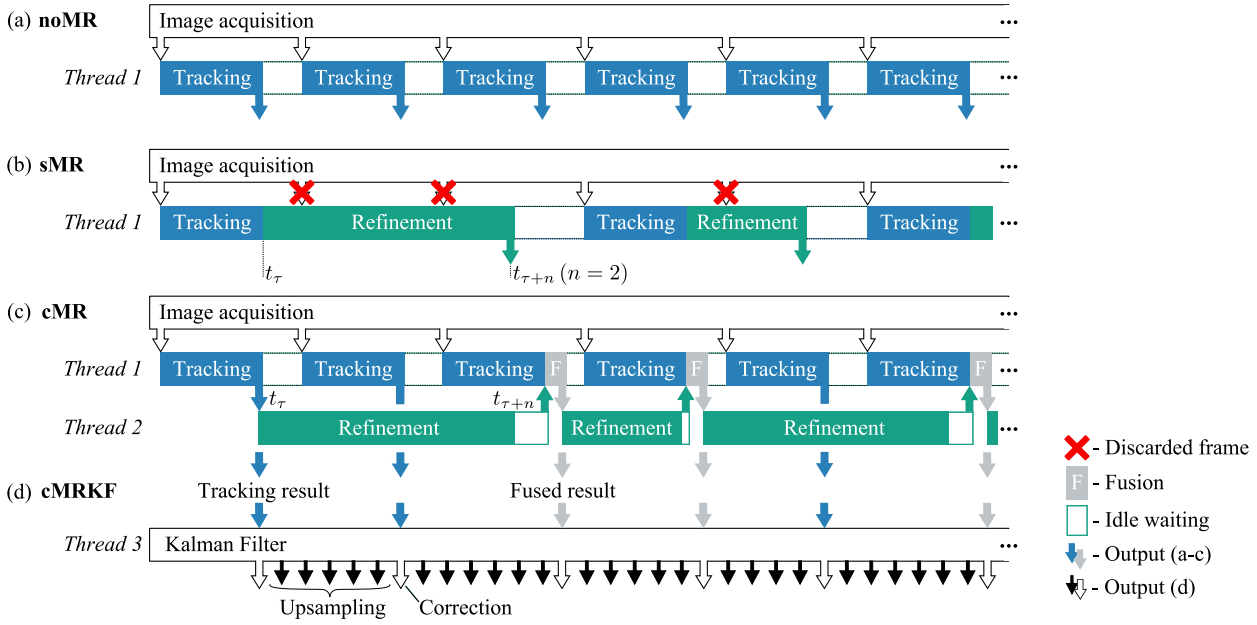

Fig. 4. Computational pipeline considering (a) tracking by Algorithm 1 without mesh refinement (noMR), (b) sequential (sMR) and (c) concurrent mesh refinement (cMR), both deploying Algorithms 1 and 2. In (c), tracking runs in Thread 1 (CPU) and the refinement in Thread 2 (GPU). Subsequently, the proposed 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).

limited solution due to the non-deterministic optimization scheme. 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 Fig. 4b). This may lead not only to significantly delayed measurements but also to tracking failure if the scene concurrently undergoes rapid motion or large local deformation. Thus, concurrent tracking and mesh refinement (cMR) with subsequent affine-invariant fusion, as illustrated in Fig. 4c, is discussed (see next section). This method is compared with noMR (equivalent to (Schoob et al., 2016)) and sMR. The latter method corresponds to the monoscopic DLK-algorithm (Zhu et al., 2009b; Du et al., 2015) 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 Fig. 4d.

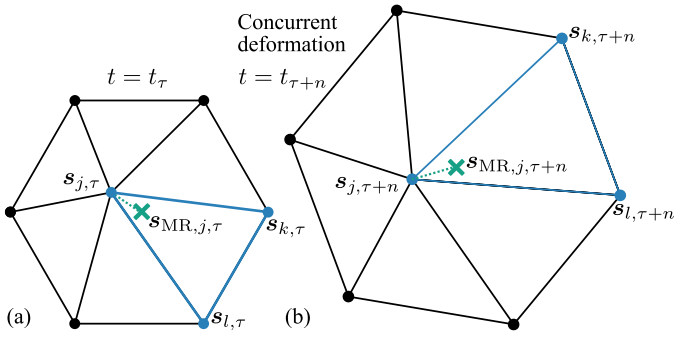

Fig. 5. Mesh in (a) initial configuration at time t_τ and (b) subsequently tracked position at time $t_{\tau+n}$. If drift occurs for vertex $\boldsymbol{s}_{j,\tau}$, the proposed mesh refinement yields the corrected position $\boldsymbol{s}_{\mathrm{MR},j,\tau}$. Fusion of delayed mesh refinement with respect to t_τ is achieved at subsequent time $t_{\tau+n}$ even if the mesh concurrently undergoes deformation, here exemplarily shown for rotation and scale change between (a) and (b).

2.5. Affine-invariant fusion of tracking and mesh refinement

Regarding sMR, motion is initially estimated according to Algorithm 1. Subsequently, Algorithm 2 is initiated at time t_τ to compensate for mesh misalignment caused, for instance, by drift (see Fig. 4b). Since the refinement processes dense texture information in a non-deterministic optimization framework, 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 Fig. 4c, 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_τ to $t_{\tau+n}$ with mesh vertex $\boldsymbol{s}_{j,\tau}$ being subject to drift (see Fig. 5a). Due to computational delay, fusion of the refinement result $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ referring to t_τ cannot be achieved until $t_{\tau+n}$. To compensate for simultaneous deformation (see Fig. 5b), affine-invariant fusion of the corrected mesh vertex position $\boldsymbol{s}_{\mathrm{MR},j,\tau}$ is performed by

$$\begin{aligned}\boldsymbol{s}_{\mathrm{MR},j,\tau+n} &= \boldsymbol{W}_i(\boldsymbol{s}_{\mathrm{MR},j,\tau}, \boldsymbol{q}_{\tau+n}) \\ &= \begin{pmatrix}\boldsymbol{s}_{j,\tau+n} & \boldsymbol{s}_{k,\tau+n} & \boldsymbol{s}_{l,\tau+n}\end{pmatrix} \begin{pmatrix}\xi_{\mathrm{MR},j,\tau} \\ \xi_{\mathrm{MR},k,\tau} \\ \xi_{\mathrm{MR},l,\tau}\end{pmatrix}, \end{aligned} \tag{34}$$

where the parametrization $\boldsymbol{q}_{\tau+n}$ is obtained from stereo-based tracking with Algorithm 1 running in Thread 1. After Thread 2 finishes Algorithm 2 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

$$0 \leq \xi_{\mathrm{MR},\{j,k,l\},\tau} \leq 1\ . \tag{35}$$

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 Eq. (34), instead of cartesian vertex positions, allows affine-invariant fusion of tracking and refinement at $t_{\tau+n}$, independent of concurrent deformation. The

entire mesh coordinates in the left and right camera view can be corrected in one step by computing

$$\boldsymbol{S}_{\mathrm{MR},i,\tau+n} = \begin{pmatrix} \boldsymbol{\xi}^{\mathrm{T}}_{\mathrm{MR},1,\tau} & \boldsymbol{0}_N \\ \vdots & \vdots \\ \boldsymbol{\xi}^{\mathrm{T}}_{\mathrm{MR},N,\tau} & \boldsymbol{0}_N \\ \boldsymbol{0}_N & \boldsymbol{\xi}^{\mathrm{T}}_{\mathrm{MR},1,\tau} \\ \vdots & \vdots \\ \boldsymbol{0}_N & \boldsymbol{\xi}^{\mathrm{T}}_{\mathrm{MR},N,\tau} \end{pmatrix} \boldsymbol{S}_i(\boldsymbol{q}_{\tau+n}) \tag{36}$$

considering the pixelwise warp function (34) in a stacked formulation of Eq. (8). Since the epipolar constraint is satisfied for both tracking and mesh refinements, it implicitly applies to Eq. (36).

While Algorithm 1 runs on the CPU in Thread 1 (Core i7-3770, Intel Corporation, Santa Clara, CA, USA), Algorithm 2 is computed asynchronously in Thread 2 deploying the CUDA framework and a GeForce GTX Titan GPU (NVIDIA, Santa Clara, CA, USA). Once Thread 2 finishes, fusion according to Eq. (36) is performed. In contrast to Algorithm 1, inlier check (35) and the subsequent matrix multiplication (36) are computed at negligible costs.

2.6. Upsampling of the motion measurements

To further reduce the latency-dependent tracking misalignment, the epipolar constraint-based parameter set (5) 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 (Vaudrey et al., 2008), we propose each vertex $\boldsymbol{s}_j$ to be tracked individually with state vector

$$\boldsymbol{x}_{j,\tau} = \left(u_{j,\tau}, v_{j,\tau}, d_{j,\tau}, \dot{u}_{j,\tau}, \dot{v}_{j,\tau}, \dot{d}_{j,\tau}\right)^{\mathrm{T}}, \tag{37}$$

where the first subscript j indicates the vertex index and the second subscript τ time step t_τ. Instead of modeling the entire mesh within a single state space model, representation (37) 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 $\left(\dot{u}_j, \dot{v}_j\right)^{\mathrm{T}}$, 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 us to find an optimal state estimate, taking process and measurement noise into account. The three motion directions $\{u, v, d\}$ can be considered as independent; thus, state representation (37) can be separated into the following three state vectors

$$\begin{aligned} \boldsymbol{x}_{\mathrm{u},j,\tau} &= \left(u_{j,\tau}, \dot{u}_{j,\tau}\right)^{\mathrm{T}} \\ \boldsymbol{x}_{\mathrm{v},j,\tau} &= \left(v_{j,\tau}, \dot{v}_{j,\tau}\right)^{\mathrm{T}} \\ \boldsymbol{x}_{\mathrm{d},j,\tau} &= \left(d_{j,\tau}, \dot{d}_{j,\tau}\right)^{\mathrm{T}}, \end{aligned} \tag{38}$$

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

A dynamic system is modeled by the discretized state and measurement equation

$$\begin{aligned} \boldsymbol{x}_{\mathrm{d},j,\tau} &= \boldsymbol{F}_\tau \boldsymbol{x}_{\mathrm{d},j,\tau-1} + \boldsymbol{w}_\tau \\ z_{\mathrm{d},j,\tau} &= \boldsymbol{H}_\tau \boldsymbol{x}_{\mathrm{d},j,\tau} + v_\tau, \end{aligned} \tag{39}$$

where the process and measurement noise are represented by normal probability distributions $\boldsymbol{w}_\tau \propto \mathcal{N}(0, \boldsymbol{Q}_\tau)$ and $v_\tau \propto \mathcal{N}(0, R_\tau)$ with covariance matrix $\boldsymbol{Q}_\tau \in \mathbb{R}^{2\times 2}$ and variance R_τ, respectively. The state transition and measurement matrices are denoted by $\boldsymbol{F}_\tau \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{H}_\tau \in \mathbb{R}^{1\times 2}$, respectively. The 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

$$\boldsymbol{F}_\tau = \begin{pmatrix} 1 & \Delta T \\ 0 & 1 \end{pmatrix}, \tag{40}$$

where ΔT is the sample time between t_τ and $t_{\tau-1}$. 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 estimate

$$\begin{aligned} \boldsymbol{x}^-_{\mathrm{d},j,\tau} &= \boldsymbol{F}_\tau \boldsymbol{x}_{\mathrm{d},j,\tau-1} \\ \boldsymbol{P}^-_\tau &= \boldsymbol{F}_\tau \boldsymbol{P}_{\tau-1} \boldsymbol{F}^{\mathrm{T}}_\tau + \boldsymbol{Q}_\tau\ . \end{aligned} \tag{41}$$

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

$$\begin{aligned} \boldsymbol{x}_{\mathrm{d},j,\tau} &= \boldsymbol{x}^-_{\mathrm{d},j,\tau} + \boldsymbol{K}_\tau \left(z_{\mathrm{d},j,\tau} - \boldsymbol{H}_\tau \boldsymbol{x}^-_{\mathrm{d},j,\tau}\right) \\ \boldsymbol{K}_\tau &= \boldsymbol{P}^-_\tau \boldsymbol{H}^{\mathrm{T}}_\tau \left(\boldsymbol{H}_\tau \boldsymbol{P}^-_\tau \boldsymbol{H}^{\mathrm{T}}_\tau + R_\tau\right)^{-1} \\ \boldsymbol{P}_\tau &= \left(\boldsymbol{I} - \boldsymbol{K}_\tau \boldsymbol{H}_\tau\right) \boldsymbol{P}^-_\tau, \end{aligned} \tag{42}$$

where $\boldsymbol{P}_\tau \in \mathbb{R}^{2\times 2}$ denotes the 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

$$\boldsymbol{Q}_\tau = \begin{pmatrix} \frac{\Delta T^4}{4} & \frac{\Delta T^3}{2} \\ \frac{\Delta T^3}{2} & \Delta T^2 \end{pmatrix} \sigma_Q^2\ , \tag{43}$$

depending on the uncertainty σ_Q. Since only position is measured, the related noise variance is defined as $R_\tau = \sigma_R^2$. 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 Fig. 4d. If no measurement is available, only state prediction (41) is performed to upsample the previous measurement with the underlying motion model and to reduce the latency-dependent misalignment. Once an image-based tracking result is available, Eq. (42) corrects the motion estimate.

3. Experimental

Initially, the methodology for accuracy assessment on *in vivo* tissue (IVT) is presented. Subsequently, the system design for online performance evaluation of the tracking with two motion patterns is illustrated. Based on integration into a surgical framework according to Fig. 1, laser ablation trials conducted on tissue substitute and porcine *ex vivo* tissue (EVT) are described.

Table 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	448	μRALP sequence with partial occlusion
SEQ8	368	μRALP sequence with large deformation

3.1. Tracking accuracy assessment on IVT

Tracking performance is assessed on laparoscopic, beating heart and laryngeal *in vivo* tissue (IVT) datasets. Five reference points $\boldsymbol{p}_{m,\mathrm{GT}}$ with $m \in \{1,\ldots,5\}$, 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 pointwise error is then given by

$$e_{\mathrm{Track}}(\boldsymbol{p}_m) = \left\| {}_{(\mathrm{L})}\boldsymbol{p}_{m,\mathrm{Track}} - {}_{(\mathrm{L})}\boldsymbol{p}_{m,\mathrm{GT}} \right\|_2, \tag{44}$$

considering the back-projected points with reference to frame $(\mathrm{CF})_{\mathrm{L}}$.

In total, eight stereo sequences are considered (see Table 1). Sequences SEQ1–5 are obtained from the laparoscopic Hamlyn dataset[1] (Mountney et al., 2010; Stoyanov et al., 2005). The first three videos, denoted by SEQ1–3, are adopted from a laparoscopic porcine procedure including a scale change, a simulated occlusion in a nearly static scene, and significant deformation, respectively. Datasets SEQ4 and SEQ5 describe more challenging beating heart tracking scenarios that have already been considered (Stoyanov et al., 2005; Richa et al., 2010). The last three videos, denoted by SEQ6–8, are captured with a stereo endoscope (VSii, Visionsense, Petach-Tikva, Israel, stereo baseline of 1 mm) in an *in vivo* laryngeal intervention in the μRALP project[2]. To summarize, the presented method is evaluated in three different surgical scenarios providing scenes with and without occlusion as well as significant deformation and changes in illumination. In the laparoscopic Hamlyn dataset, the tracked region is located at a distance of 170 mm (camera baseline ∼ 5 mm). For the beating heart and the laryngeal *μ*RALP sequences, the distance amounts to 40 mm (baseline ∼ 5 mm) and 20 mm (baseline ∼ 1 mm) on average, respectively. The analyzed sequences are listed in Table 1.

During the evaluation study, Algorithm 1 was parametrized with $\lambda_{\mathrm{D}} = 0.01$, $\beta = 0.03$ and a triangle width of 35 pixels. Regarding Algorithm 2, the Huber threshold for the rank-transform residuals is set to $\sigma_H = 10$. The number of pyramidal levels has been fixed to three with a maximum of 20 iterations per level and a termination criterion of $\epsilon = 0.03$.

Regarding the tracking performance of cMRKF on the *in vivo* data, results are compared with not only those of noMR and sMR but also those of state-of-the-art algorithms, namely, an implementation of the TPS-based non-rigid tracking as a further direct method (Richa et al., 2010), and the hierarchical multi-affine (HMA) feature-matching toolbox (Puerto-Souza and Mariottini, 2013). The stereoscopic TPS method was reimplemented including specular highlight filtering as well as CUDA optimization. For providing an acceptable tradeoff between accuracy and runtime, a set of 3 × 3 control points was used. The HMA algorithm is discussed for two different strategies. Features are matched either with respect to the initial frame (HMAi) or between consecutive frames of the image sequence (HMAc). In order to establish left-right correspondence for the monoscopic HMA algorithm and to initialize the TPS model in the right view, dense surface reconstruction was employed (Schoob et al., 2015a).

3.2. Assessment of the latency-dependent tracking error

In addition to the IVT trials, an in-depth validation of the latency-dependent tracking misalignment should demonstrate the superior performance of cMR(KF) compared with noMR and sMR. To provide accurate ground truth (GT) for each frame, a tissue sample is positioned on a high-precision, parallel-kinematic platform (Hexapod H-824.G11, Physik Instrumente (PI), Karlsruhe, Germany) and translated with a repeatability of ± 0.5 μm. Stereo images are acquired with a stereo camera (2 × 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 slightly converged with a baseline of 37 mm at a distance of 60 mm to the sample surface. A schematic overview of the setup is shown in Fig. 6a.

For simplicity, tissue deformation is not considered in this part of the evaluation, since acquiring ground truth in real time is complex. Instead, rigid movements of the sample are performed while GT is measured from the hexapod encoder data. 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 $(\mathrm{CF})_{\mathrm{H},0}$ can be calculated by

$${}_{(\mathrm{H},0)}\tilde{\boldsymbol{p}}_{m,\mathrm{GT}} = {}^{\mathrm{S}}\boldsymbol{T}^{-1}_{\mathrm{H},0}\,{}_{(\mathrm{S})}\tilde{\boldsymbol{p}}_{m,\mathrm{GT}}, \tag{45}$$

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

$${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H},0} = {}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H}}\,{}^{\mathrm{H}}\boldsymbol{T}_{\mathrm{H},0}, \tag{46}$$

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

$${}^{\mathrm{S}}\boldsymbol{T}_{\mathrm{H}} = {}^{\mathrm{L}}\boldsymbol{T}^{-1}_{\mathrm{S,init}}\,{}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{H},0}\,{}^{\mathrm{H,init}}\boldsymbol{T}^{-1}_{\mathrm{H},0}, \tag{47}$$

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 considered to be equal to the initial hexapod rotation with respect to $(\mathrm{CF})_{\mathrm{L}}$. The image-based tracking result is finally mapped by

$${}_{(\mathrm{H},0)}\tilde{\boldsymbol{p}}_{m,\mathrm{track}} = {}^{\mathrm{L}}\boldsymbol{T}^{-1}_{\mathrm{H},0}\,{}_{(\mathrm{L})}\tilde{\boldsymbol{p}}_{m,\mathrm{track}} \tag{48}$$

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 (Tsai and Lenz, 1989). The latency-dependent misalignment (LD) is then assessed with respect to GT by the error function

$$e_{\mathrm{LD}}(\boldsymbol{p}_m) = \left\| {}_{(\mathrm{H},0)}\boldsymbol{p}_{m,\mathrm{track}} - {}_{(\mathrm{H},0)}\boldsymbol{p}_{m,\mathrm{GT}} \right\|_2 . \tag{49}$$

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 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 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 mm and

[1] http://hamlyn.doc.ic.ac.uk/vision/
[2] http://www.microralp.eu/

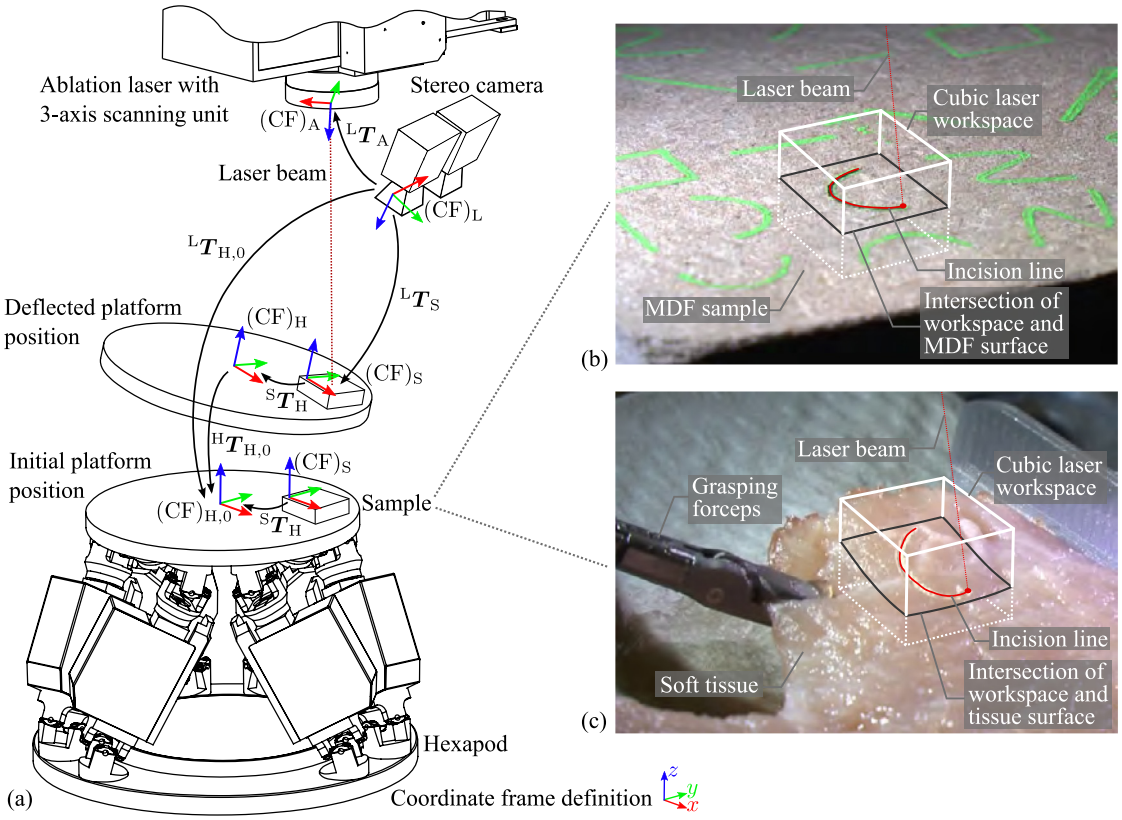

Fig. 6. Experimental design is shown in (a) with a rigid setup deploying a stereo camera, a surgical laser, and a parallel robot for positioning tasks to assess tracking performance. Motion estimation and laser ablation trials conducted on tissue substitute (MDF) and *ex vivo* tissue (EVT) samples are shown in (b,c). For both specimens, the surface has to be positioned properly in the cubic laser workspace.

a maximum velocity of 2.1 mm/s. The second scenario, which is called axial, is defined by movements with an amplitude of 4 mm at 1 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 (Schoob et al., 2015b).

For each motion pattern, two types of tissue are considered in the experimental study. As illustrated in Fig. 6b, 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 Fig. 6c). The active sensor area of the camera was cropped to 400 × 400 pixels enabling an image acquisition frame rate of 80 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 × 4 triangles with an edge length of 75 pixels was chosen. Motion upsampling rate was set to 200 Hz in accordance with the hexapod encoder sampling rate.

3.3. Laser ablation framework

To demonstrate vision-guided laser control, ablation trials were conducted on MDF and EVT. Therefore, the experimental setup in Fig. 6a additionally comprises a surgical Er:YAG laser ($\lambda = 2.94\,\mu\mathrm{m}$, DPM-15, Pantec Engineering AG, Ruggell, Liechtenstein) and a three-axis scanning unit (VarioScan and HurryScan, SCANLAB, Puchheim, Germany). The camera field of view is optimized with respect to the area of intersection between tissue surface and laser scanning range, which is defined by a cube of 10 mm in each direction. To estimate the laser-to-camera transform ${}^{\mathrm{L}}\boldsymbol{T}_{\mathrm{A}}$ (see Fig. 6a), circle grids are ablated onto a planar surface and detected in the stereo view. After back-projection to object space, the laser axis is computed by principal component analysis (PCA) and laser workspace orientation is estimated by point-based registration (Schoob et al., 2015b).

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 state-of-the-art systems, e.g., the Digital AcuBlade™ Scanning Micromanipulator (Lumenis, Yokneam, Israel). 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 $v_{\mathrm{S}} = 200\,\mathrm{mm\,s^{-1}}$ 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)[3] (Quigley et al., 2009).

3.4. Laser ablation trials on MDF

Straight and curved lines were stamped with green ink onto the MDF sample surface that was positioned in the laser focal range using the hexapod platform (see Fig. 6b). Tracking and ablation were simultaneously performed considering both the lateral and axial motion patterns. The root mean square error (RMSE) of path

[3] http://www.ros.org/

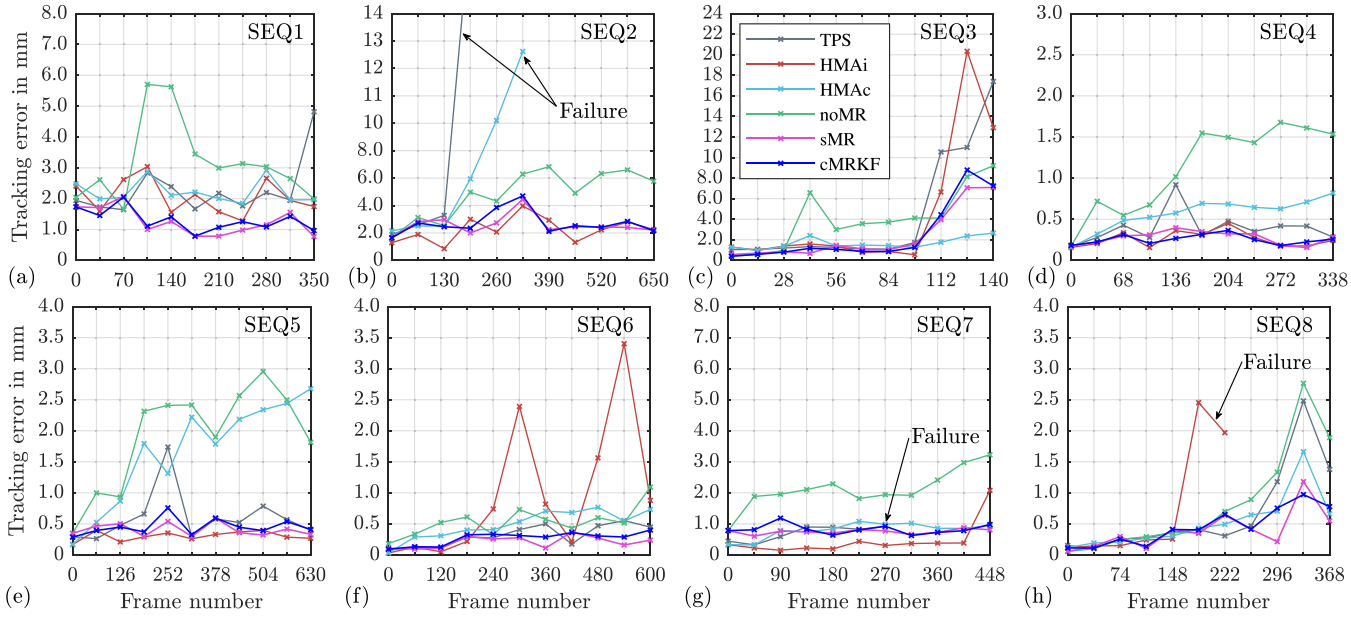

Fig. 7. Tracking error (RMSE) in mm for analyzed methods and IVT sequences SEQ1–8 (a–h). Accuracy is evaluated with respect to ground truth measured for eleven particular frames per sequence (distributed equally along each sequence).

tracing was computed between the initial shape and associated ablation, both segmented by thresholding. Additionally, laser ablation on a static sample was performed to quantify the impact of the laser-to-camera registration error.

3.5. 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 deploying a stylus-based tablet interface (Schoob et al., 2015c). Laser 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 tissue sample is moved in the axial direction. Finally, the ablated paths were analyzed under microscopic imaging in terms of 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 Fig. 6c, 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.

4. Results

4.1. Tracking accuracy assessment on IVT

Fig. 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 2.

In comparison with TPS, HMAi, HMAc, and noMR, the results demonstrate superior performance of the tracking with mesh refinement when either sMR or cMRKF is applied. Since we observed no differences between cMR and cMRKF during the IVT validation, we skip presenting the results of cMR 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 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 Fig. 7). In addition, as the supplemental video highlights, 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 Fig. 8a, where the instrument tip enters the tracked region. When the MHD-based detection scheme (see Fig. 8b) is incorporated into the reweighting process according to Eq. (33), robustness to partial occlusions, such as those caused by instruments or laser ablation with significant carbonization, can be improved. In comparison with method noMR (i.e. SEQ3–5), drift is successfully eliminated. Decomposing the RMSE for the cMRKF method reveals that the error in z-direction predominates. For sequence SEQ1, the RMSE amounts to $(e_x, e_y, e_z) = (0.22, 0.26, 1.31)$ mm, revealing that the z-error is approximately five times higher compared with that in the other two spatial directions. Since the depth resolution increases with the baseline-to-distance ratio, a reduced predominance in the z-direction is revealed for SEQ4 and SEQ6, yielding $(e_x, e_y, e_z) = (0.10, 0.10, 0.22)$ mm and $(e_x, e_y, e_z) = (0.08, 0.12, 0.25)$ mm, respectively.

Computation time is exemplarily discussed for SEQ8, which exhibits significant deformation, and is depicted in Fig. 9 and Table 3. Initially, the TPS method runs at almost constant 13 ms per frame, since only slight motion occurs, which can be tracked within a few iterations. Since tracking of tissue deformation requires a higher

Table 2
Tracking error in mm measured for the IVT sequences. Accuracy values are valid until tracking failure (F). **Bold** numbers represent the best performance, whereas italic numbers denote the second best performance.

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

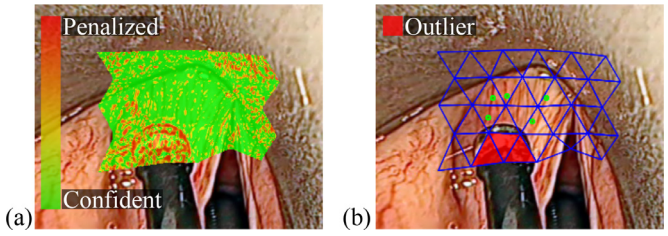

Fig. 8. Instrument-induced partial occlusion in SEQ7 penalized by implementing (a) norm-like Huber function and (b) MHD-based detection scheme.

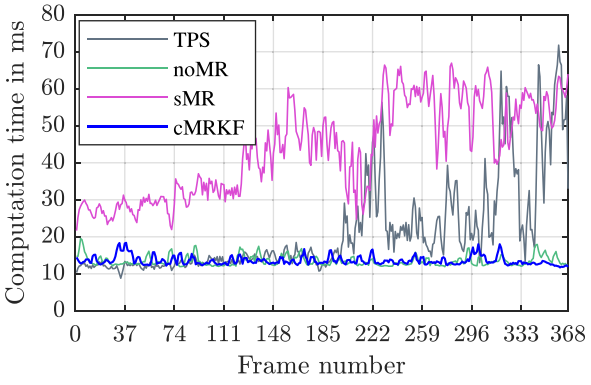

Fig. 9. Runtime in milliseconds for SEQ8 (8 × 4 triangle mesh).

number of iterations to converge, the computation time drastically increases up to 70 ms. Even though it is more accurate, the sMR method is as non-deterministic as the TPS approach. This observation is substantiated by the high standard deviation of 14.5 ms (see Table 3).

Table 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

Even though no restriction on the computation time was imposed to ensure convergence, the 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 ± 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, we were not able to reproduce the HMA runtime reported in the original work; thus, we assume an average time of 50 ± 20 ms per frame, as presented in (Puerto-Souza and Mariottini, 2013). 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 Figs. 10, 11, 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.

4.2. 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

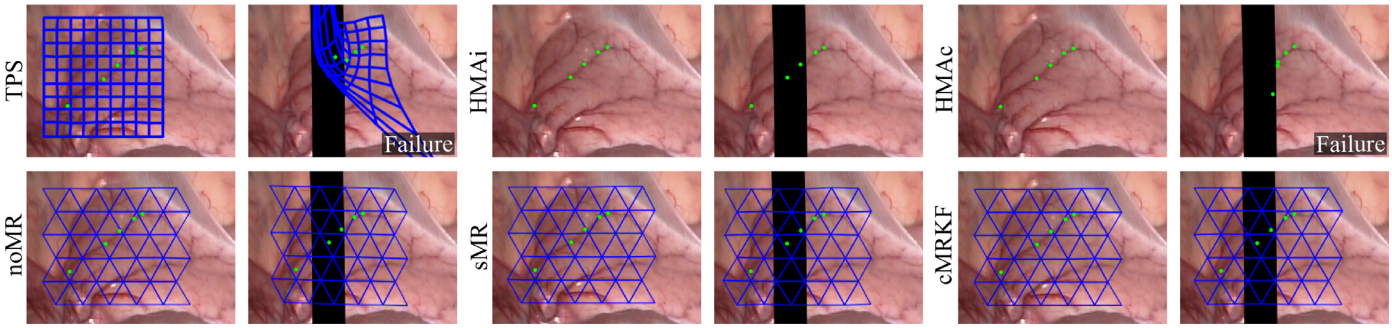

Fig. 10. Comparison of tracking results in laparoscopic sequence SEQ2. For each sequence, frame 2 (left) and frame 254 (right) are shown. Five landmarks (green dots) are tracked with respect to ground truth. A partial occlusion is simulated with a synthesized vertical bar moving from left to right. Tracking failure was detected for methods TPS and HMAc. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

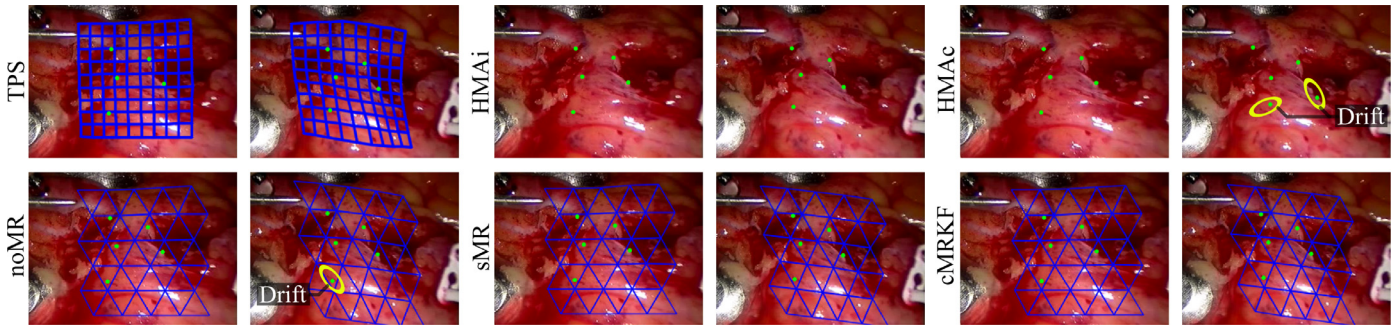

Fig. 11. Comparison of tracking results in beating heart sequence SEQ4. For each sequence, frame 2 (left) and frame 323 (right) are shown. Five landmarks (green dots) are tracked with respect to ground truth. Drift of certain landmarks was observed for methods HMAc and noMR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

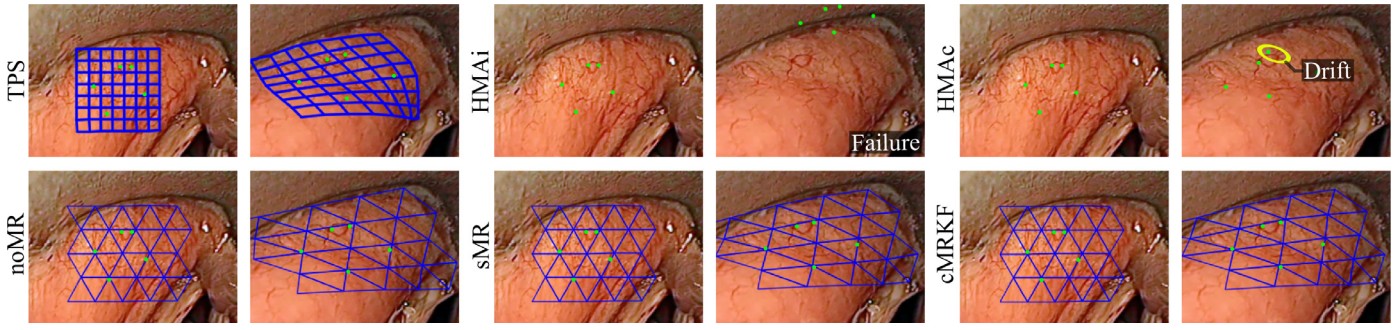

Fig. 12. Comparison of tracking results in laryngeal sequence SEQ8. For each sequence, frame 2 (left) and frame 216 (right) are shown. Five landmarks (green dots) are tracked with respect to ground truth. Drift was observed for method HMAc. Tracking failure was detected for method HMAi. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results of tracking the lateral MDF motion are shown in Fig. 13, including the position over time, the associated ground truth and the tracking error. 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 Fig. 13a,b). By contrast, tracking with subsequent mesh refinement (sMR) compensates for drift; however, it drastically increases the tracking error (see Fig. 13c,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 for noMR whith simultaneous compensation for drift is achieved (see Fig. 13e,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 Fig. 13g,h.

The error curves of Fig. 13 are summarized in the form of box plots (see Fig. 14a) (Mc Gill et al., 1978). The interquartile range (IQR), defined from the bottom to the top of the box, contains data points between the 25th and the 75th percentile, respectively, whereas the error median is represented by the notch. The upper whisker includes data within 1.5 IQR of the upper quartile, whereas the lower whisker contains data within 1.5 IQR of the lower quartile. Outliers are marked by a cross if they are not between the whiskers. 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 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 and 1.29 mm in the lateral and 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 an 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 4, are higher than those in the case of tracking on the MDF specimen.

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 × 200 pixels, the associated computation time for tracking is listed in Table 5. In particular, for the PFN optimization scheme (Algorithm 1), 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 × 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.

4.3. Laser ablation trials on MDF

The results of path tracing on the MDF specimen are listed in Table 6. Regarding the static scenario, the ablation misalignment is below 0.07 mm, which correlates with the laser-to-camera registration error reported in our previous study (Schoob et al., 2015b). In accordance to the cMRKF-based tracking error presented

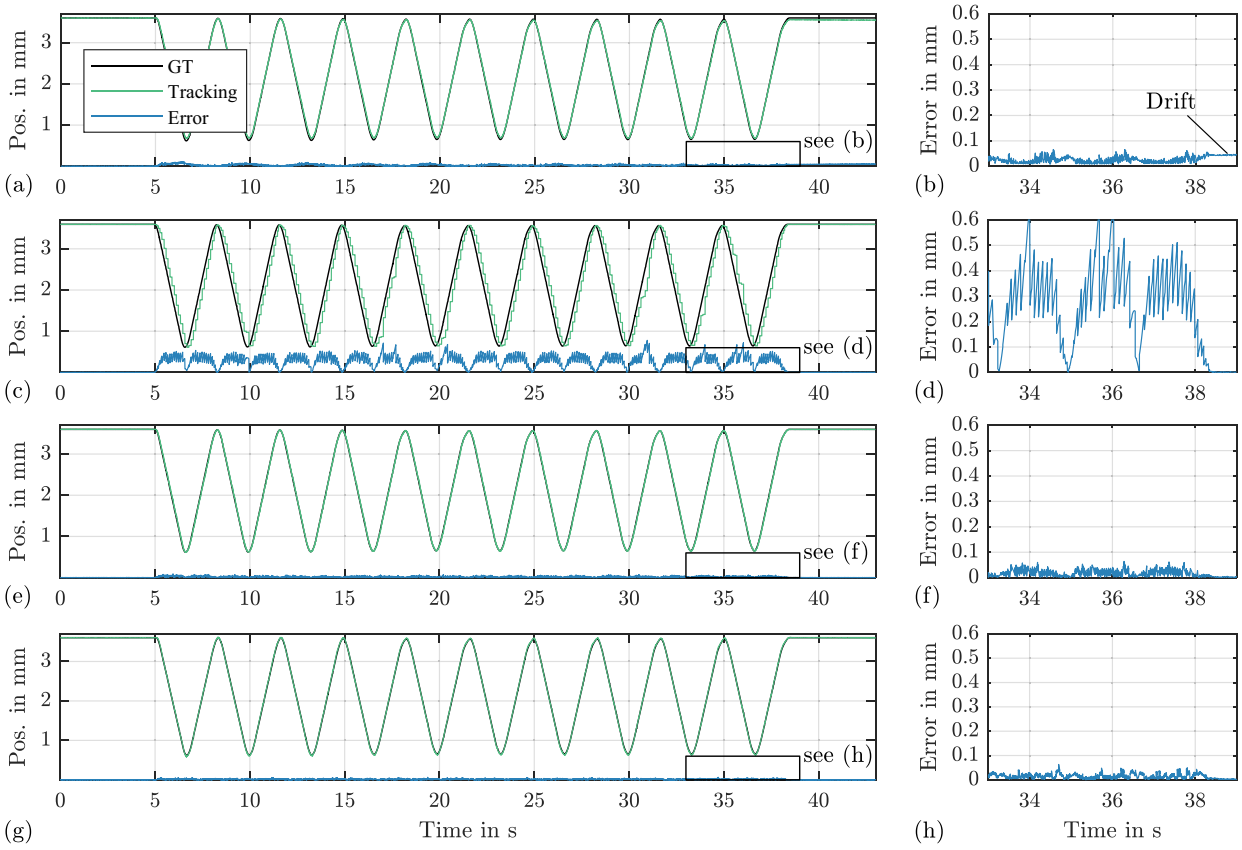

Fig. 13. Results of tracking the MDF sample for the lateral motion pattern. The position and associated ground truth (GT) trajectory plotted over time for methods (a) noMR, (c) sMR, (e) cMR, and (g) cMRKF. The latency-dependent tracking error is additionally shown in the associated zoomed view.

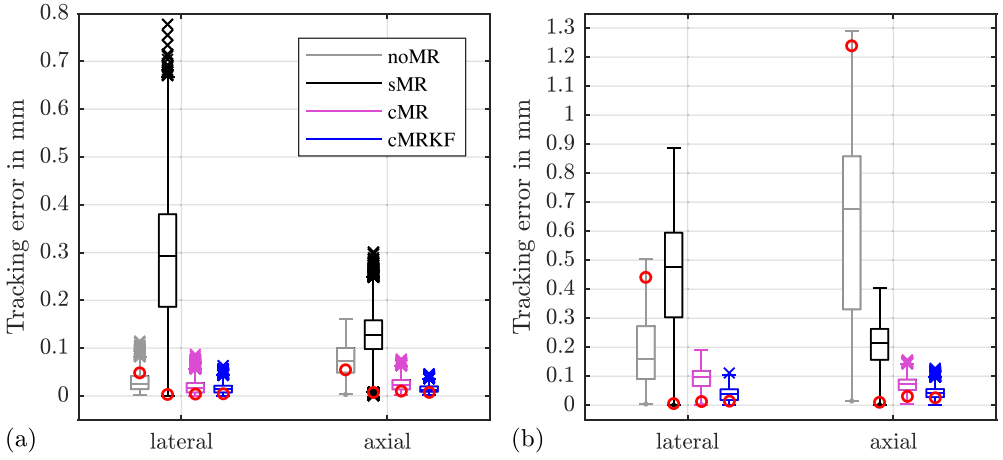

Fig. 14. Box plot illustrating the tracking error measured for (a) the MDF specimen and (b) the porcine EVT sample. For each method, the results of the lateral and axial motion pattern are shown. The final misalignment (drift) after returning to the hexapod home position is indicated by a red circle. Outliers are marked by the symbol × . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 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. **Bold** numbers represent the best performance, whereas Italic numbers denote the second best performance.

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

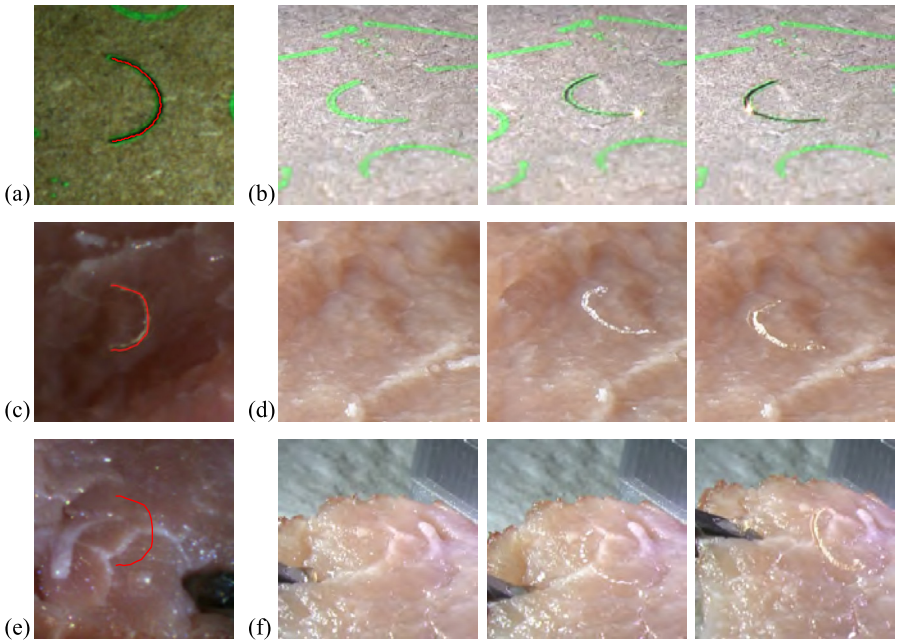

Fig. 15. Laser ablation of the curved scan pattern. In the first row, results of ablation onto the MDF specimen while moving into lateral direction are shown by (a) the left camera view including segmented incision (red line) and by (b) three snapshot images of the motion sequence acquired with an external, 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Computation time in ms as a function of the horizontal triangle edge length when tracking an image region of 200 × 200 pixels.

Triangle edge length (px)	25	50	75	100
Mesh dimension	16 × 12	8 × 6	6 × 4	4 × 3
Tracking (Alg. 1)	751.1	28.6	9.3	6.3
Refinement (Alg. 2)	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

Table 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

Table 7
Videos demonstrating tracking on MDF and soft tissue (IVT, EVT).

Name	Description
IVT	Tracking on IVT datasets
MDF	Ablation on MDF under lateral and axial motion
EVT–1	Ablation on EVT under lateral and axial motion
EVT–2	Ablation on EVT comparing sMR and cMRKF
EVT–3	Ablation on EVT comparing w/o and w/ laser focusing
EVT–4	Ablation on EVT while manipulating with forceps

in Table 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 Fig. 15a, three snapshot images of acquired video sequence are depicted in Fig. 15b, clearly illustrating the progressively ablated incision. Microscopic images of the straight and curved line, demonstrating precise path tracing, are provided in Fig. 16a. The aforementioned ablation trials are included in the supplemental video material (see Table 7).

4.4. Laser ablation trials on EVT

The results of path tracing on porcine EVT are listed in Table 6. The associated snapshots of the lateral motion sequence are shown in Fig. 15c,d. Compared with the MDF trials, the increase in the ablation error correlates with the larger tracking deviation, as listed in Table 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 Fig. 16b,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 Fig. 16d. 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 our tracking-based results in Fig. 16c with laser ablation without image-based focus adjustment when moving in the axial direction. As illustrated in Fig. 16e, carbonization at the incision edges can be observed as a result of non-optimal energy exposure to the tissue. This may significantly influence the desired incision quality and lead to increased trauma and healing duration of the tissue. Thus, proper focusing by maintaining constant distance to the tissue is mandatory for laser surgery in a dynamic soft tissue environment.

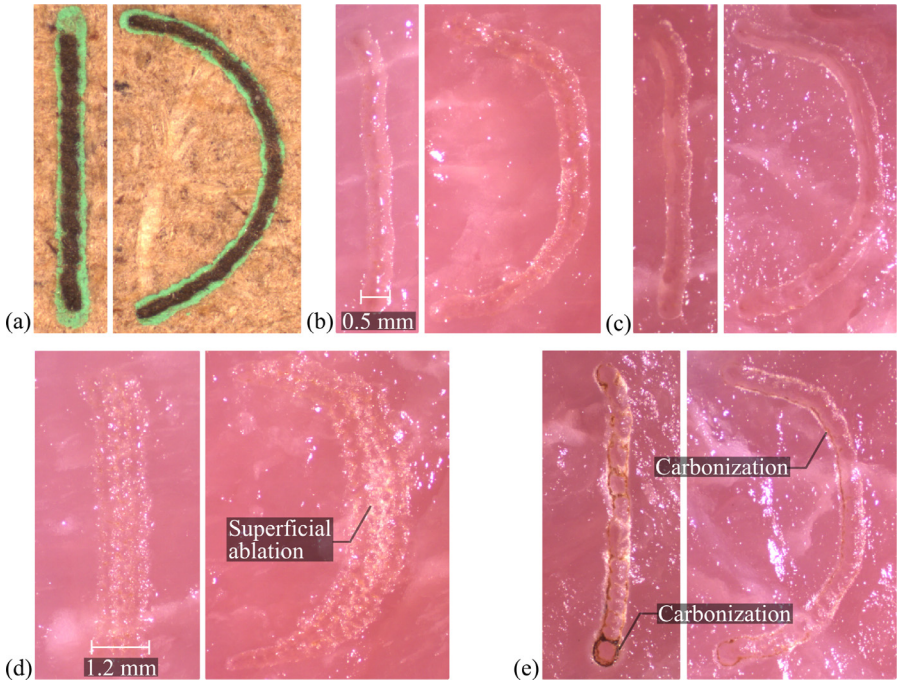

Fig. 16. Qualitative results of the laser ablation trials on the MDF and the EVT sample under lateral (a,b) and axial movements (c) deploying concurrent tracking scheme cMRKF. In comparison with (b), sequential mesh refinement (sMR), as depicted in (d), leads to poor incision quality characterized by 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.

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 Fig. 15e–f). Exemplary sequences are captured and included in the supplemental video material (see Table 7). The associated microscopic examination clearly indicates that improved incision quality can be reproduced even on tissue undergoing deformation.

5. Conclusion

In this article, 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 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 that of the approach presented previously (Schoob et al., 2016). In addition, an experimental design is described to assess the latency-dependent tracking misalignment. 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 can be successfully integrated into the visual feedback loop, facilitating online adjustment of the desired 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 will focus on the investigation of different online-capable feature detection and matching techniques extending our method and allowing for global retargeting of the tracked region after total occlusion or re-entering into the field of view. Furthermore, transfer of the framework from the presented lab setup to an endoscopic laser system is required for clinical trials.

Acknowledgements

The research received funding from the European Seventh Framework Programme FP7-ICT under grant agreement µRALP - no 288663. We thank Giorgio Peretti from the Department of Otorhinolaryngology, University of Genoa, Italy, for providing *in vivo* laryngeal data.

Electronic supplementary material

The online version contains supplementary video material.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2017.06.004

References

Baker, S., Gross, R., Matthews, I., Ishikawa, T., 2003. Lucas-Kanade 20 Years On: a Unifying Framework: Part 2. Tech. rep. cmu-ri-tr-03-01. Robotics Institute, Pittsburgh, PA.

Bouguet, J.-Y., 2000. Pyramidal Implementation Of The Lucas Kanade Feature Tracker. Tech. rep. Microprocessor Research Labs, Intel Corporation.

Brunet, F., Gay-Bellile, V., Bartoli, A., Navab, N., Malgouyres, R., 2011. Feature-driven direct non-rigid image registration. Int. J. Comput. Vis. 93 (1), 33–52.

Chang, P.-L., Stoyanov, D., Davison, A.J., 2013. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In: Med Image Comput Comput Assist Interv, pp. 42–49.

Collins, T., Bartoli, A., Bourdel, N., Canis, M., 2016. Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos. In: Med Image Comput Comput Assist Interv, pp. 404–412.

Dagnino, G., Mattos, L., Caldwell, D., 2015. A vision-based system for fast and accurate laser scanning in robot-assisted phonomicrosurgery. Int. J. Comput. Assist. Radiol. Surg. 10 (2), 217–229.
Du, X., Clancy, N., Arya, S., Hanna, G., Kelly, J., Elson, D., Stoyanov, D., 2015. Robust surface tracking combining features, intensity and illumination compensation. Int. J. Comput. Assist. Radiol. Surg. 10 (12), 1915–1926.
Dutter, R., Huber, P.J., 1981. Numerical methods for the nonlinear robust regression problem. J. Stat. Comput. Simul. 13 (2), 79–113.
Filzmoser, P., Ruiz-Gazen, A., Thomas-Agnan, C., 2013. Identification of local multivariate outliers. Stat. Pap. 55 (1), 29–47.
Giannarou, S., Visentini-Scarzanella, M., Yang, G.-Z., 2013. Probabilistic tracking of affine-invariant anisotropic regions. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 130–143.
Hager, G.D., Belhumeur, P.N., 1998. Efficient region tracking with parametric models of geometry and illumination. IEEE Trans. Pattern Anal. Mach. Intell. 20 (10), 1025–1039.
Haouchine, N., Cotin, S., Peterlik, I., Dequidt, J., Lopez, M.S., Kerrien, E., Berger, M.-O., 2015. Impact of soft tissue heterogeneity on augmented reality for liver surgery. IEEE Trans. Vis. Comput. Graph. 21 (5), 584–597.
Lau, W., Ramey, N., Corso, J., Thakor, N., Hager, G., 2004. Stereo-based endoscopic tracking of cardiac surface deformation. In: Med Image Comput Comput Assist Interv., vol. 3217, pp. 494–501.
Mattos, L.S., Deshpande, N., Barresi, G., Guastini, L., Peretti, G., 2014. A novel computerized surgeon-machine interface for robot-assisted laser phonomicrosurgery. Laryngoscope 124 (8), 1887–1894.
Mc Gill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of box plots. Am. Stat. 32 (1), 12–16.
Mountney, P., Stoyanov, D., Yang, G.-Z., 2010. Three-dimensional tissue deformation recovery and tracking. IEEE Signal Process. Mag. 27 (4), 14–24.
Ortmaier, T., Gröger, M., Boehm, D.H., Falk, V., Hirzinger, G., 2005. Motion estimation in beating heart surgery. IEEE Trans. Biomed. Eng. 52 (10), 1729–1740.
Pilet, J., Lepetit, V., Fua, P., 2008. Fast non-rigid surface detection, registration and realistic augmentation. Int. J. Comput. Vis. 76 (2), 109–122.
Preiswerk, F., De Luca, V., Arnold, P., Celicanin, Z., Petrusca, L., Tanner, C., Bieri, O., Salomir, R., Cattin, P.C., 2014. Model-guided respiratory organ motion prediction of the liver from 2d ultrasound. Med. Image Anal. 18 (5), 740–751.
Prokopetc, K., Bartoli, A., 2016. A comparative study of transformation models for the sequential mosaicing of long retinal sequences of slit-lamp images obtained in a closed-loop motion. Int. J. Comput. Assist. Radiol. Surg. 1–10.
Puerto-Souza, G., Mariottini, G.-L., 2013. A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images. IEEE Trans. Med. Imaging 32 (7), 1201–1214.
Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., 2009. Ros: an open-source robot operating system. IEEE Int Conf Robot Autom – Workshop on Open Source Software.
Renevier, R., Tamadazte, B., Rabenorosoa, K., Tavernier, L., Andreff, N., 2016. Endoscopic laser surgery: design, modeling and control. IEEE ASME Trans. Mechatron. PP (99). 1–1
Richa, R., Poignet, P., Liu, C., 2010. Three-dimensional motion tracking for beating heart surgery using a thin-plate spline deformable model. Int. J. Rob. Res. 29 (2–3), 218–230.
Royer, L., Krupa, A., Dardenne, G., Le Bras, A., Marchand, E., Marchal, M., 2017. Real-time target tracking of soft tissues in 3d ultrasound images based on robust visual information and mechanical simulation. Med. Image Anal. 35, 582–598.
Rubinstein, M., Armstrong, W., 2011. Transoral laser microsurgery for laryngeal cancer: a primer and review of laser dosimetry. Lasers Med. Sci. 26 (1), 113–124.
Sauvée, M., Poignet, P., Triboulet, J., Dombre, E., Malis, E., Demaria, R., 2006. 3D heart motion estimation using endoscopic monocular vision system. Model. Control Biomed. Syst. 6, 141–146.
Schoob, A., Kundrat, D., Kahrs, L., Ortmaier, T., 2015. Comparative study on surface reconstruction accuracy of stereo imaging devices for microsurgery. Int. J. Comput. Assist. Radiol. Surg. 1–12.
Schoob, A., Kundrat, D., Kleingrothe, L., Kahrs, L., Andreff, N., Ortmaier, T., 2015. Tissue surface information for intraoperative incision planning and focus adjustment in laser surgery. Int. J. Comput. Assist. Radiol. Surg. 10 (2), 171–181.
Schoob, A., Laves, M.-H., Kahrs, L.A., Ortmaier, T., 2016. Soft tissue motion tracking with application to tablet-based incision planning in laser surgery. Int. J. Comput. Assist. Radiol. Surg. 1–13.
Schoob, A., Lekon, S., Kundrat, D., Kahrs, L.A., Mattos, L.S., Ortmaier, T., 2015. Comparison of tablet-based strategies for incision planning in laser microsurgery. SPIE Med. Imaging Int. Soc. Opt. Photonics. 94150J–94150J
Shi, J., 1994. Good features to track. In: Computer Vision and Pattern Recognition,1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. IEEE, pp. 593–600.
Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE Trans Med Imaging 32 (7), 1153–1190.
Stoyanov, D., Darzi, A., Yang, G.Z., 2004. Dense 3D Depth Recovery for Soft Tissue Deformation During Robotically Assisted Laparoscopic Surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 41–48.
Stoyanov, D., Mylonas, G.P., Deligianni, F., Darzi, A., Yang, G.Z., 2005. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In: Med Image Comput Comput Assist Interv, pp. 139–146.
Stoyanov, D., Rayshubskiy, A., Hillman, E., 2012. Robust registration of multispectral images of the cortical surface in neurosurgery. In: Biomedical Imaging (ISBI),2012, 9th IEEE International Symposium on. IEEE, pp. 1643–1646.
Stoyanov, D., Yang, G.-Z., 2009. Soft tissue deformation tracking for robotic assisted minimally invasive surgery. In: Engineering in Medicine and Biology Society, 2009, pp. 254–257. EMBC 2009. Annual International Conference of the IEEE. IEEE
Suwelack, S., Röhl, S., Bodenstedt, S., Reichard, D., Dillmann, R., dos Santos, T., Maier-Hein, L., Wagner, M., Wünscher, J., Kenngott, H., 2014. Physics-based shape matching for intraoperative image guidance. Med. Phys. 41 (11), 111901.
Tan, D.J., Holzer, S., Navab, N., Ilic, S., 2014. Deformable template tracking in 1ms. British Mach. Vis. Conf.
Tang, H.-W., Brussel, H.V., Sloten, J.V., Reynaerts, D., De Win, G., Cleynenbreugel, B.V., Koninckx, P.R., 2006. Evaluation of an intuitive writing interface in robot-aided laser laparoscopic surgery. Comput. Aided Surg. 11 (1), 21–30.
Tsai, R.Y., Lenz, R.K., 1989. A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. IEEE Trans. Rob. Autom. 5 (3), 345–358.
Vaudrey, T., Badino, H., Gehrig, S., 2008. Integrating disparity images by incorporating disparity rate. In: International Workshop on Robot Vision, pp. 29–42.
Wong, W.-K., Yang, B., Liu, C., Poignet, P., 2013. A quasi-spherical triangle-based approach for efficient 3-d soft-tissue motion tracking. IEEE ASME Trans. Mechatron. 18 (5), 1472–1484.
Yang, B., Wong, W.-K., Liu, C., Poignet, P., 2014. 3D soft-tissue tracking using spatial–color joint probability distribution and thin-plate spline model. Pattern Recognit. 47 (9), 2962–2973.
Yang, S., Lobes, L.A., Martel, J.N., Riviere, C.N., 2015. Handheld-automated microsurgical instrumentation for intraocular laser surgery. Lasers Surg. Med. 47 (8), 658–668.
Ye, M., Giannarou, S., Meining, A., Yang, G.-Z., 2016. Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. Med. Image Anal. 30, 144–157.
Yip, M., Lowe, D., Salcudean, S., Rohling, R., Nguan, C., 2012. Tissue tracking and registration for image-guided surgery. IEEE Trans. Med. Imaging 31 (11), 2169–2182.
Zabih, R., Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: Comput Vis ECCV, vol. 801. Springer, pp. 151–158.
Zhu, J., Gool, L.V., Hoi, S.C.H., 2009. Unsupervised face alignment by robust nonrigid mapping. In: IEEE Int Conf Comput Vis, pp. 1265–1272.
Zhu, J., Lyu, M., Huang, T., 2009b. A fast 2d shape recovery approach by fusing features and appearance. IEEE Trans. Pattern. Anal. Mach. Intell. 31 (7), 1210–1224.