

Endoscopic scene labelling and augmentation using intraoperative pulsatile motion and colour appearance cues with preoperative anatomical priors

Masoud S. Nosrati¹ · Alborz Amir-Khalili² · Jean-Marc Peyrat³ · Julien Abinahed³ · Osama Al-Alao⁴ · Abdulla Al-Ansari⁴ · Rafeef Abugharbieh² · Ghassan Hamarneh¹

Received: 7 July 2015 / Accepted: 13 November 2015 / Published online: 12 February 2016
© CARS 2016

Abstract

Purpose Despite great advances in medical image segmentation, the accurate and automatic segmentation of endoscopic scenes remains a challenging problem. Two important aspects have to be considered in segmenting an endoscopic scene: (1) noise and clutter due to light reflection and smoke from cutting tissue, and (2) structure occlusion (e.g. vessels occluded by fat, or endophytic tumours occluded by healthy kidney tissue).

Methods In this paper, we propose a variational technique to augment a surgeon's endoscopic view by segmenting visible as well as occluded structures in the intraoperative endoscopic view. Our method estimates the 3D pose and deformation of anatomical structures segmented from 3D preoperative data in order to align to and segment corresponding structures in 2D intraoperative endoscopic views. Our preoperative to intraoperative alignment is driven by, first, spatio-temporal, signal processing based vessel pulsation cues and, second, machine learning based analysis of colour and textural visual cues. To our knowledge, this is the first work that utilizes vascular pulsation cues for guiding preoperative to intraoperative registration. In addition, we incorporate a tissue-specific (i.e. heterogeneous) physically based deformation model into our framework to cope with

the non-rigid deformation of structures that occurs during the intervention.

Results We validated the utility of our technique on fifteen challenging clinical cases with 45 % improvements in accuracy compared to the state-of-the-art method.

Conclusions A new technique for localizing both visible and occluded structures in an endoscopic view was proposed and tested. This method leverages both preoperative data, as a source of patient-specific prior knowledge, as well as vasculature pulsation and endoscopic visual cues in order to accurately segment the highly noisy and cluttered environment of an endoscopic video. Our results on in vivo clinical cases of partial nephrectomy illustrate the potential of the proposed framework for augmented reality applications in minimally invasive surgeries.

Keywords Robotic surgery · Partial nephrectomy · Image-guided surgery · Segmentation · 3D pose estimation · Endoscopy · Patient-specific model · Occluded vessels · Kidney

Introduction

Minimally invasive surgeries (MIS) are appreciated for their many advantages over traditional open surgery including decreased risk of infection due to minimal incisions, and faster recovery times for patients [15]. The shortcomings of MIS are mainly associated with the loss of direct 3D view of the surgical scene as well as cumbersome and non-intuitive tool manipulation. With the advent of robotic MIS systems, many of these shortcomings have been alleviated. Greater precision, improved dexterity and enhanced 3D immersive visualization for surgeons are three important advantages of robotic over traditional MIS.

✉ Masoud S. Nosrati
smasoudn@gmail.com; smn6@sfu.ca

¹ Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

² BiSICL, University of British Columbia, Vancouver, BC, Canada

³ Qatar Robotic Surgery Centre, Qatar Science and Technology Park, Doha, Qatar

⁴ Urology Department, Hamad General Hospital, Hamad Medical Corporation, Doha, Qatar

One of the application areas in which robotic MIS is being expanded is partial nephrectomy. Approximately 208,000 new cases of kidney cancer are diagnosed in the world each year [10]. The goal in partial nephrectomy, an effective treatment for localized renal cancers, is to excise only the cancerous regions and spare as much healthy tissue as possible to preserve kidney function. According to Gill et al. [14], a partial nephrectomy procedure is organized into five main stages: (1) bowel mobilization; (2) Hilar dissection and control; (3) identification and demarcation of tumour margins; (4) resection of tumour; and (5) reconstruction of the kidney (renorrhaphy). Among these, hilar dissection and tumour demarcation stand out as two critical and daunting stages requiring significant expertise.

Hilar dissection allows the surgeon access to the renal hilum, where the flow of blood into and out of the kidney is controlled by clamping the vessels. Improper vessel control due to overlooked accessory renal vessels can cause significant bleeding during resection [28]. The identification and demarcation stage is also critical as, during this stage, the surgeon localizes the tumour mass and identifies the resection margins. Proper demarcation speeds up the succeeding stage of tumour resection, improves the preservation of kidney function by sparing as much healthy tissue as possible, and avoids tumour recurrence by correctly delineating all of the cancerous tissue from the healthy ones.

Both of these two stages are complex tasks. In the hilar dissection and vessel clamping stage, some vessels might be accidentally missed due to substantial variability in patients vasculature and the amount of fat which surrounds the kidney and hides the vessels [31]. Tumour identification and demarcation is difficult due to noise associated with clutter such as bleeding or smoke from cutting, poor endoscopic image colour/texture contrast between different structures, occluding surgical tools, and limited 3D visibility of the structures of interest where only surfaces are observable from the camera feed. If the surgeon's level of experience is limited, the incidence of missed vessels or incorrect demarcation may significantly jeopardize the surgical outcome.

To minimize mistakes during the operation, surgeons typically rely on previously viewed preoperative 3D scans, e.g. computed tomography (CT), and mentally reconstruct locations of various structures during surgery. Transferring this mental abstraction from 3D to 2D data is an error-prone procedure especially if the surgeon's level of experience is limited. To facilitate this, many efforts have been made towards augmenting the endoscopic views. These methods vary from directly segmenting the endoscopic scene (e.g. using level sets) to registering the preoperative data onto the intraoperative endoscopic scene.

Related works

Reducing the difficulty of intraoperative navigation has been attempted by various approaches that rely on multi-modal registration to align the preoperative data onto the intraoperative view by, for example, fusing a 3D preoperative image volume with 2D intraoperative ultrasound [11], intraoperative MR [16], or 2D X-ray [25,33]. In this work, we focus on augmenting the endoscopic video as it remains the staple modality in MIS. Some recent works ignored preoperative data and segmented the endoscopic view using active contour-based methods [12,13] or parameter-sensitive morphological operations and thresholding techniques [21]. Since these approaches rely only on colour/intensity information, they often fail due to noise and clutter from bleeding and smoke. In addition, these methods focused on segmenting only one object in an endoscopic scene.

Other techniques proposed to register 3D preoperative data on 2D intraoperative view(s) manually, e.g. by using invasive fiducials and performing tracking [17,18,26,27,29]. While the registration in these methods is performed manually, the methods proposed in Yim et al. [32] and Merritt et al. [20] are able to automatically find the 3D pose. Yet, none of the aforementioned methods can handle the free-form deformation of tissues that usually happens due to respiratory motion and/or surgical intervention. Furthermore, these techniques are unable to segment critical occluded structures, e.g. vessels covered by fat.

To account for non-rigid tissue deformation, Agudo et al. [1] exploited the well-known FEM model analysis and proposed an online framework to model the non-rigid deformations of shapes in a video sequence. However, their method has been designed for a single camera setting. In our previous work [23], we incorporated generative statistical shape models of tissues in their 3D pose tracking and endoscopic video segmentation framework. Although the statistical model used is capable of encoding non-rigid deformations, these deformations are based on a heterogeneous training population and are not patient-specific. In addition, our previous method required manual 3D–2D registration for the first frame of the video and is incapable of localizing hidden vessels.

To localize vessels in the endoscopic view, different methods have been proposed. These methods include both hardware solutions, such as near infrared fluorescence imaging [30], and algorithmic methods based on colour/intensity information from the endoscope that highlight vasculature based on perfusion models [8]. However, due to sensitivity to the dose administered and the cost involved, hardware solutions are not widely used. Moreover, none of these methods [8,30] are able to identify occluded vessels that are hidden under thick layers of fat. A detailed survey of such vessel localization techniques was presented in our recent

paper [2]. In our recent works [2,3] we proposed a purely image-based method for labelling occluded vasculature by detecting minute pulsatile vessel motion. Feedback from our clinical collaborators revealed that the raw visualization of our fuzzy labels alone are not ideal for guidance as (1) they are susceptible to noise, (2) they cannot identify parts of the vasculature that are deeply hidden under other tissue, and (3) the visualizations are not easy to interpret compared to methods based on pre- to intra-operative alignment. This provided the motivation for us to integrate the complimentary advantages of our previously proposed techniques in [2] and [23] into a single framework.

Two important aspects have to be considered in segmenting an endoscopic video: noise/clutter due to light reflection from uneven tissue surfaces and smoke from cutting tissue, and structure occlusion either by tools or other tissues (e.g. vessels occluded by fat, or endophytic tumours occluded by healthy kidney tissue). In this paper, we address these two problems by proposing a variational technique to augment

the endoscopic surgical view. Our framework estimates the 3D pose and deformations of multiple structures in the preoperative data and uses that to aid the segmentation of *multiple visible* and *occluded structures* in an endoscopic scene. Our key contribution in this work is how we leverage vessel pulsation and endoscopic visual cues to guide 3D preoperative to endoscopic intraoperative alignment. Our surgical image guidance system combines signal processing and machine learning cues. The first performs spatio-temporal colour waveform analysis to detect *vessel pulsation*, and the second uses random decision forests (RF) to learn colour and textural visual patterns of tissue types in endoscopy. The above contributions are used to guide a multi-structure, preoperative to intraoperative alignment process that allow for both rigid and heterogeneous physically based, patient-specific non-rigid deformations. Table 1 presents a detailed summary of the capabilities of state-of-the-art methods compared to our proposed technique. To demonstrate the potential utility of our method to surgeons, we evaluated our method with a ret-

Table 1 Categorization and comparison between state-of-the-art methods for intraoperative guidance and our proposed method

	Organ	Data		Method	Cues	Multiple obj.	Biomechanical deformation	Occluded obj.	Auto. pose estimation
		Pre-op.	Intra-op.						
Figueiredo et al. [12]	C	×	EN	Active contour	CO	✓	×	×	–
Figueiredo et al. [13]	C	×	EN	Active contour	CO	×	×	×	–
Mewes et al. [21]	S	×	EN	Thresholding	GO, TX, CH	×	×	×	–
Estépar et al. [11]	K/CA	CT	US	LEPART	ED	✓	×	×	×
Gill et al. [16]	P	MR	MR	2D–3D reg.	IN	×	×	×	✓
Pickering et al. [25]	T/F	CT	FL	2D–3D reg.	IN	×	×	×	✓
Zikic et al. [33]	H	CBR	FL	2D–3D reg.	IN	×	×	×	✓
Pratt et al. [26]	K	CT	EN	3D–3D (stereo) reg.	ML	✓	×	×	×
Yim et al. [32]	VF	CT	EN	3D–2D reg.	IN	×	×	×	✓
Merritt et al. [20]	L	CT	EN	3D–2D reg.	IN	×	×	×	✓
Puerto et al. [27]	K	CT	EN	Feature tracking	AL	✓	×	×	×
Tobis et al. [30]	K	×	EN	Fluorescence imaging	C	×	×	×	–
Teber et al. [29]	K	CT	EN	3D–2D reg.	ML	✓	×	×	×
Amir-Khalili et al. [2]	V	×	EN	Motion segmentation	PL	×	×	✓	–
Nosrati et al. [23]	K	CT	EN	3D–2D reg.	CH, TX	✓	×	×	×
Our method	K	CT	EN	3D–2D reg.	CH, TX, PL	✓	✓	✓	✓

LEPART low-pass spectral phase correlation with harmonic selection. *op* operative. *obj* object. *reg* registration. *Auto* automatic. Data: *CT* computed tomography, *US* ultrasound, *MR* magnetic resonance, *EN* endoscopy, *FL* fluoroscopy, *CBR* cone beam reconstruction. Organs: *K* kidney, *CA* coeliac-aorta, *P* prostate, *T/F* tibia/femur, *H* head, *C* colon, *S* stomach, *V* vasculature, *VF* vocal folds, *L* Lung. Cues: *ED* edge *IN* intensity, *CO* colour, *CH* colour histogram, *TX* texture, *ML* manual landmark, *AL* automatic landmark, *PL* pulsation

respective study of fifteen challenging robot-assisted partial nephrectomy (RAPN) clinical cases.

Method

Problem statement and notation

Given 3D preoperative data, let $\mathcal{S}_{\text{pre}} = \{S_1, \dots, S_N\}$ be the set of N segmented structures of interest in the preoperative spatial domain $\Omega_{\text{pre}} \subset \mathbb{R}^3$, where S_i represents the surface of the i th structure. Having M camera views of the surgical scene, let $I_m : \Omega_{2D}^m \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ represent the 3-channel RGB image of m th camera view. We adopt the 3D geometry defined by first camera ($m = 1$) as the reference domain of the surgical scene, $\Omega_{\text{srg}} \subset \mathbb{R}^3$. Our objective is to augment the surgical intraoperative view by localizing and visualizing the N structures of interest in the 2D endoscopic images (I_1, \dots, I_M). To do so, we transform and deform the objects of interest in 3D such that their projections onto the 2D camera views (e.g. their silhouettes) align with the corresponding structures in I_1, \dots, I_M . Figure 1 illustrates the overview of our pipeline.

We implicitly represent the boundary of each structure in I_m by a level set function $\phi_m^i : \Omega_{2D}^m \rightarrow \mathbb{R}$ such that

$$\begin{cases} \phi_m^i(\mathbf{x}) > 0, & \mathbf{x} \text{ is inside the } i\text{th structure in } I_m \\ \phi_m^i(\mathbf{x}) = 0, & \mathbf{x} \text{ is on the boundary of the } i\text{th structure} \\ \phi_m^i(\mathbf{x}) < 0, & \mathbf{x} \text{ is outside the } i\text{th structure in } I_m. \end{cases} \quad (1)$$

Let $\mathbf{P}_\ell^{\text{pre},i} = (X_\ell^{\text{pre},i}, Y_\ell^{\text{pre},i}, Z_\ell^{\text{pre},i}) \in S_i$ be the coordinates of the ℓ th point on S_i . The level set ϕ_m^i is calculated as:

$$\phi_m^i(\mathbf{x}) = \text{SDM}\left(\partial\left(\mathcal{P}_m\left(T\left(\mathbf{P}^{\text{pre},i}\right); \pi_m\right)\right)\right), \quad (2)$$

where T is a spatial transformation consisting of two parts: a non-rigid component and a rigid component that maps the 3D models (\mathbf{P}^{pre}) from Ω_{pre} to the reference surgical domain Ω_{srg} . $\mathcal{P}_m : \Omega_{\text{srg}} \rightarrow \Omega_{2D}^m$ is the projection from the surgical scene frame of reference to Ω_{2D}^m given the corresponding camera parameters $\pi = \{\pi_1, \dots, \pi_M\}$, ∂ is the spatial 2D derivative of the projected model (i.e. resulting in the boundary of the structures), and $\text{SDM}(\cdot)$ is the signed distance map. In less technical terms, we position and deform the segmented 3D model (using T), virtually image (or project) the 3D model using the camera parameters (π), detect the boundary of the projected image (using ∂), and finally represent the boundary using level sets. We define T as:

$$T(\mathbf{P}^{\text{pre},i}) = \mathbf{R}(\mathbf{P}^{\text{pre},i} + \Psi^i \mathbf{u}^i)_{[3 \times n_i]} + \mathbf{t}, \quad (3)$$

where the columns of Ψ^i are the i th tissue-specific modes of vibration and \mathbf{u}^i are the corresponding weights that are used to non-rigidly deform the segmented preoperative models $\mathbf{P}^{\text{pre},i}$ (see “Patient-specific heterogeneous deformation model” section for more details). \mathbf{R} and \mathbf{t} are the rotation matrix and translation vector in 3D space, respectively, giving the rigid component of transformation T . Note that the number of elements in $\mathbf{P}^{\text{pre},i}$ is $3 \times n_i$ (n_i is the number of points in S_i); however, they are stacked as a long $3n_i \times 1$ column in (3). The subscript $(\cdot)_{[3 \times n_i]}$ is the reshaped version of $(\mathbf{P}^{\text{pre},i} + \Psi^i \mathbf{u}^i)$ from $3n_i \times 1$ to $3 \times n_i$. Also, note that $\Psi^i \mathbf{u}^i$ deforms the whole i th tissue (or object S_i) and not just a single point on S_i . We emphasize that \mathbf{R} , \mathbf{t} , and \mathbf{u}^i in (3) are the same for the 3D models regardless of which viewpoint of the M cameras is considered. Having N structures appear in M camera images, we define and minimize the following energy functional E to simultaneously align the preoperative models with the intraoperative images and segment them:

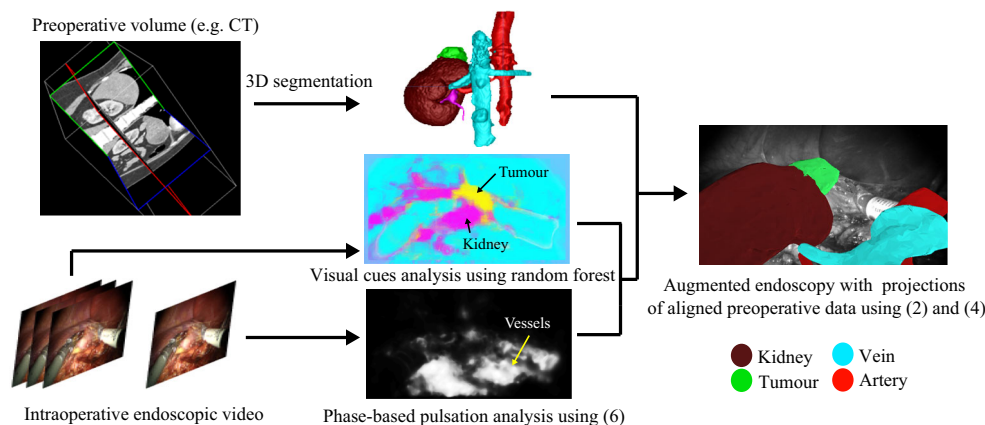


Fig. 1 Our surgical image guidance system leverages vessel pulsation and visual appearance cues to guide the 3D preoperative to endoscopic intraoperative alignment

$$E(\mathbf{T}, \mathbf{U}, \Phi; \mathbf{P}^{\text{pre}}, I_1, \dots, I_M, \pi) = \sum_{n=1}^N \sum_{m=1}^M \int_{\Omega_{2D}^m} (2\rho_m^n - 1)(\mathbf{x}) H(\phi_m^n(\mathbf{x})) d\mathbf{x}, \tag{4}$$

where $H(\cdot)$ is the Heaviside function, $\rho_m^n(\mathbf{x})$ are the regional terms that provide the cues towards which the level set is moved during the optimization and measure the agreement of the image pixels \mathbf{x} with the learnt statistical models (see ‘‘Data terms: appearance of structures in endoscopic images’’ section) of the n th structure in I_m , $\Phi = \{\phi_1^1, \dots, \phi_1^N, \dots, \phi_M^1, \dots, \phi_M^N\}$, and $\mathbf{U} = \{\mathbf{u}^1, \dots, \mathbf{u}^N\}$. The above energy function is similar to the Chan–Vese model [7] with the main difference lying in the optimization; where in the Chan–Vese method, the energy function is minimized with respect to the level set function ϕ whereas in our case, we optimize E with respect to the pose and shape parameters. To find the optimum pose and deformations of structures, we optimize E with respect to \mathbf{T} and \mathbf{U} (with Φ is updated accordingly).

The success of the proposed optimization depends on how accurate we model the 3D non-rigid deformation and how well we identify candidate pixels belonging to different structures in the endoscopic scenes. Some structures are hidden (e.g. occluded vessels), while others have complicated visual appearance. In the following sections, we explain how we choose our data terms [ρ_m^n in (4)] and how we generate the structures’ deformation and the final segmentation.

Data terms: appearance of structures in endoscopic images

In RAPN, the focus is on three organs: the vasculature (v), kidney (k), and tumour (t). In this context, we set the number of objects $N = 3$ in (4). The regional term of vessels, kidney, and tumour are represented with $\rho^1 = \rho^v$, $\rho^2 = \rho^k$, and $\rho^3 = \rho^t$, respectively.

Vascular pulsatile motion

Computing the regional term ρ^v by appearance alone is difficult as blood vessels are typically hidden under a layer of fat. These regions may, however, be identified by their characteristic pulsatile motion, which is detectable but invisible to the naked eye. This regional term is computed by first extracting pulsatile motion features from local phase information using our method proposed in [3]. In this section we briefly describe the framework for extracting these features and the corresponding data term used to drive our segmentation.

A video frame captured from camera m denoted as $I_m(\mathbf{x}, t)$ that maps the given pixel $\mathbf{x} \in \Omega_{2D}^m$ at time t to an intensity value can be represented as function of local displacements $d(\mathbf{x}, t)$ with respect to the first frame of the video

$I_m(\mathbf{x}, 0) = f(\mathbf{x})$ such that $I_m(\mathbf{x}, t) = f(\mathbf{x} + d(\mathbf{x}, t))$. Our regional term ρ^v is extracted from the local motions specified by $d(\mathbf{x}, t)$ and to approximate these motions from the video, we perform a wavelet decomposition of the video into sub-bands such that

$$I_m(\mathbf{x}, t) = f(\mathbf{x} + d(\mathbf{x}, t)) \approx \sum_{\omega=-\infty}^{\infty} A_{\omega} e^{i\omega(\mathbf{x} + \hat{d}(\mathbf{x}, t))}, \tag{5}$$

with each sub-band representing a complex sinusoid $S_{\omega}(\mathbf{x}, t) = A_{\omega} e^{i\omega(\mathbf{x} + \hat{d}(\mathbf{x}, t))}$ at spatial frequency ω and $\hat{d}(\mathbf{x}, t)$ is our approximation to the local motion. The local phase of each sub-band is defined as $\Phi_{\omega}(\mathbf{x}, t) = \text{arg}(S_{\omega}) = \omega(\mathbf{x} + \hat{d}(\mathbf{x}, t))$. Since the local $\omega \hat{d}(\mathbf{x}, t)$ is the only motion related component of the phase that varies with time, we can isolate it from the zero-frequency component $\omega \mathbf{x}$ by applying a mean-free temporal bandpass filter

$$\mathcal{H}(\mathbf{x}, t) = 2B_H \text{sinc}(2B_H t) - 2B_L \text{sinc}(2B_L t), \tag{6}$$

where B_L and B_H are the low and high frequency cut-off, respectively. The response of the temporal bandpass filter is denoted by $B_{\omega}(\mathbf{x}, t) = \Phi_{\omega} * \mathcal{H} = \omega \hat{d}_{\mathcal{H}}(\mathbf{x}, t)$, where $\hat{d}_{\mathcal{H}}$ are components of the motion that are in the passband of the filter. The passband of the filter is tuned to the typical heart rate of a patient so that we can isolate components of the local motion that are synchronous with the heart rate and hence to vascular pulsation.

The bandpassed phases B_{ω} are then denoised and combined across scales and orientations using the methods outlined in [2] to obtain fuzzy labels

$$L = \frac{1}{C} \sum_{\forall \omega} \frac{|\tilde{Q}_{\omega}|}{2\pi\omega}, \tag{7}$$

where C is a normalizing factor and \tilde{Q} is the denoised B_{ω} . To attenuate these false positives and obtain better visual cues of the hidden vessels, we process L in a temporal window (i.e. 2s video) by summing L in time and denoising the results with a spatial median filter. We denote the resulting denoised fuzzy labels by L^{avg} . Then, we calculate the probability of a pixel belonging to a pulsating vessel (artery and vein) as:

$$P^v(\mathbf{x} | L^{\text{avg}}) = \frac{1}{2} + \frac{1}{\pi} \arctan \left(\frac{L^{\text{avg}}(\mathbf{x}) - 0.3}{\epsilon} \right), \tag{8}$$

where we chose ϵ to be 0.1. The above equation maps L^{avg} to the probabilistic range of [0, 1]. The value 0.3 in (8), chosen empirically, intensifies any value more than 0.3 and discards small values in L^{avg} . The data term corresponding to vessels is defined as $\rho^v = -\log(P^v(\mathbf{x} | L^{\text{avg}}))$.

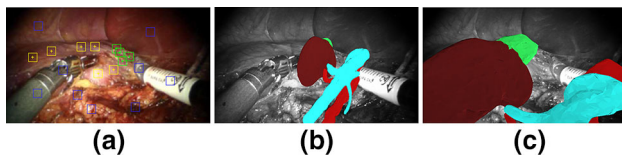


Fig. 2 Patch selection and initialization. **a** Patch selection (blue background/tools, yellow kidney, green tumour). **b** Probability map of background, kidney and tumour. **c** Initial pose (before convergence). **d** Recovered pose and shape (after convergence)

Learning of tissue appearance

To calculate the data term of a visible structure, we extract a variety of image features (\mathcal{A}_m) from each image m and train a RF classifier to distinguish between different structures. Particularly, we capture the textural patterns and colour variations via local colour histograms of the normalized RGB and YCbCr channels. The probability of pixel \mathbf{x} belonging to kidney ($P^k(\mathbf{x})$) and tumour ($P^t(\mathbf{x})$) are estimated by training a RF consisting of N_t binary decision trees. To train the RF, we select few 20×20 patches in I_m , $i = 1, \dots, M$ from different structures. In practice, surgeons may (virtually) select these patches with the help of surgical tools. We emphasize that unlike feature-based methods, e.g. [27], our method does not require any correspondence between 3D CT and the 2D intraoperative data. Figure 2a shows a sample seeding on an example endoscopic scene of real clinical data. After training, for each pixel \mathbf{x} , the feature channels, $\mathcal{A}_m(\mathbf{x})$, are propagated through each RF tree resulting in the probability $P^{n,j}(\mathbf{x}|\mathcal{A}_m(\mathbf{x}))$, for the j th tree and n th structure. These probabilities are combined into a forest's joint probability $P^n(\mathbf{x}|I_m(\mathbf{x})) = \frac{1}{N_t} \sum_{j=1}^{N_t} P^{n,j}(\mathbf{x}|\mathcal{A}_m(\mathbf{x}))$ to determine the probability of \mathbf{x} belonging to n th structure. Figure 2b illustrates examples of regions probability for the frame shown in Fig. 2a. The regional terms of kidney and tumour are then calculated as: $\rho_m^k = -\log P^k(\mathbf{x}|I_m(\mathbf{x}))$ and $\rho_m^t = -\log P^t(\mathbf{x}|I_m(\mathbf{x}))$, respectively.

Patient-specific heterogeneous deformation model

Due to the noisy endoscopic images, the regional terms alone are not able to provide robust cues for guiding the 3D–2D alignment. To obtain a reliable result, we constrain the space of possible transformations to patient-specific shape models obtained from the preoperative data (\mathbf{P}^{pre}). To account for non-rigid deformation of structures, we include the tissue-specific modes of vibration (Ψ^i) in our framework as described in “Problem statement and notation” section.

The modes of vibration Ψ^i are obtained by solving the generalized eigendecomposition problem: $\mathbf{K}^i \Psi^i = \mathbf{M}^i \Psi^i \Lambda^i$, where \mathbf{K} and \mathbf{M} are the stiffness and mass matrices, respectively, and Λ^i is a diagonal matrix of eigenvalues associated with the eigenvectors (Ψ^i) where higher eigenval-

ues corresponds to higher frequency. In contrast to traditional variational methods that update the level set function representing the segmentation to adapt it to image data, the use of Ψ favours shape deformations that are biomechanically plausible (as dictated by \mathbf{K} and \mathbf{M}).

As we increase the number of modes, more deformations are allowed, which is useful for accurately modelling the deformations. Given all possible modes of shape variability, we are able to produce/recover all possible deformations of that shape. However, as we include more deformation modes, the computational complexity increases and recovering the proper deformation parameters becomes more difficult. In addition, too many modes of variations may end up with extreme and unrealistic deformations. We empirically found that choosing more than 6 modes results in an unjustified increase in complexity, as deformations may become unrealistic and no noticeable improvement in accuracy is observed. According to Pentland et al. [24], since the lowest frequencies correspond to rigid body motions and global deformations, and we already encoded global deformations, translation and rotation, through \mathbf{R} and \mathbf{t} (Eq. (3)), we selected eigenvectors corresponding to larger eigenvalues to focus on finer deformation details.

We use the average Hounsfield unit (HU) associated with each structure in the preoperative CT to approximate the stiffness of each tissue. This way the stiffness of each structure is proportional to its tissue density. We used a simple linear elastic model to deform the objects of interest to lower the computational complexity; however, exploring more advanced elastic models might improve the results but with the cost of complexity. In the next section, we show how we optimize E in (4).

Optimization of the multi-structure pose and deformation

We optimize (4) with respect to \mathbf{T} , i.e. with respect to $\mathbf{R} = \{\alpha, \beta, \gamma\}$, $\mathbf{t} = \{t_x, t_y, t_z\}$, and $\mathbf{u} = \{u_1, \dots, u_q\}$, where α , β , and γ are the rotation angles around the x , y , and z axes, respectively and t_x , t_y , and t_z are the translations in x , y , and z directions. The derivative of E with respect to each of the unknown parameters is:

$$\frac{\partial E}{\partial \xi_\ell} = \sum_{n \in \{v, k, t\}} \sum_{m=1}^M \int_{\Omega_{2D}^m} \rho_m^n(\mathbf{x}) \frac{\partial H(\phi_m^n(\mathbf{x}))}{\partial \xi_\ell}, \quad (9)$$

where $\xi_\ell = \{\alpha, \beta, \gamma, t_x, t_y, t_z, u_1, \dots, u_q\}$ and

$$\begin{aligned} \frac{\partial H(\phi_m^n(\mathbf{x}))}{\partial \xi_\ell} &= \frac{\partial H(\phi_m^n(\mathbf{x}))}{\partial \phi_m^n} \left(\frac{\partial \phi_m^n}{\partial x} \frac{\partial x}{\partial \xi_\ell} + \frac{\partial \phi_m^n}{\partial y} \frac{\partial y}{\partial \xi_\ell} \right) \\ &= \delta(\phi_m^n) \left[\frac{\partial \phi_m^n}{\partial x} \frac{\partial \phi_m^n}{\partial y} \right] \begin{bmatrix} \frac{\partial x}{\partial \xi_\ell} \\ \frac{\partial y}{\partial \xi_\ell} \end{bmatrix}. \end{aligned} \quad (10)$$

In (10), $\delta(\cdot)$ is the Dirac delta function. We use the centred finite difference to calculate $\frac{\partial \phi}{\partial x}$ and $\frac{\partial \phi}{\partial y}$. Every 2D point $\mathbf{x} = (x, y)$ in I_m has at least one corresponding 3D point $\mathbf{P}^{\text{srg}} = (X^{\text{srg}}, Y^{\text{srg}}, Z^{\text{srg}})$ in the surgical domain Ω_{srg} . We calibrated the stereo camera and obtained the intrinsic camera parameters. For $m = 1$, \mathbf{x} and \mathbf{P}^{srg} are related by:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_{1x} & 0 & c_{1x} \\ 0 & f_{1y} & c_{1y} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X^{\text{srg}} \\ Y^{\text{srg}} \\ Z^{\text{srg}} \end{bmatrix}, \tag{11}$$

where $\pi_1 = (f_{1x}, f_{1y}, c_{1x}, c_{1y})$ are the first camera parameters. Hence, we have

$$\frac{\partial x}{\partial \xi_\ell} = \frac{f_{1x}}{Z^{\text{srg}^2}} \left(Z^{\text{srg}} \frac{\partial X^{\text{srg}}}{\partial \xi_\ell} - X^{\text{srg}} \frac{\partial Z^{\text{srg}}}{\partial \xi_\ell} \right). \tag{12}$$

We similarly calculate $\frac{\partial y}{\partial \xi_\ell}$. Also, each 3D point \mathbf{P}^{srg} is related to \mathbf{P}^{pre} by $\mathbf{P}^{\text{srg}} = \mathbf{R}\mathbf{P}^{\text{pre}} + \mathbf{t}$. Therefore, for $\xi_\ell \in \{\alpha, \beta, \gamma, t_x, t_y, t_z\}$, $\frac{\partial X^{\text{srg}}}{\partial \xi_\ell}$, $\frac{\partial Y^{\text{srg}}}{\partial \xi_\ell}$ and $\frac{\partial Z^{\text{srg}}}{\partial \xi_\ell}$ in (12) are easily calculated upon the choice of transformation function. To update the non-rigid deformation of structures, we need to calculate the derivatives of E with respect to the shape parameters u^i . These shape parameters u^i and the energy functional are related through the 2D coordinate x in E (Eq.(4)) which are derived from the corresponding 3D point in the surgical domain $\mathbf{P}^{\text{srg}} = (X^{\text{srg}}, Y^{\text{srg}}, Z^{\text{srg}})$ and the preoperative 3D point \mathbf{P}^{pre} . According to (3), \mathbf{P}^{pre} depends on the weights of the modes of vibration, u^i . Therefore, for the shape parameters $\xi_\ell \in \{u_1, \dots, u_k\}$, the derivative of a 3D point \mathbf{P}^{srg} in the surgical domain with respect to ξ_ℓ is $\frac{\partial \mathbf{P}^{\text{srg}}}{\partial \xi_\ell} = \mathbf{R} \cdot \psi_\ell$, where ψ_ℓ is the ℓ th mode of vibration in Ψ . To avoid any irrational shape deformation, we limit the shape parameters to vary not more than three times the standard deviation (Λ^i). For multiple camera views ($m > 1$), the extrinsic parameters ($\mathbf{R}_m^{\text{ext}}, \mathbf{t}_m^{\text{ext}}$) have to be considered in calculating the derivatives, i.e. \mathbf{R} in the above equations is replaced by $\mathbf{R}_m^{\text{ext}}\mathbf{R}$. The boundary of the segmented structures in I_m are the zero level set of their corresponding level set functions ϕ_m^n that are obtained by (2) after finding the optimal \mathbf{T} .

Materials and experiments

For validation, we applied our framework to fifteen different clinical cases of robot-assisted partial nephrectomy. All endoscopic videos were acquired by a da Vinci[®] Si surgical system (Intuitive Surgical, California, USA), and each frame was resized to 480×270 pixels for efficiency. The default parameters suggested in our previous works [2,3] were used to detect the vascular motion cues. We used $N_t = 70$ trees to train the RF for learning the appearance of kidney and tumour. Higher values of N_t did not improve accuracy but increased complexity.

We used the patient-specific 3D segmented kidney, tumour, and vasculature models and to set the tissue-specific modes of vibration, for simplicity, we assumed that the structures can be modelled as a set of unit masses mutually interconnected, i.e. \mathbf{M} is the identity matrix and can be removed from the eigendecomposition equation in our case. We set the stiffness of each tissue to be proportional to its corresponding HU (higher HU means higher density and hence higher stiffness). HU is calculated from the preoperative DICOM meta-data as $\text{HU} = \text{CT pixel value} \times \mu_s + \mu_i$, where μ_s and μ_i are the *rescale slope* and *rescale intercept* values that are stored in the CT meta-data. We manually initialized \mathbf{T} such that the projection of 3D models intersect the organs. This initialization does not need to be close to the solution. Figure 2c shows the initial pose, which despite being not well placed, results in a reasonable pose as shown in Fig. 2d. However, we emphasize that an irrational initialization will result in a wrong pose estimation due to our local optimization framework. We should also mention that if most of the surface of the objects is occluded in the 2D scene, our method cannot find the correct pose. For our experiments, we asked surgeons to stop moving the tools for ~ 10 s so our method can compute the regional term.

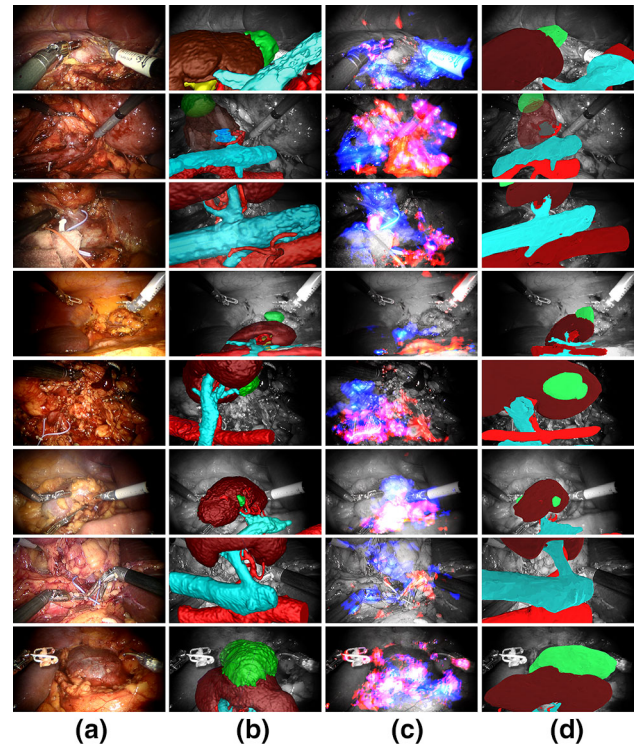


Fig. 3 First eight cases of the qualitative comparison of our proposed method with state-of-the-art method [3]. **a** Original endoscopic image. **b** The ground truth of venous (cyan), arterial (red), kidney (brown) and tumour (green) structures provided in [3]. **c** Segmentation results of vessels using [3]. **d** Our results. Kidney and tumour are shown in yellow and green, respectively

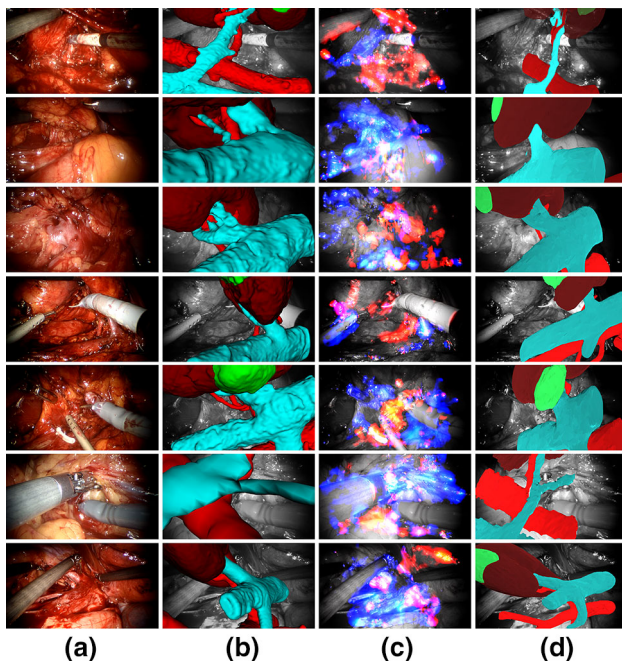


Fig. 4 Final seven cases of the qualitative comparison of our proposed method with state-of-the-art method [3]. **a** Original endoscopic image. **b** The ground truth of venous (cyan), arterial (red), kidney (brown) and tumour (green) structures provided in [3]. **c** Segmentation results of vessels using [3]. **d** Our results. Kidney and tumour are shown in yellow and green, respectively

We compared the segmentations obtained through our guidance system with the same ground truth presented in [2,3] (Figs. 3, 4). Note how the noisy segmentations in (c) are improved in (d) by incorporating the preoperative prior information. We also quantitatively compared our proposed method with [2] in Table 2.

The average runtime of our unoptimized MATLAB code to process the vessel pulsation in a 4-s clip (120 frames) was 65 s. The runtime for pose estimation and segmenting the structures depends on the initial pose of the organs. The average runtime to find the pose and segment the structures for an initialization similar to Fig. 2c is ~ 16 s on a standard 3.40 GHz CPU.

Discussion and conclusions

We proposed a new technique for localizing both visible and occluded structures in an endoscopic view by estimating the

3D pose and deformations of structures of interest in the 3D surgical space. Our framework leverages both preoperative data, as a source of patient-specific prior knowledge, as well as vasculature pulsation (by analysing the local phase information) and endoscopic visual cues (by training a random decision forest) in order to accurately segment the highly noisy and cluttered environment of an endoscopic video. To handle the non-rigid deformation of different structures, we incorporated a tissue-specific physically based deformation model. To make the non-rigid deformation of each structure closer to reality, we used the HU value of each structure in the preoperative CT and assigned a specific stiffness to each deformable model. Our results on in vivo clinical cases of partial nephrectomy illustrate the potential of the proposed framework for augmented reality applications in MIS.

There are several directions to extend this work. Our variational framework is highly parallelizable, and we do not foresee any obstacles towards a GPU implementation for real-time pose estimation and endoscopic video segmentation. In addition, we believe that leveraging stereo views as well as encoding depth information into the proposed energy functional can improve the performance.

We attribute some of the observable differences between the ground truth, and our results to both the local optimization framework we used and also to the error in the alignment of the ground truth. As mentioned in Amir-Khalili et al. [2], due to the fact that the preoperative model was rigidly aligned to the endoscopic video, an alignment error of 4–7 mm exist in cases where the organs have been significantly retracted by the surgical instruments or mobilization of other organs. We believe that despite the visible differences between the two, our current solution is one step closer to an ideal solution compared to the ground truth as our current method allows for non-rigid modes of vibration. Generating a ground truth that accounts for the non-rigid deformations due to mobilization and retraction requires volumetric intraoperative imaging such as cone beam CT or possibly implanting fiducials. The use of such imaging techniques is not feasible as it exposes the patient and clinicians to ionizing radiation and implanting fiducials is intrusive and invasive and hence not recommended.

Also, in our future work, we will explore the use of an additional shape variation component that is orthogonal to the restricted shape model, as described by Andrews

Table 2 Quantitative comparison for kidney (K) and vessel (V) segmentation: our method versus [2]

Method	DSC		TPR		FPR		F1-measure		Accuracy	
	K	V	K	V	K	V	K	V	K	V
Amir-Khalili et al. [2]	–	0.41	–	0.74	–	0.40	–	0.41	–	0.60
Our method	0.70	0.61	0.70	0.56	0.07	0.06	0.70	0.61	0.88	0.87

The best results are highlighted in bold

DSC dice similarity coefficient, TPR true positive ratio, FPR false positive ratio

and Hamarneh [4], since this allows for exploring larger shape variability without noticeable increase in complexity. Although we limited the modes of vibration to vary not more than three times the corresponding eigenvalue (to avoid any irrational shape deformation), we still might get similar projection from two different 3D deformations. This is due to the fact that we lose information during the 3D to 2D transformation. We believe that this is another interesting future direction that worth investigation.

Given that in this proposed method we used a local optimization technique, leveraging our own group and others that have worked on convexification techniques [4–6,9,19,22] can make the method less sensitive (or insensitive) to initialization.

Finally, improved estimates of elasticity parameters (e.g. using elastography imaging) will likely more accurately constrain the space of non-rigid deformations.

Acknowledgments This publication was made possible by NPRP Grant #4-161-2-056 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent This articles does not contain patient information.

References

- Agudo A, Agapito L, Calvo B, Montiel J (2014) Good vibrations: a modal analysis approach for sequential non-rigid structure from motion. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1558–1565
- Amir-Khalili A, Hamarneh G, Peyrat JM, Abinahed J, Al-Alao O, Al-Ansari A, Abugharbieh R (2015) Automatic segmentation of occluded vasculature via pulsatile motion analysis in endoscopic robot-assisted partial nephrectomy video. *Med Image Anal* 25(1):103–110
- Amir-Khalili A, Peyrat JM, Abinahed J, Al-Alao O, Al-Ansari A, Hamarneh G, Abugharbieh R (2014) Auto localization and segmentation of occluded vessels in robot-assisted partial nephrectomy. In: Medical image computing and computer-assisted intervention (MICCAI), pp 407–414
- Andrews S, Hamarneh G (2015) The generalized log-ratio transformation: learning shape and adjacency priors for simultaneous thigh muscle segmentation. *IEEE Trans Med Imaging* 34(9):1773–1787
- Andrews S, McIntosh C, Hamarneh G (2011) Convex multi-region probabilistic segmentation with shape prior in the isometric log-ratio transformation space. In: IEEE international conference on computer vision (ICCV). IEEE, pp 2096–2103
- Brown E, Chan T, Bresson X (2009) Convex formulation and exact global solutions for multi-phase piecewise constant Mumford–Shah image segmentation. *UCLA CAM report*, pp 09–66
- Chan TF, Vese L et al (2001) Active contours without edges. *IEEE Trans Image Process* 10(2):266–277
- Crane NJ et al (2010) Visual enhancement of laparoscopic partial nephrectomy with 3-charge coupled device camera: assessing intraoperative tissue perfusion and vascular anatomy by visible hemoglobin spectral response. *J Urol* 184(4):1279–1285
- DeLong A, Boykov Y (2009) Globally optimal segmentation of multi-region objects. In: IEEE international conference on computer vision (IEEE ICCV), pp 285–292
- Escudier B, Kataja V et al (2010) Renal cell carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 21(Suppl 5):v137–v139
- Estépar RSJ, Westin CF, Vosburgh KG (2009) Towards real time 2D to 3D registration for ultrasound-guided endoscopic and laparoscopic procedures. *Int J Comput Assist Radiol Surg* 4(6):549–560
- Figueiredo IN, Figueiredo PN, Stadler G, Ghattas O, Araujo A (2010) Variational image segmentation for endoscopic human colonic aberrant crypt foci. *IEEE Trans Med Imaging* 29(4):998–1011
- Figueiredo IN, Moreno JC, Prasath VBS, Figueiredo PN (2012) A segmentation model and application to endoscopic images. In: Image analysis and recognition. Springer, pp 164–171
- Gill IS, Desai MM, Kaouk JH, Meraney AM, Murphy DP, Sung GT, Novick AC (2002) Laparoscopic partial nephrectomy for renal tumor: duplicating open surgical techniques. *J Urol* 167(2):469–476
- Gill IS, Kavoussi LR, Lane BR, Blute ML, Babineau D, Colombo JR Jr, Frank I, Permpongkosol S, Weight CJ, Kaouk JH et al (2007) Comparison of 1,800 laparoscopic and open partial nephrectomies for single renal tumors. *J Urol* 178(1):41–46
- Gill S, Abolmaesumi P, Vikal S, Mousavi P, Fichtinger G (2008) Intraoperative prostate tracking with slice-to-volume registration in MR. In: International conference of the society for medical innovation and technology, pp 154–158
- Hernes N, Toril A, Lindseth F, Selbekk T, Wolff A, Solberg OV, Harg E, Rygh OM, Tangen GA, Rasmussen I et al (2006) Computer-assisted 3D ultrasound-guided neurosurgery: technological contributions, including multimodal registration and advanced display, demonstrating future perspectives. *Int J Med Robot Comput Assist Surg* 2(1):45–59
- Hummel J, Figl M, Bax M, Bergmann H, Birkfellner W (2008) 2D/3D registration of endoscopic ultrasound to CT volume data. *Phys Med Biol* 53(16):4303
- McIntosh C, Hamarneh G (2009) Optimal weights for convex functionals in medical image segmentation. In: Advances in visual computing. Springer, pp 1079–1088
- Merritt SA, Rai L, Higgins WE (2006) Real-time CT-video registration for continuous endoscopic guidance. In: Medical imaging. International Society for Optics and Photonics, pp 614313–614313
- Mewes PW, Neumann D, Licegevic O, Simon J, Juloski AL, Angelopoulou E (2011) Automatic region-of-interest segmentation and pathology detection in magnetically guided capsule endoscopy. In: Medical image computing and computer-assisted intervention (MICCAI 2011). Springer, pp 141–148
- Nosrati MS, Andrews S, Hamarneh G (2013) Bounded labeling function for global segmentation of multi-part objects with geometric constraints. In: IEEE international conference on computer vision (ICCV), pp 2032–2039
- Nosrati MS, Peyrat JM, Abinahed J, Al-Alao O, Al-Ansari A, Abugharbieh R, Hamarneh G (2014) Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors. In: Medical image computing and computer-assisted intervention (MICCAI), pp 324–331

24. Pentland A, Sclaroff S (1991) Closed-form solutions for physically based shape modeling and recognition. *IEEE Trans Pattern Anal Mach Intell (IEEE TPAMI)* 7:715–729
25. Pickering MR, Muhit AA, Scarvell JM, Smith PN (2009) A new multi-modal similarity measure for fast gradient-based 2D–3D image registration. In: *Engineering in medicine and biology society. EMBC 2009. Annual international conference of the IEEE. IEEE*, pp 5821–5824
26. Pratt P, Mayer E, Vale J, Cohen D, Edwards E, Darzi A, Yang GZ (2012) An effective visualisation and registration system for image-guided robotic partial nephrectomy. *J Robot Surg* 6(1):23–31
27. Puerto-Souza GA, Mariottini GL (2013) Toward long-term and accurate augmented-reality display for minimally-invasive surgery. In: *IEEE international conference on robotics and automation (ICRA). IEEE*, pp 5384–5389
28. Singh I (2009) Robot-assisted laparoscopic partial nephrectomy: current review of the technique and literature. *J Min Access Surg* 5(4):87
29. Teber D et al (2009) Augmented reality: a new tool to improve surgical accuracy during laparoscopic partial nephrectomy? Preliminary in vitro and in vivo results. *Eur Urol* 56(2):332–338
30. Tobis S et al (2011) Near infrared fluorescence imaging with robotic assisted laparoscopic partial nephrectomy: initial clinical experience for renal cortical tumors. *J Urol* 186(1):47–52
31. Urban BA et al (2001) Three-dimensional volume-rendered CT angiography of the renal arteries and veins: normal anatomy, variants, and clinical applications. *RadioGraphics* 21(2):373–386
32. Yim Y, Wakid M, Kirmizibayrak C, Bielowicz S, Hahn J (2010) Registration of 3D CT data to 2D endoscopic image using a gradient mutual information based viewpoint matching for image-guided medialization laryngoplasty. *J Comput Sci Eng* 4(4):368–387
33. Zikic D, Glocker B, Kutter O, Groher M, Komodakis N, Khamene A, Paragios N, Navab N (2010) Markov random field optimization for intensity-based 2D–3D registration. In: *SPIE medical imaging. International Society for Optics and Photonics*, pp 762334–762334