A Methodology and Clinical Dataset with Ground-truth to Evaluate Registration Accuracy Quantitatively in Computer-assisted Laparoscopic Liver Resection

N. Rabbani[a] ; L. Calvet[a,b,d]; Y. Espinel[a]; B. Le Roy[a,c]; M. Ribeiro[a,b]; E. Buc[a,b] and A. Bartoli[a]

[a]EnCoV, Institut Pascal, Clermont-Ferrand; [b]CHU Clermont-Ferrand; [c]CHU Saint-Etienne; [d]IRIT, University of Toulouse

ARTICLE HISTORY
Compiled September 27, 2021

ABSTRACT
Augmented Reality (AR) can assist Laparoscopic Liver Resection (LLR) by registering a preoperative 3D model with laparoscopic images. Evaluating the accuracy of the registration methods requires measuring Target Registration Error (TRE). Previous work evaluates TRE on simulated, phantom and animal data but not on clinical data. Our contribution is a methodology for groundtruth acquisition using Laparoscopic Ultrasound (LUS) in clinical LLR, two evaluation criteria, a four-patient dataset and an evaluation of two existing registration methods. We acquire groundtruth by first calibrating a LUS probe. During surgery, the LUS is coregistered with the laparoscope and its observations made transferable to 3D laparoscope coordinates. We propose two evaluation criteria: an inclusion criterion and a measure of TRE. The inclusion criterion is binary: it is passed if and only if all the LUS tumour profiles lie within the registration-predicted tumour augmented by the oncologic margin of 1 cm. The TRE is computed as the average minimal distance between each LUS tumour profile and the registration-predicted tumour volume. The average position error in our LUS registered images is estimated as about 1mm, which is far better than the measured errors for the state-of-the-art registration methods, making our dataset relevant for their evaluation. We ran a preliminary evaluation of two registration methods. Both methods failed the inclusion criterion for all the patients. The TRE measurements show that the registration-predicted tumours are out of the 1 cm oncologic margin with reasonable standard deviation over the dataset. We conclude that improvements in the accuracy of registration methods is needed for accurate gesture guidance. Patients will be added to our dataset; together with our methodology, they will form a benchmark for future registration methods. The dataset is available publicly at `http://igt.ip.uca.fr/~ab/code_and_datasets/datasets/llr_reg_evaluation_by_lus`.

KEYWORDS
Computer-Assisted Intervention, Laparoscopic Liver Resection, Augmented Reality, Target Registration Error, Ultrasound Probe Calibration

1. Introduction

Laparoscopic Liver Resection (LLR) is challenging because of the hidden intra-parenchymatous structures. Augmented Reality (AR) is a promising assistance

CONTACT N. Rabbani. Email: navid_rabbani@yahoo.com

method, which works by registering a 3D model reconstructed from preoperative CT onto the laparoscopic images. The registration must align and deform the preoperative 3D model adequately. Evaluating the registration quantitatively is challenging, because groundtruth is not available in clinical conditions. Previous registration methods were thus only evaluated quantitatively on simulated, phantom and animal data. We propose a methodology for the acquisition of groundtruth using Laparoscopic Ultrasound (LUS) in clinical LLR, an initial dataset for four patients, two evaluation criteria and a preliminary evaluation of two existing registration methods.

Research in computer-aided laparoscopy mainly concerns 3D reconstruction (Maier-Hein et al. 2014) and preoperative 3D model registration, which is a requirement to enable AR. It is especially challenging in LLR, because the liver has a high degree of deformation. The existing solution methods deform the preoperative 3D model to fit the target laparoscopic images under biomechanical constraints (Adagolodjo et al. 2017; Koo et al. 2017; Haouchine et al. 2016; Robu et al. 2018).

Evaluating the technical performance of the registration methods requires one to measure a Target Registration Error (TRE). Ideally, this would be the discrepancy between the actual location of an intraparenchymatous tumour and its prediction by the registration. Previous work uses four types of evaluation setups. The first type is the in-silico setup, which involves a simulated surgical setup. The second type is the phantom setup, which is a physical replica of the organ that is typically 3D printed. The third type is the animal setup, which uses animal parts (hence, ex-vivo) or live animals (hence, in-vivo). The fourth type is the clinical setup, involving real surgical data. Each type of setups has pros and cons, in terms of realism and complexity. Clearly, none of the first three types may replace a full-fledged evaluation in clinical conditions, which appears as the ultimate evaluation setup. However, owing to the difficulty of measuring the groundtruth location of intraparenchymatous structures, previous work has not reported any TRE measurements in clinical conditions. Instead, the results in clinical conditions are typically qualitative (Adagolodjo et al. 2017; Haouchine et al. 2016; Koo et al. 2017; Robu et al. 2018) or represent a Fiducial Registration Error (FRE) (Thompson et al. 2018), which is known to underestimate TRE. In addition, most datasets are kept private, which does not allow the comparison of methods on standardised grounds. Table 1 summarises the public datasets. We observe that there does not exist a dataset with groundtruth for the position of intraparenchymatous structures obtained in clinical conditions. Our goal in this paper is to describe a methodology and preliminary results towards fulfilling this need in LLR.

Table 1. Public datasets with ground-truth. The last column indicates if a standardised evaluation criterion and script are provided.

Dataset	*Data type*	*Groundtruth*	*Landmarks*	*Script*
Maier-Hein et al. (2014) (Open-CAS)	Ex-vivo animal	3D reconstruction	External	Yes
DePoLL (Modrzejewski et al. 2019)	In-vivo animal	Registration	External & internal	Yes
Suwelack et al. (2014) (Open-CAS)	In-silico & phantom	Registration	External & internal	No
Proposed	Clinical	Registration	Internal	Yes

We bring five main contributions. First, we propose a complete framework for data collection, with LUS probe tracking and calibration, from which LUS observations are transferable to 3D laparoscope coordinates. Second, we extend a reference US calibration method to include the optimisation of camera poses and reprojection error. Third, we propose a novel calibration phantom design which outperforms previous ones, by modelling and minimising the effect of uncertainty. Fourth, equipped with the above framework and methods, we propose the first clinical dataset with groundtruth

location obtained from LUS for intra-parenchymatous structures in LLR. Fifth, we propose criteria to evaluate TRE for these structures when predicted from preoperative image data and an evaluation of existing augmented reality methods. Our dataset and evaluation script form the first means to evaluate preoperative data registration in LLR and in laparoscopy in general. They were made publicly available for the benefit of the research community

The paper starts with a review of previous work in section 2, followed by the proposed method in section 3, data collection and experimental results in section 4 and conclusions in section 5.

2. Previous Work

Registration evaluation. TRE is the commonly agreed evaluation metric in registration problems. However, there is a lack of datasets for its computation in clinical LLR, mostly because localising the target intra-parenchymatous liver structures is challenging. Existing work (Plantefève et al. 2016; Koo et al. 2017; Haouchine et al. 2016) use in-silico evaluation, which suffers from lack of realism of the organ biomechanics and numerical approximations in biomechanical simulations. Espinel et al. (2020) use a fabricated phantom, which also suffers lack of realism. Thompson et al. (2018); Modrzejewski et al. (2019) propose ex-vivo and in-vivo evaluations on animals. In (Modrzejewski et al. 2019), metal markers are placed on the organ and used for TRE computation. It is a valuable effort because the animal organs are more realistic than phantoms. However, the conditions of LLR performed on animals remain different in terms of organ access and visibility, and organ anatomy compared to LLR performed on humans. Thompson et al. (2018) propose an evalution on patient data. They however do not compute TRE. Instead, they measure the visible misalignment, whose correlation to the real TRE is debatable. There was another attempt to evaluate liver registration in (Clements et al. 2016), in the context of open surgery. In this case, the inner organs are not occluded, and the authors could use a laser range scanner to produce an intraoperative 3D model of the liver, to which they registered the preoperative 3D model obtained by segmenting CT scans (Dumpuri et al. 2010; Cash et al. 2007, 2003). For evaluation, they used US images which are coregistered with the intraoperative model. The US probe was tracked using optical trackers. Using laser range scanners and optical trackers is not possible in laparoscopic surgery because one cannot infer the relative pose between these devices inside the abdomen from the outside pose. The evaluation was done using the mean closest-point distances between the feature contours from the intraoperative US images and the preoperative model, which represents an oversimplification of the TRE.

Having ground-truth of the hidden intra-parenchymatous structures at disposal in real surgery conditions is yet highly desirable. These data can only be acquired during surgery. Bernhardt et al. (2015) used intraoperative CT registered to laparoscopic images. However, they did not evaluate their groundtruth accuracy, neither did they publicly release a dataset and evaluation methodology.

LUS probe tracking. Several methods were proposed to localise and track medical devices in laparoscopic surgery. Liu et al. (2014); Cheung et al. (2010); Hayashi et al. (2015) use electromagnetic trackers. However, this technology is not well adapted to the operating theatre because of the presence of ferromagnetic objects distorting the electromagnetic field and leading to localisation errors (Poulin and Amiot 2002).

Optical trackers are also used (Cenni et al. 2016; Carbajal et al. 2013; Poon and Rohling 2005; Wang et al. 2018). They are accurate and commercial solutions are now available. However, they are not suitable for obtaining the LUS probe head pose in our context. Indeed, the LUS probes used in LLR are generally articulated, as their head is mechanically controlled by the surgeon and, being located inside the cavity, kept invisible from the tracker.

LUS probe calibration. The general US calibration problem is to compute the location of each US pixel in some probe coordinate system. A thorough review of the literature was conducted (Mercier et al. 2005; Hsu et al. 2009). Calibration methods use dedicated phantoms whose geometry must be known with high precision. The phantom must include two key features. First, they must allow one to coregister the phantom and the US probe, for example by means of checkerboards usable in pose computation. Second, they must provide sharp signal changes in the US image, for example by means of wires visible in the US plane. The calibration methods mainly differ in terms of phantom shapes, including for example points, wires and planes. The existing methods lack a systematic evaluation of the uncertainty and design optimisation towards optimal calibration results.

3. Proposed Evaluation Methodology and Data Collection Methods

We acquire groundtruth using LUS, which we coregister with the laparoscope, as shown in figure 1. For that, we use a printed checkerboard sticker on the distal end of the LUS probe, defining the LUS coordinates. Our setup has three main steps. First, we reconstruct a 3D model of the sticker. Second, we calibrate the probe using a 3D printed phantom inspired by (Lasso et al. 2014). Third, we compute the LUS pose, which is the transformation between the laparoscope and the LUS probe.

The 3D reconstruction of the probe head sticker uses SfM (AliceVision-Meshroom 2021) with a high end camera. The image checkerboard corners are manually found and refined to subpixel accuracy (Bouguet 2000). The corners 3D positions are then computed by triangulation (Hartley and Zisserman 2003). Both camera poses and corners 3D positions are refined by bundle adjustment (Hartley and Zisserman 2003).

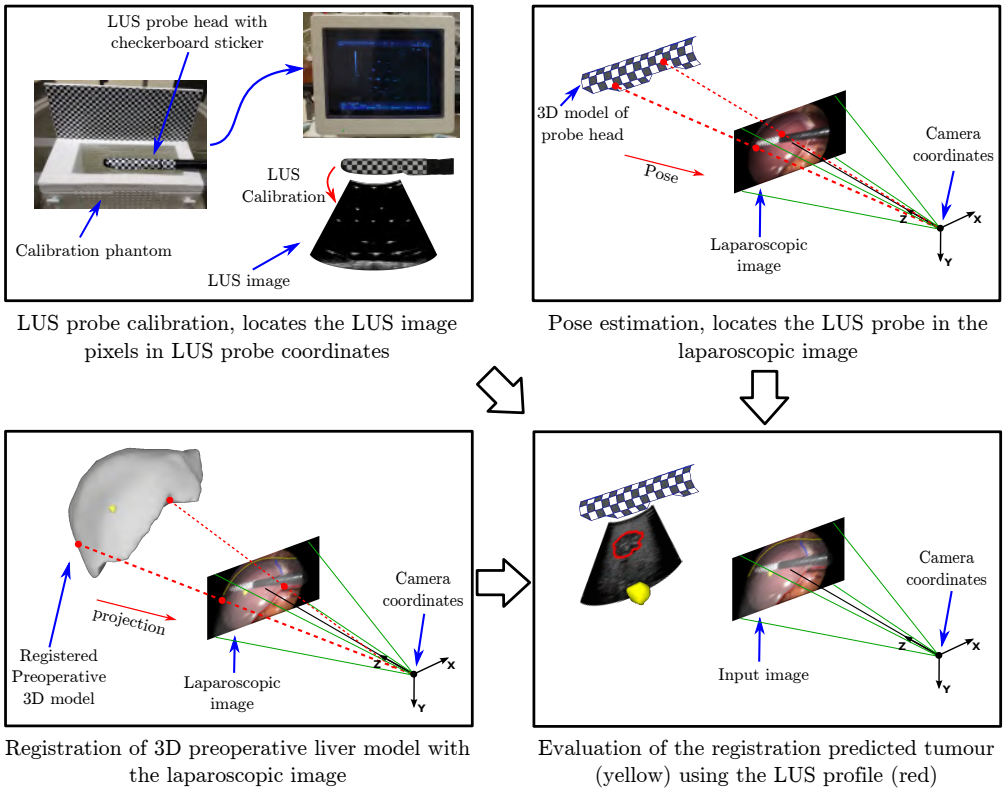
3.1. Probe Calibration

The setup used for probe calibration is shown in figure 2a. We use a high end camera to film both the probe head and the phantom. A planar checkerboard of known geometry is stuck on the phantom. Images are acquired from viewpoints ensuring that the probe head and the checkerboard are both visible and can thus be coregistered.

We consider a 3D point of phantom coordinates $\mathbf{X}_{\text{ph}}$ lying on the US plane. The point has US image pixel coordinates $\left[u_{\text{us}}, v_{\text{us}}\right]^\top$ and US probe coordinates $\mathbf{X}_{\text{us}} = \left[s_L u_{\text{us}}, 0, s_A v_{\text{us}}, 1\right]^\top$, where s_A and s_L are the pixel-to-meter scale factors. The US probe y-axis is chosen normal to the US plane. We denote a rigid transformation from coordinates a to coordinates b as T_b^a. We thus write $\mathbf{X}_{\text{ph}}$ as:

$$\mathbf{X}_{\text{ph}} = (\mathsf{T}_{\text{ca}}^{\text{ph}})^{-1}\mathsf{T}_{\text{ca}}^{\text{pr}}\mathsf{T}_{\text{pr}}^{\text{us}}\mathbf{X}_{\text{us}}, \tag{1}$$

where $\mathsf{T}_{\text{ca}}^{\text{ph}}$ and $\mathsf{T}_{\text{ca}}^{\text{pr}}$ are the poses of the phantom and the probe respectively in camera



LUS probe calibration, locates the LUS image pixels in LUS probe coordinates

Pose estimation, locates the LUS probe in the laparoscopic image

Registration of 3D preoperative liver model with the laparoscopic image

Evaluation of the registration predicted tumour (yellow) using the LUS profile (red)

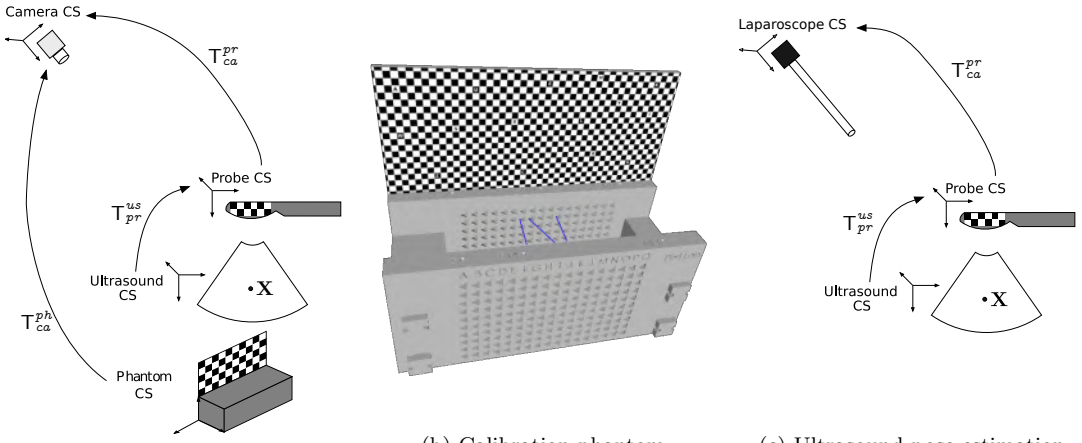
Figure 1. Overall methodology for groundtruth acquisition and registration evaluation.

coordinates, which are estimated from the imaged checkerboard corners. We use the phantom proposed in (Carbajal et al. 2013), shown in figure 2b, which includes N-wires structures. The wires intersect the US plane in points which appear as blobs in the US image. Equation (1) shows that we can calibrate the probe, namely compute s_A, s_L and $\mathsf{T}^{\text{us}}_{\text{pr}}$, from a set of point correspondences $\{(\mathbf{X}_{\text{ph}}, \mathbf{X}_{\text{us}})\}$. This forms a nonlinear optimisation problem over s_A, s_L and $\mathsf{T}^{\text{us}}_{\text{pr}}$. An initialisation is computed in closed-form from 3 point correspondences (Boctor et al. 2004). The initialisation is then refined, following (Carbajal et al. 2013), by minimising the nonlinear least-squares cost:

$$\text{IPE} = \sum_{i=1}^{N} \sum_{j \in \mathcal{M}_i} \|\mathbf{X}^{i,j}_{\text{ph}} - (\mathsf{T}^{\text{ph}}_{\text{ca}})^{-1} \mathsf{T}^{\text{pr}}_{\text{ca}} \mathsf{T}^{\text{us}}_{\text{pr}} \mathbf{X}^{i,j}_{\text{us}}\|^2, \tag{2}$$

over $\mathsf{T}^{\text{us}}_{\text{pr}}$,$s_A$ and s_L, where IPE stands for in-plane-error, $i \in [1, N]$ is the image index, $\mathcal{M}_i$ the wire indices whose intersection can be reliably extracted in the US image i and $\mathbf{X}^{i,j}_{\text{us}}$ the coordinates of the 3D point intersection of the US image i with the wire j. This method does not take the uncertainty of camera poses into account.

We propose an extension of the above optimisation which includes the refinement of the camera poses while introducing two additional cost terms formed by the repro-



(a) Ultrasound probe calibration (b) Calibration phantom (c) Ultrasound pose estimation

Figure 2. Ultrasound probe calibration and pose estimation; CS stands for 'coordinate system'.

jection errors associated to the checkerboard corners:

$$\alpha \sum_{i=1}^{N} \sum_{j \in \mathcal{L}_i} \| f(\mathsf{K}\mathsf{T}_{\text{ca}}^{\text{pr},i} \mathbf{O}_{\text{pr}}^{j}) - \mathbf{o}_{\text{ca}}^{i,j} \|^2 + \beta \sum_{i=1}^{N} \sum_{j \in \mathcal{P}_i} \| f(\mathsf{K}\mathsf{T}_{\text{ca}}^{\text{ph},i} \mathbf{Q}_{\text{ph}}^{j}) - \mathbf{q}_{\text{ca}}^{i,j} \|^2 + \text{IPE}, \quad (3)$$

where $\mathcal{L}_i$ and $\mathcal{P}_i$ are the indices of the probe and phantom imaged checkerboard corners respectively extracted in image i; $f([u, v, w]^\top) = [u/w, v/w]^\top$ is the canonical perspective projection modelling the camera; K is the camera calibration matrix; $\mathsf{T}_{\text{ca}}^{\text{pr},i}$ and $\mathsf{T}_{\text{ca}}^{\text{ph},i}$ are the probe and phantom poses respectively in camera coordinates, $\mathbf{O}_{\text{pr}}^{j}$ and $\mathbf{Q}_{\text{ph}}^{j}$ are the 3D coordinates of the corners for the probe and phantom checkerboards respectively; $\mathbf{o}_{\text{ca}}^{i,j}$ and $\mathbf{q}_{\text{ca}}^{i,j}$ are the imaged checkerboard corners for the probe and the phantom respectively, and α and β are hyperparameters fixed empirically to $\alpha = 2.73$ and $\beta = 2.92$. The initial estimates for $\mathsf{T}_{\text{ca}}^{\text{ph}}$ and $\mathsf{T}_{\text{ca}}^{\text{pr}}$ are obtained using EPnP (Lepetit et al. 2009). The optimisation is performed using Levenberg-Marquardt.

The proposed solution with camera pose refinement outperforms the original solution of minimising IPE on its own. Results are shown in table 3 and discussed in section 3.3.

3.2. Phantom Design Evolution with Improved N-wires Configurations

The geometry of the phantom proposed by Carbajal et al. (2013) allows one to easily customise the wire configuration. We use this property to design new wire configurations which better constrain the calibration solution. For that, we propose to propagate and minimise uncertainty. The values of the first two terms in equation (3) are independent of the wire configuration and we thus consider the IPE term (2) only. Concretely, we study uncertainty propagation for a single acquisition, referred to as *US plane pose estimation* in the sequel, which is the problem of finding the US plane pose which best models the location of the wire intersections in the US image. A wire configuration minimising uncertainty in the estimation of $\mathsf{T}_{\text{pr}}^{\text{us}}$ for a single image is expected to also reduce it for an image collection and therefore to produce improved US probe calibration.

Figure 3 shows how uncertainty propagates in US probe calibration. We denote by $\mathbf{B}_i$ the exact intersection location of wire i in the US image. We assume that the wire extraction uncertainty follows an additive *i.i.d.* bivariate normal distribution with zero-mean and covariance matrix Σ_B:

$$\mathbf{\Delta}_i^B \sim \mathcal{N}(0, \Sigma^B) \qquad \Sigma^B = \begin{bmatrix} \sigma_{B,A}^2 & 0 \\ 0 & \sigma_{B,L}^2 \end{bmatrix}, \tag{4}$$

where $\sigma_{B,A}^2$ and $\sigma_{B,L}^2$ are the axial and lateral variances. We also model the uncertainty on the wire endpoints, which is induced by the limited 3D printing resolution of the phantom and wire bending effects near the phantom holes. Let $\mathbf{W}_i$ be the vector of endpoint coordinates of the wire i:

$$\mathbf{W}_i = \left[X_{1,i}, Y_{1,i}, Z_{1,i}, X_{2,i}, Y_{2,i}, Z_{2,i}\right]^\top. \tag{5}$$

We assume that the endpoint uncertainty follows an additive *i.i.d.* multivariate normal distribution with zero-mean and isotropic covariance matrix Σ^W:

$$\mathbf{\Delta}_i^W \sim \mathcal{N}(0, \Sigma^W) \qquad \Sigma^W = \sigma_W^2 \mathrm{I}_{6\times 6}. \tag{6}$$

We propose to propagate uncertainty through the US plane pose estimation from the wire intersections extracted in the US image. Let $\mathbf{P}$ be the vector of parameters of the US plane pose:

$$\mathbf{P} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z]^\top, \tag{7}$$

where $[\theta_x, \theta_y, \theta_z]^\top$ are the Euler angles of the rotation and $[t_x, t_y, t_z]^\top$ the translation. The plane pose estimation is performed by minimising the IPE cost (2). We assume that the uncertainty of the output:

$$\mathbf{\Delta}^P = \mathrm{g}(\mathbf{B}_i + \mathbf{\Delta}_i^B, \mathbf{W}_i + \mathbf{\Delta}_i^W) - \mathrm{g}(\mathbf{B}_i, \mathbf{W}_i), \tag{8}$$

follows an additive multivariate normal distribution $\mathbf{\Delta}^P \sim \mathcal{N}(\mu, \Sigma^P)$, where g is the pose function, taking as inputs the wire intersections $\{\mathbf{B}_i\}_{i\in\mathcal{M}_i}$ in the US image and the wire endpoints $\{\mathbf{W}_i\}_{i\in\mathcal{M}_i}$ and providing as outputs the plane pose $\mathbf{P}$ minimising the IPE cost (2). Our objective is to find a wire configuration minimising the output uncertainty Σ^P.

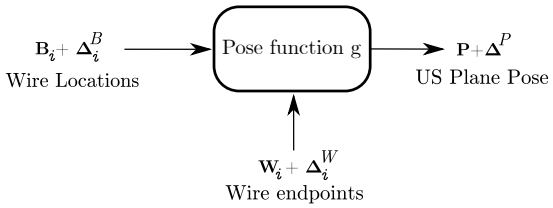
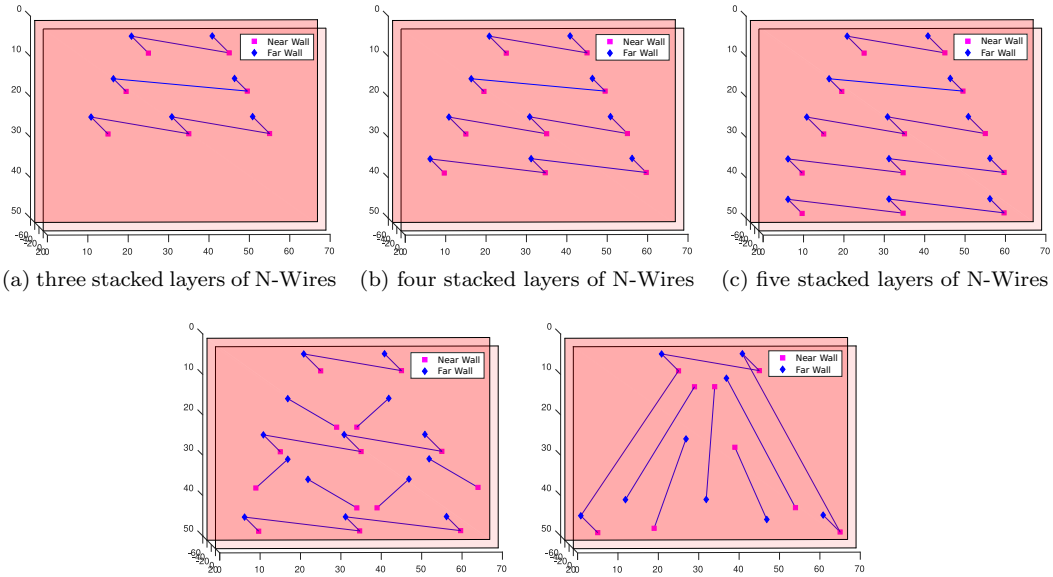


Figure 3. Propagation of uncertainty the US calibration phantom.

Concretely, we find wire configurations that minimise the generalised variance $|\Sigma^P|$, namely the determinant of the covariance matrix Σ^P, which is a common measure in uncertainty propagation (Wilks 1960). We propose to estimate Σ^P using a Monte



(a) three stacked layers of N-Wires (b) four stacked layers of N-Wires (c) five stacked layers of N-Wires

(d) proposed structure with three layers of N-Wires and six oblique wires in between

(e) proposed structure with flat and slanted N-Wires and five oblique wires in between

Figure 4. The five evaluated wire structures in the US calibration phantom design (the axes are in mm)

Carlo approach (Ogilvie 1984). The principle is to generate input samples randomly following their distributions, perform the deterministic plane pose computation and infer the parameters of the output distribution from the output samples. Concretely, we use 1000 input samples with $\sigma_{B,A} = \sigma_{B,L} = 0.45\,\text{mm}$ and $\sigma_W = 0.5\,\text{mm}$. These values were chosen empirically based on observations made during our experiments. The computed output samples allow us to find the mean, variance and covariance for each of the six outputs, giving Σ^P and finally, the sought $|\Sigma^P|$.

We tested five wire configurations, labelled (a) to (e), shown in figure 4. The first three ones (a-c) are typical stacked N-Wires with three, four and five layers respectively, inspired from (Carbajal et al. 2013). The fourth and fifth structures (d,e) are proposed by us: they differ from the first three mainly in the fact that they show slant over two dimensions of the 3D space, namely in the x and z directions rather than in the x direction only. An N-wire pattern is always present, to facilitate plane pose initialisation, as described in section 3.1.

The plane pose uncertainty computed for each of the five structures are given in table 2. For the first three structures (a-c), this shows that, as expected, adding layers to a typical stacked N-wire structure decreases pose uncertainty. The uncertainty strongly drops, from 9.7351 to 0.0032. The last two configurations (d,e) show the lowest output uncertainty, namely 0.0015 and 0.0003. The calibration evaluation of section 3.3 and the dataset acquisition of section 4 use configuration (d). This is because we discovered configuration (e) only after data acquisition was completed.

3.3. Validation of the Calibration and Data Acquisition Setup

We evaluated our data acquisition setup in realistic conditions, nearly identical to LLR in terms of hardware and laparoscope-to-probe distance varying within $[10, 15]\,\text{cm}$. A

Table 2. Pose uncertainty statistics (mm and deg) computed for the five structures of figure 4.

Id	Uncertainty in position						Uncertainty in orientation						$\lvert\Sigma^P\rvert$
	X		Y		Z		α		β		γ		
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	
a	-0.064	1.8	0.005	3.6	0.179	1.7	-0.051	3.4	0.023	0.6	-0.125	3.5	9.7351
b	0.032	1.1	-0.109	2.8	0.094	1.2	0.089	2.2	0.019	0.4	-0.113	2.4	0.0636
c	0.026	0.9	-0.084	2.2	0.074	0.8	0.001	1.8	-0.020	0.3	-0.077	1.7	0.0032
d	-0.017	0.8	-0.016	2.4	0.026	0.8	-0.025	1.5	0.002	0.3	-0.002	1.9	0.0015
e	0.006	0.6	-0.026	2.2	0.068	0.8	0.004	1.0	0.005	0.4	-0.031	1.9	0.0003

laparoscope was used to acquire optical images of the calibrated LUS probe. The probe acquires US images of a cross-wire phantom, similar to the one used for calibration but with wire-crossing points of known location to be used as landmarks. During the acquisition, the probe was positioned to make the wire-crossing points visible in the US image. The proposed evaluation measures the deviation between the groundtruth and the calibration-predicted landmarks in US pixel coordinates. The predictions are obtained by chaining the probe pose $\hat{\mathsf{T}}^{\text{pr}}_{\text{ca}}$ found from EPnP with the calibration results $\hat{\mathsf{T}}^{\text{us}}_{\text{pr}}$, $\hat{s}_A$ and $\hat{s}_L$.

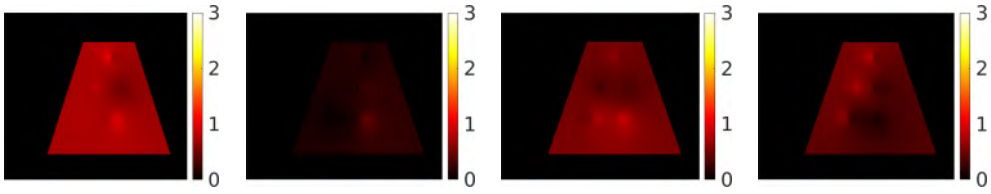
Concretely, the location of the wire-crossing points $\mathbf{X}^*_{\text{ph}}$ used as ground truth are transferred from phantom coordinates to camera coordinates as $\mathbf{X}^*_{\text{ca}} = \mathsf{T}^{\text{ph}}_{\text{ca}}\mathbf{X}^*_{\text{ph}}$, where $\mathsf{T}^{\text{ph}}_{\text{ca}}$ is obtained from a checkerboard attached to the phantom, as for prior probe calibration. The location of the wire-crossing points extracted in the US image are transferred to camera coordinates using $\mathbf{X}_{\text{ca}} = \mathsf{T}^{\text{pr}}_{\text{ca}}\hat{\mathsf{T}}^{\text{us}}_{\text{pr}}\hat{\mathbf{X}}_{\text{us}}$. The final error, which is the discrepancy between the measured and predicted points, is then $\epsilon = \|\mathbf{X}_{\text{ca}} - \mathbf{X}^*_{\text{ca}}\|_2$. This represents an upper bound on the real error, as the poses $\mathsf{T}^{\text{ph}}_{\text{ca}}$ are estimates and thus subject to noise. We performed and evaluated three independent probe calibrations. For each calibration, the proposed evaluation was performed. Each evaluation involved 20 images with different probe poses and various wire-crossing point depths in the US images. We report statistics on ϵ in table 3, along with the error components over the US axial, lateral and out-of-plane directions. For comparison purposes, we report the same statistics for calibrations obtained with the existing IPE-based optimisation (Carbajal et al. 2013). Figure 5 shows the error distribution for one calibration over the US image; the other calibration runs have a similar error distribution. This distribution was obtained by interpolating from the error measured at the wire-crossing points over the entire US image. We use the inverse distance weighting method for interpolation.

We can draw two important observations from these results. First, the calibration errors are in the order of 1 mm. Hence, the proposed setup is suited for the evaluation of state-of-the-art registration in LLR, as 1 mm represents between 3% to 12% of the registration errors, as given in table 4. Second, they show that our calibration method outperforms (Carbajal et al. 2013). Hence, there is a benefit in fine-tuning the probe and phantom poses.

The measured error depends on both LUS calibration and pose estimation. Pose estimation is known to be more susceptible to error along the camera axis. In our tests, in order to have a better view of the LUS probe checkerboard, the laparoscope was pointed towards the side and top of the LUS probe and hence the EPnP-induced error is projected dominantly in the axial and out-of-plane directions. It is in agreement with the results in Table 3.

Table 3. Error statistics (mm) for the proposed calibration method and the method in (Carbajal et al. 2013).

	Method	Total error				Avg. directional error		
		Avg.	STD	Min	Max	Lateral	Axial	Out-of-plane
1st calibration	Proposed	1.26	0.29	0.73	1.78	0.42	0.71	0.83
	Carbajal et al. (2013)	1.37	0.41	0.50	2.30	0.52	0.80	0.81
2nd calibration	Proposed	0.71	0.20	0.25	1.03	0.15	0.40	0.49
	Carbajal et al. (2013)	1.30	0.25	1.01	1.86	0.63	0.82	0.76
3rd calibration	Proposed	1.44	0.47	0.95	2.27	0.58	1.27	0.24
	Carbajal et al. (2013)	1.55	0.46	0.66	2.34	0.54	1.30	0.60



(a) Total error (b) Axial error (c) Lateral error (d) Out-of-plane error

Figure 5. Interpolated error distribution (mm) over the US image for the 2nd calibration run.

4. Evaluation Criteria, Dataset and Registration Evaluation Results

In AR assisted LLR, the registration method predicts the tumours in laparoscope 3D coordinates, from the preoperative 3D model and intraoperative laparoscopic images. We discuss two evaluation criteria for this prediction, the proposed dataset and an evaluation of existing registration methods.

4.1. The Inclusion Criterion

We propose a binary inclusion criterion which is positive (hence, passed) if and only if registration-predicted tumour augmented by the typical oncologic margin of 1 cm contains all the LUS tumour profiles. Specifically, for image i, $i \in [1, N]$, over a total of N images, we define the predicted tumour volume T_i, the augmented volume T_i^a and the LUS tumour profile transferred to laparoscope coordinates P_i. The inclusion criterion is then:

$$\mathrm{IC} = (P_1 \subset T_1^a) \wedge \cdots \wedge (P_N \subset T_N^a)\,. \tag{9}$$

The inclusion criterion must be met for precision surgical guidance. It is however difficult to pass and does not bring much quantitative information on the actual precision, especially if failed, as opposed to TRE. The inclusion criterion must be met for precision surgical guidance. The rationale for it is as follows. Using augmented reality as the primary method of resection guidance means that the surgeon resects a volume containing T_i^a. For a successful R0 resection comprising all malignant tissues, the part P_i detected as malignant by LUS must thus be included in the resected volume, hence in T_i^a. This directly leads to the above definition of the inclusion criterion, as the conjunction of the inclusion of each LUS profile in the corresponding registration-predicted volume. The inclusion criterion is difficult to pass. Failure is to be interpreted as the unreliability of the registration-predicted tumour volume, in the sense that trusting it for resection would involve that malignant tissue could be left behind. The inclusion

criterion is thus a fundamental safety criterion, which must be passed for a method to be used in clinical practice. It does not bring much quantitative information on the actual precision, especially if failed, as opposed to TRE. The latter however does not allow a registration method to qualify for safe clinical practice. The inclusion criterion and TRE are thus complementary.

4.2. The Target Registration Error

We propose to estimate TRE from the groundtruth obtained from LUS. We define TRE as the average over the tumour volume of the deformation field $\phi_i : \mathbb{R}^3 \to \mathbb{R}^3$ that exists between the LUS and registration-predicted tumours:

$$\text{TRE} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\int_{\partial P_i} dX} \int_{\partial P_i} \|\phi_i(X)\|_2^2 \, dX, \tag{10}$$

where ∂P_i is the boundary curve of P_i. This integral cannot be solved analytically and we thus discretise it by sampling ∂P_i with M_i points $\{P_{ij}\}$:

$$\text{TRE} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sum_{j=1}^{M_i} \Delta P_{ij}} \sum_{j=1}^{M_i} \|\phi_i(P_{ij})\|_2^2 \, \Delta P_{ij}, \tag{11}$$

where $\Delta P_{ij} = \|P_{i(j+1)} - P_{ij}\|_2$. Without loss of generality, we assume evenly spaced discretisations with $M_1 = \cdots = M_N = M$ and $\Delta P_{i1} = \Delta P_{i2} = ... = \Delta P_{iM} = \Delta P_i$:

$$\text{TRE} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{1}{M\Delta P_i} \sum_{j=1}^{M} \|\phi_i(P_{ij})\|_2^2 \, \Delta P_i = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|\phi_i(P_{ij})\|_2^2. \tag{12}$$

We assume that the deformation field is kept constant during data acquisition, hence $\phi_1(X) = \phi_2(X) = ...\phi_N(X) = \phi(X)$. We break it down into two additive fields as $\phi(X) = \rho(X; R, t) + \epsilon(X)$, with a rigid field $\rho(X; R, t)$ which explains the overall discrepancy by a rotation R and translation t in the vicinity of the tumour, and a residual field $\epsilon(X)$. With this decomposition, we can obtain the deformation field using an Iterative Closest Point (ICP) method following (Beasley et al. 1999) to estimate R, t. Introducing the closest point T_{ij} to $RP_{ij} + t$, we minimise the energy:

$$\begin{aligned} E_\epsilon &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|\epsilon(P_{ij})\|_2^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|\phi(P_{ij}) - \rho(P_{ij})\|_2^2 \\ &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|(T_{ij} - P_{ij}) - (RP_{ij} + t - P_{ij})\|_2^2 \\ &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \|T_{ij} - (RP_{ij} + t)\|_2^2. \end{aligned} \tag{13}$$

We give pseudo-code for our method in algorithm 1. The rigid field is initialised to an identity rotation and the average displacement of the centre of gravity, represented by operator cog, for the translation. We then alternate the computation of closest points

given the rigid field and update of the rigid field given the correspondences as the closest points. The update is solved by Horn's closed form solution (Horn et al. 1988) to minimise E_ϵ. Convergence is determined by thresholding the incremental average closest points displacement by a small constant taken as 1×10^{-3} mm.

The metric computed here is a specific implementation of TRE. We develop this implementation from the most general definition of TRE, with the help of assumptions based on the nature of our ground-truth data. As the metric is computed between volumes and profiles by finding closest-point correspondences, this implementation could be named Closest-Point-Volume-Profile TRE. To keep it short, we however simply refer to it as TRE.

Algorithm 1: TRE estimation

Set $R_0 \leftarrow \mathrm{I}_{3\times3}$, $t_0 \leftarrow \frac{1}{N}\sum_{i=1}^{N}(\mathrm{cog}(T_i) - \mathrm{cog}(P_i))$ and $k \leftarrow 0$
repeat
  Set $T_{ij}^k \leftarrow$ closest point on T_i to $R_k P_{ij} + t_k$, $i \in [1, N]$, $j \in [1, M]$
  Set $k \leftarrow k + 1$
  Set $(R_k, t_k) \leftarrow \arg\min_{R,t} E_\epsilon$
until convergence
Set TRE $\leftarrow \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\|T_{ij}^{k-1} - P_{ij}\|_2$

4.3. Clinical Dataset

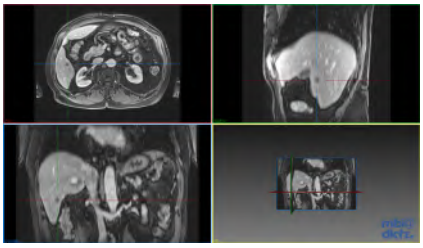
Our dataset currently includes LLR procedures in four patients with one tumour each. The data collection is supported by an ethical approval with ID IRB00008526-2019-CE58 issued by CPP Sud-Est VI in Clermont-Ferrand, France.

For each patient, it contains four key elements, illustrated in figure 6: (a) a preoperative CT, (b) a preoperative 3D model of the liver shape and its inner tumours obtained by manual segmentation of the CT made by an experienced surgeon, (c) a set of intraoperative laparoscopic images and (d) corresponding LUS images acquired synchronously. The laparoscopic images show the anterior liver, completely or partially. The tumours were segmented by an experienced surgeon in each LUS image and are provided as binary masks. The dataset also includes all the data related to probe calibration and pose estimation. Our Matlab implementation is provided, which produces the groundtruth and computes the two evaluation criteria. It thus facilitates the evaluation of any custom registration method.

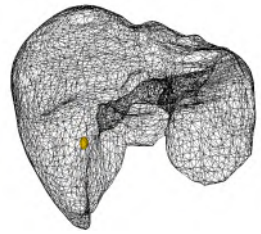
4.4. Evaluation Results and Discussion

We evaluated two existing registration methods (Adagolodjo et al. 2017; Koo et al. 2017) over the complete proposed dataset with the two proposed criteria. Figure 7a shows an example registration-predicted tumour from method (Adagolodjo et al. 2017) coregistered with the LUS groundtruth profile. Figure 7b shows an example AR result for visualisation purposes only, showing the registration-predicted tumour and LUS plane augmented in the laparoscopic image.

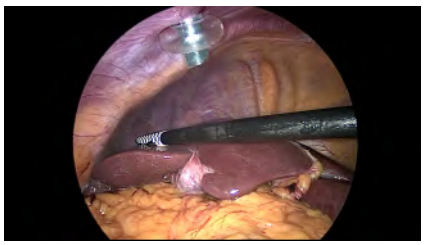
Table 4 summarises the evaluation results. Both methods failed the inclusion criterion for all four patients. This is not surprising since the criterion is difficult to meet and these methods only use one image to solve registration, hence are prone to high registration uncertainty. TRE shows that both methods are more accurate for patients



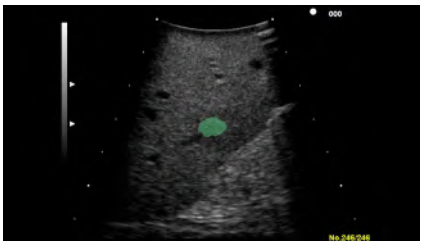
(a) Preoperative CT



(b) Preoperative 3D model



(c) Laparoscopic images



(d) Segmented LUS images

Figure 6. Dataset composition for each patient.

Table 4. Evaluation results for two existing registration methods and tumour depth (the unit is mm).

	Patient 1		Patient 2		Patient 3		Patient 4		Overall	
	IC	TRE	IC	TRE	IC	TRE	IC	TRE	IC	TRE
Adagolodjo et al. (2017)	Fail	8.25	Fail	37.25	Fail	28.40	Fail	15.83	0 of 4	22.43
Koo et al. (2017)	Fail	9.49	Fail	38.95	Fail	25.04	Fail	18.35	0 of 4	22.95
Tumour depth	6.61		11.53		18.62		1.31		Avg: 9.52	

1 and 4. This is explained by the tumour being more superficial than in patients 2 and 3. Overall, method (Adagolodjo et al. 2017) outperforms method (Koo et al. 2017) by a slight margin.

In both methods, the registration is constrained using anatomical landmarks. It can thus be expected that the regions near these landmarks are less prone to errors than away from them. The deep tumours are located in the posterior part of the liver, which is located far from the observable landmarks, and can thus undergo a higher TRE, as consistently reflected in the experimental results.

5. Conclusions

We have proposed the first methodology to evaluate preoperative data registration in LLR quantitatively, with the key idea of collecting groundtruth tumour positions using LUS. For that, we have proposed a complete methodology, involving LUS probe calibration and colocalisation with the laparoscope. We have used a calibration phantom, for which we have proposed new designs, following a principled statistical uncertainty modelling. We have validated groundtruth acquisition in realistic lab experiments, showing an accuracy upper-bounded by 1 mm, hence perfectly compatible with the evaluation task at hand. Equipped with this methodology, we have collected a dataset of four LLR procedures and proposed two evaluation criteria, which we have used to perform an evaluation of two existing AR methods. The proposed methodology and

dataset form a step toward standardising the quantitative evaluation of computer-assisted laparoscopy. Future work includes additional data collection and evaluation of additional existing registration methods such as Modrzejewski et al. (2019).

Acknowledgments. This work was partly funded by CNRS under the Hepataug pre-industrial project and by Cancéropôle CLARA under the AIALO project.

References

Adagolodjo Y, Trivisonne R, Haouchine N, Cotin S, Courtecuisse H. 2017. Silhouette-based pose estimation for deformable organs application to surgical augmented reality. In: IROS. IEEE. p. 539–544.

AliceVision-Meshroom. 2021. Available from: `alicevision.org`.

Beasley RA, Stefansic JD, Herline AJ, Guttierez L, Galloway Jr RL. 1999. Registration of ultrasound images. In: SPIE Medical Imaging 1999: Image Display; vol. 3658. p. 125–132.

Bernhardt S, Nicolau SA, Bartoli A, Agnus V, Soler L, Doignon C. 2015. Using shading to register an intraoperative ct scan to a laparoscopic image. In: CARE. Springer. p. 59–68.

Boctor E, Viswanathan A, Choti M, Taylor RH, Fichtinger G, Hager G. 2004. A novel closed form solution for ultrasound calibration. In: IEEE ISBI. p. 527–530.

Bouguet JY. 2000. Matlab camera calibration toolbox. Caltech Technical Report.

Carbajal G, Lasso A, Gómez Á, Fichtinger G. 2013. Improving n-wire phantom-based freehand ultrasound calibration. IJCARS. 8(6):1063–1072.

Cash DM, Miga MI, Glasgow SC, Dawant BM, Clements LW, Cao Z, Galloway RL, Chapman WC. 2007. Concepts and preliminary data toward the realization of image-guided liver surgery. Journal of Gastrointestinal Surgery. 11(7):844–859.

Cash DM, Sinha TK, Chapman WC, Terawaki H, Dawant BM, Galloway RL, Miga MI. 2003. Incorporation of a laser range scanner into image-guided liver surgery: surface acquisition, registration, and tracking. Medical Physics. 30(7):1671–1682.

Cenni F, Monari D, Desloovere K, Aertbeliën E, Schless SH, Bruyninckx H. 2016. The reliability and validity of a clinical 3d freehand ultrasound system. Computer methods and programs in biomedicine. 136:179–187.

Cheung CL, Wedlake C, Moore J, Pautler SE, Peters TM. 2010. Fused video and ultrasound images for minimally invasive partial nephrectomy: a phantom study. In: MICCAI 2010. p. 408–415.

Clements LW, Collins JA, Weis JA, Simpson AL, Adams LB, Jarnagin WR, Miga MI. 2016. Evaluation of model-based deformation correction in image-guided liver surgery via tracked intraoperative ultrasound. SPIE Medical Imaging. 3(1):015003.

Dumpuri P, Clements LW, Dawant BM, Miga MI. 2010. Model-updated image-guided liver surgery: preliminary results using surface characterization. Progress in biophysics and molecular biology. 103(2-3):197–207.

Espinel Y, Özgür E, Calvet L, Le Roy B, Buc E, Bartoli A. 2020. Combining visual cues with interactions for 3d–2d registration in liver laparoscopy. Annals of Biomedical Eng:1–16.

Haouchine N, Roy F, Untereiner L, Cotin S. 2016. Using contours as boundary conditions for elastic registration during minimally invasive hepatic surgery. In: IROS. p. 495–500.

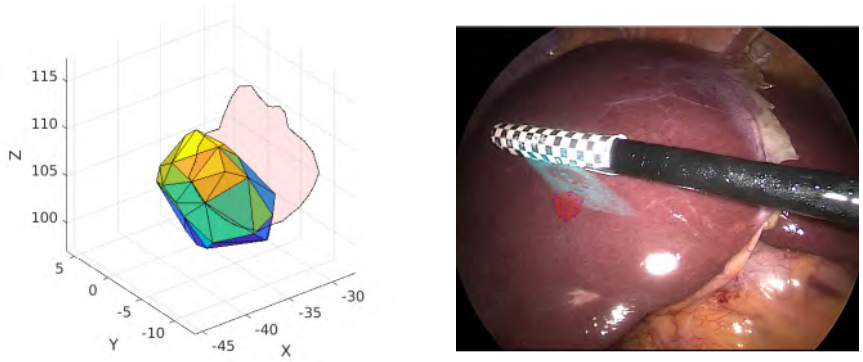
Hartley R, Zisserman A. 2003. Multiple view geometry in computer vision. Cambridge university press.

Hayashi Y, Igami T, Hirose T, Nagino M, Mori K. 2015. Development and clinical application of surgical navigation system for laparoscopic hepatectomy. In: SPIE Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling; vol. 9415. p. 94151X.

Horn BK, Hilden HM, Negahdaripour S. 1988. Closed-form solution of absolute orientation using orthonormal matrices. JOSA A. 5(7):1127–1135.

Hsu PW, Prager RW, Gee AH, Treece GM. 2009. Freehand 3d ultrasound calibration: a review.

In: Advanced imaging in biology and medicine. Springer; p. 47–84.
Koo B, Özgür E, Le Roy B, Buc E, Bartoli A. 2017. Deformable registration of a preoperative 3d liver volume to a laparoscopy image using contour and shading cues. In: MICCAI. p. 326–334.
Lasso A, Heffter T, Rankin A, Pinter C, Ungi T, Fichtinger G. 2014. Plus: Open-source toolkit for ultrasound-guided intervention systems. IEEE Trans Biomed Eng. 61(10):2527–2537.
Lepetit V, Moreno-Noguer F, Fua P. 2009. Epnp: An accurate o (n) solution to the pnp problem. International journal of computer vision. 81(2):155.
Liu X, Kang S, Wilson E, Peters CA, Kane TD, Shekhar R. 2014. Evaluation of electromagnetic tracking for stereoscopic augmented reality laparoscopic visualization. In: Workshop on Clinical Image-Based Procedures. Springer. p. 84–91.
Maier-Hein L, Groch A, Bartoli A, Bodenstedt S, Boissonnat G, Chang PL, Clancy N, Elson DS, Haase S, Heim E, et al. 2014. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. IEEE Trans on medical imaging. 33(10):1913–1930.
Mercier L, Langø T, Lindseth F, Collins DL. 2005. A review of calibration techniques for freehand 3-d ultrasound systems. Ultrasound in medicine & biology. 31(4):449–471.
Modrzejewski R, Collins T, Seeliger B, Bartoli A, Hostettler A, Marescaux J. 2019. An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions. IJCARS. 14(7):1237–1245.
Ogilvie JF. 1984. A monte-carlo approach to error propagation. Computers & chemistry. 8(3):205–207.
Plantefève R, Peterlik I, Haouchine N, Cotin S. 2016. Patient-specific biomechanical modeling for guidance during minimally-invasive hepatic surgery. Annals of biomedical engineering. 44(1):139–153.
Poon TC, Rohling RN. 2005. Comparison of calibration methods for spatial tracking of a 3-d ultrasound probe. Ultrasound in medicine & biology. 31(8):1095–1108.
Poulin F, Amiot LP. 2002. Interference during the use of an electromagnetic tracking system under or conditions. Journal of biomechanics. 35(6):733–737.
Robu MR, Ramalhinho J, Thompson S, Gurusamy K, Davidson B, Hawkes D, Stoyanov D, Clarkson MJ. 2018. Global rigid registration of ct to video in laparoscopic liver surgery. IJCARS. 13(6):947–956.
Suwelack S, Röhl S, Bodenstedt S, Reichard D, Dillmann R, dos Santos T, Maier-Hein L, Wagner M, Wünscher J, Kenngott H, et al. 2014. Physics-based shape matching for intraoperative image guidance. Medical physics. 41(11):111901.
Thompson S, Schneider C, Bosi M, Gurusamy K, Ourselin S, Davidson B, Hawkes D, Clarkson MJ. 2018. In vivo estimation of target registration errors during augmented reality laparoscopic surgery. IJCARS. 13(6):865–874.
Wang L, Wang T, Liu H, Hu L, Han Z, Liu W, Guo N, Qi Y, Xu Y. 2018. An automated calibration method of ultrasonic probe based on coherent point drift algorithm. IEEE Access. 6:8657–8665.
Wilks SS. 1960. Multidimensional statistical scatter. Contributions to probability and statistics:486–503.



(a) Registration-predicted tumour (mesh) and LUS tumour profile (planar surface), units are mm

(b) Laparoscopic image with registration-predicted tumour (red) and LUS plane (cyan)

Figure 7. Sample registration result with coregistered LUS groundtruth.