Hand-Eye Calibration of Surgical Instrument for Robotic Surgery Using Interactive Manipulation

Fangxun Zhong, *Student Member, IEEE*, Zerui Wang, Wei Chen, Kejing He, Yaqing Wang, *Member, IEEE*, and Yun-Hui Liu, *Fellow, IEEE*

***Abstract*—Conventional robot hand-eye calibration methods are impractical for localizing robotic instruments in minimally-invasive surgeries under intra-corporeal workspace after pre-operative set-up. In this letter, we present a new approach to autonomously calibrate a robotic instrument relative to a monocular camera without recognizing calibration objects or salient features. The algorithm leverages interactive manipulation (IM) of the instrument for tracking its rigid-body motion behavior subject to the remote center-of-motion constraint. An adaptive controller is proposed to regulate the IM-induced instrument trajectory, using visual feedback, within a 3D plane which is observable from both the robot base and the camera. The eye-to-hand orientation and position are then computed via a dual-stage process allowing parameter estimation in low-dimensional spaces. The method does not require the exact knowledge of instrument model or large-scale data collection. Results from simulations and experiments on the da Vinci Research Kit are demonstrated via a laparoscopy resembled set-up using the proposed framework.**

***Index Terms*—Calibration and identification, sensor-based control, surgical robotics: laparoscopy, medical robots and systems.**

I. Introduction

THE trend of researching robotic minimally-invasive surgery (RMIS) is towards supervised autonomy of sub-task execution [1]. Explorations have so far been made to automate non-critical procedures including tumour localization [2], endoscope positioning [3], suturing [4], [5], etc., aiming for standardizing outcomes and reducing human workload. Performing a delicate surgical task autonomously requires precise instrument positioning to provide safe and reliable interaction with the surgical field, which arises the need for knowing the instrument's pose from external sensors to facilitate sensor-based robot control. In RMIS, this is achievable via online instrument tracking using the endoscope as a versatile sensor. However, it is prone to detection failure due to dynamic visual conditions during complex manipulation steps [6], [7], reducing its long-term reliability. A more practical solution is to perform hand-eye calibration such that the pose can be continuously retrieved via robot kinematics data [8].

The hand-eye calibration problem has been well addressed with applications to industrial robot arms, but yet has gained limited traction in RMIS. A major reason is that most existing methods rely on external visual patterns [9], [10], which are not applicable to RMIS as the instruments have reached intra-corporeal space upon pre-operative set-up before the calibration part. The confined instrument workspace and camera's field of view also limit the feasibility of multi-group data acquisition. To cater for minimally-invasive set-up, robotic surgical instruments own articulated structure whose motions are constrained by the remote center-of-motion (RCM) [11]. To achieve hand-eye calibration via direct recognition of the instrument's motion properties becomes of great potential to improve its practicality in RMIS.

In this paper, we propose a new autonomous hand-eye calibration framework for robotic instruments in RMIS. The study firstly applies interactive manipulation (IM) into hand-eye calibration to enrich visual sensory information of the instrument (observed from a monocular camera) for data acquisition. The contribution of this work is three-fold. First, a new parametrization method is introduced to characterize the IM-induced instrument trajectory, subject to 3-degree-of-freedom (3-DoF) RCM-constrained motions, by proposing the interactive feature plane (IFP). Next, an adaptive controller is designed to online regulate the spatial properties of IFP via visual tracking of the instrument's rigid-body motion behavior, such that the IFP could be mathematically derived from both the camera frame and the robot base. Finally, we develop a new computation method to retrieve the orientation and position term of eye-hand transformation using a dual-step computation in low-dimensional spaces based on the settled IFPs. The proposed technique has the following advantages:

- The instrument motions are fully automated during the calibration. Camera motions are not required.
- No external calibration objects, salient features, the instrument's exact CAD models or offline training process are required for our calibration method.
- The method is free of large-scale data collection, 3-DoF IM-induced joint motions suffice the online processing.

Manuscript received September 10, 2019; accepted January 5, 2020. Date of publication January 20, 2020; date of current version February 4, 2020. This letter was recommended for publication by Associate Editor P. Valdastri and Editor E. De Momi upon evaluation of the reviewers' comments. This work is supported in part by the HK RGC TRS under T42-409/18-R, in part by the CUHK-SJTU Joint Research Fund Project under Grant 4750352, in part by the CUHK T Stone Robotics Institute VC Fund 4930745 , CUHK, and in part by Shenzhen Science and Technology Program Grant KQTD 20140630150243062. *(Corresponding author: Zerui Wang.)*

The authors are with the T Stone Robotics Institute and Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, HKSAR, China (e-mail: fxzhong@mae.cuhk.edu.hk; zerui.j.wang@gmail.com; wchen@cuhk.edu.hk; yaqingwang@cuhk.edu.hk; yhliu@mae.cuhk.edu.hk).

This article has supplementary downloadable multimedia material available at https://ieeexplore.ieee.org provided by the authors.

Digital Object Identifier 10.1109/LRA.2020.2967685

Two groups of IFPs are sufficient for full recovery of the eye-hand transformation within minimal workspace.
- Recovery of the orientation and position terms is now via low-dimensional (3-DoF) spaces using online visual feedback which alleviates the error propagation from orientation to position in instrument localization.

II. Related Work

Numerous studies have addressed surgical instrument localization based on online sensory information. Attempts have been made using analytic methods to recover the instrument's pose via its geometric appearances [3], [12]–[14], whose results are highly sensitive to detection noises. One popular approach is the tracking-by-detection strategy which relies on salient features and/or virtual rendering of the instrument's CAD model [15]–[17]. However, the use of high-dimensional optimization via online feature tracking is neither computationally efficient nor reliable for real-time pose monitoring. The works in [18]–[22] suggest end-to-end pose estimation using learning-based methods which involve offline training. Meanwhile, considering kinematics data from robot joint encoders is also reported by [23]–[26] to increase the estimation accuracy of the instrument's pose (see [7] for a more comprehensive review of surgical instrument localization).

There are also works focusing on the hand-eye calibration of instruments for robust pose data acquisition with weak dependence on external sensors. For example, Mourgues *et al.* [27] developed an eye-hand calibration method using a robot-actuated stereo endoscope for patient-side data visualization. Schmidt *et al.* [28] solved the hand-eye calibration between the endoscope and the surgical robot using dual-quaternion transformation representation. Similar method is adopted by [8] and [29] upon multi-group data collection. In addition, Pachtrachai *et al.* calibrated the eye-hand information of a stereo laparoscope mounted on an industrial robot arm (or eye-in-hand set-up) using an adjoint transformation, with [30] and without [31] the RCM constraint, respectively. Notably, all these algorithms rely on an external calibration object to provide stable visual patterns. To avoid this, the work in [32] proposed to align the synthetic data from the instrument's CAD model with shape-based tracking results to solve the eye-in-hand transformation. A lowest 20-mm/10-degree position/rotation error of the estimation accuracy is achieved. Wang *et al.* [33] used the projected centerline of the instrument shaft to compute the in-between pose of two RCM-constrained robot arms on the da Vinci Surgical System (dVSS, one with the instrument, the other the endoscope). However, the algorithm depends on data collection of different robot configurations (256 groups in the experiment) to achieve a $\sim$10 mm position error.

III. Preliminaries

A. Robot Kinematics

It is well-known that the end-effector velocity of a 6-DoF serial robot manipulator (joint positions denoted by $\mathbf{q} = [q_1 \ \dots \ q_6]^\intercal$) in 3D Cartesian space is derived using the Jacobian matrix $\mathbf{J}(\mathbf{q}) \in \mathbb{R}^{3\times 6}$ as follows:

$$\dot{\mathbf{x}} = \mathbf{J}(\mathbf{q})\dot{\mathbf{q}} \tag{1}$$

where $\dot{\mathbf{x}} \in \mathbb{R}^3$ and $\dot{\mathbf{q}} \in \mathbb{R}^6$ denote the position velocity and the joint velocity, respectively. Consider a robotized surgical instrument implemented to the da Vinci Research Kit (dVRK) for RMIS without loss of generality. We assign the origin of robot base frame $\mathcal{F}_0(O_0, x_0, y_0, z_0)$ to the RCM point to eliminate kinematic parameters of the (unmoved) passive joints. The instrument's first three DoFs (q_1 and q_2 for rotation, q_3 for translation) generate RCM-constrained motions through a world-fixed pivot point without axial rotations. If one defines a new generalized coordinate vector $\mathbf{q}_s = [q_1 \ q_2]^\intercal$, then, for any configurations yielding $q_i = 0, \ \forall i \in \{4, 5, 6\}$ and $\dot{q}_3 = 0$, the end-effector (or the instrument tip) velocity with respect to the robot base reduces to the following form:[1]

$$\dot{\mathbf{x}} = \lambda(q_3)\mathbf{J}_s(\mathbf{q}_s)\dot{\mathbf{q}}_s \tag{2}$$

where $\mathbf{J}_s(\mathbf{q}_s) \in \mathbb{R}^{3\times 2}$ is a low-dimensional Jacobian matrix, $\lambda(q_3)$ denotes the distance of the instrument tip related to q_3 that passes the RCM. This implies that the orientation of the instrument shaft is solely determined by varying $\mathbf{q}_s$, which is an important property to investigate the RCM-constrained motion behavior for our modelling in Section IV.

B. Problem Formulation

We aim to solve the homogeneous transformation matrix of the camera with respect to the robot base (or the hand-eye transformation), from which the forward kinematics of the robotic instrument is precisely known. A monocular camera is used in our method whose intrinsic parameters are calibrated in advance. During the calibration step, the instrument is kinematically controlled with its RCM position remaining unchanged but beyond the camera's field of view. This corresponds to a normal pre-operative set-up in RMIS. The articulated structure and the (cylindrical) instrument shaft is fully and partially observable by the camera, respectively. To track the rigid-body motion behavior of the instrument, we assume the centerline of the instrument region detected from the 2D image to be exactly the projected line of the instrument's shaft center. Note importantly that, ideally, they might not be necessarily coincided with each other in the image due to perspective projection and thus it will introduce theoretical calibration error. However, we will demonstrate that such approximation leads to a simplified modelling robust to image feedback and does not significantly affect the calibration accuracy.

IV. Modelling

A. Interactive Manipulation (IM)

A control strategy that introduces deliberate actions to the external sensor(s) and/or the manipulating target to reveal

[1]Except further explanations, the appearance of $\mathbf{q}_s$ in our subsequent modelling implies the configuration that $q_i = 0, \ \forall i \in \{4, 5, 6\}$ and $\dot{q}_3 = 0$. This enforces the robotic instrument tip to land on the instrument shaft to facilitate visual inspection of its rigid-body motion behavior by the camera.

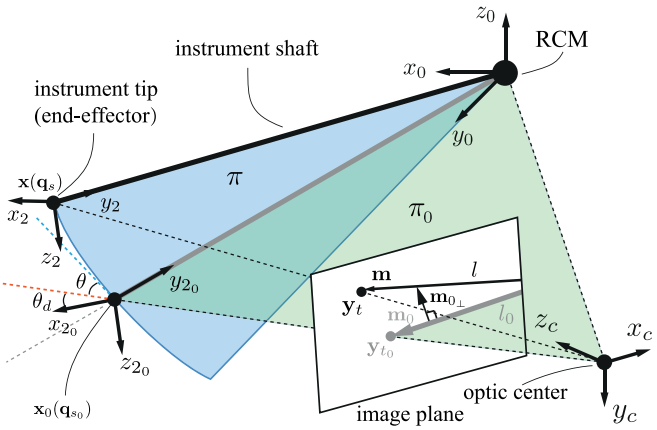

Fig. 1. Geometric interpretation of the robot-camera model. The instrument shaft (solid black line with $\mathbf{x}(\mathbf{q}_s)$ moves from its original position (solid gray line) $\mathbf{x}_0(\mathbf{q}_{s_0})$. The motion (parametrized by θ) is regulated until its swept IFP π (in blue) is settled towards the pre-defined plane π_0 from image feedback, with a converged θ_d (as l moves to l_0 in the image plane).

additional sensory feedback for task-relevant input, which is otherwise not available, is referred to as the interactive perception [34]. It has been widely applied to robot manipulation tasks involving physical interactions with the environment [35], [36]. Here, we propose the concept of IM that endows the robotic instrument with a pre-set motion trajectory whose spatial property in 3D Cartesian space also interacts with the visual data subject to a feedback controller. To start with, we propose the following feature vector

$$\mathbf{s} = \begin{bmatrix} \theta & \phi \end{bmatrix}^\intercal. \tag{3}$$

To utilize IM, we generate a pendulum resembling trajectory for $\mathbf{x}$ in 3D Cartesian space by moving only the first two DoFs. Derive $\mathbf{x}$ subject to (2) with respect to its initial configuration $\mathbf{q}_{s_0}$ as follows:

$$\mathbf{x} = \lambda(q_3)\mathbf{R}_0(\mathbf{q}_{s_0})\underbrace{\mathbf{R}_u(\mathbf{u}(\theta),\phi)\mathbf{v}_t}_{\mathbf{v}_u} \tag{4}$$

where $\mathbf{R}_0(\mathbf{q}_{s_0}) \in \mathbb{R}^{3\times3}$ denotes the constant initial rotation between the frame ($\mathcal{F}_2$) from the base frame ($\mathcal{F}_0$), as illustrated in Fig. 1. The vector $\mathbf{v}_t = [0\ -1\ 0]^\intercal$ is the fixed instrument position under $\mathcal{F}_2$ from forward kinematics. Note that $\mathbf{x}_0$ is pre-set to be safe for IM-induced motions in intra-corporeal space and is observable by the camera. The matrix $\mathbf{R}_u(\theta,\ \phi) \in \mathbb{R}^{3\times3}$ describes the resultant instrument shaft rotation relative to $\mathbf{R}_0$, from which θ depicts the plane orientation under frame $\mathcal{F}_0$, and ϕ parametrizes the rotation magnitude, both with respect to the initial robot configuration, respectively. We give the form of $\mathbf{u}(\theta)$ as

$$\mathbf{u}(\theta) = \begin{bmatrix} -\cos\theta & 0 & -\sin\theta \end{bmatrix}^\intercal \tag{5}$$

which suggests the IM-induced trajectory of the instrument shaft subject to $\mathbf{q}_s$ stays within nowhere but a 3D plane π whose normal vector is exactly $\mathbf{n}_\pi = \mathbf{R}_0(\mathbf{x}\times\mathbf{u}(\theta))$. We name the plane π as the *interactive feature plane* (IFP) whose spatial property θ is online adjustable via visual inspection of the instrument's rigid-body motion behavior. Taking derivative of (4) and substitute it to (2) yields

$$\dot{\mathbf{q}}_s = \mathbf{M}(\mathbf{q}_s,\theta,\phi)\dot{\mathbf{s}} \tag{6}$$

where $\mathbf{M}(\cdot) \in \mathbb{R}^{2\times2}$ is further derived by

$$\mathbf{M}(\cdot) = \mathbf{J}_s^{-1}(\mathbf{q}_s)\mathbf{R}_0(\mathbf{q}_{s_0})\begin{bmatrix} \dfrac{\partial\mathbf{v}_u}{\partial\theta} & \dfrac{\partial\mathbf{v}_u}{\partial\phi} \end{bmatrix} \tag{7}$$

Up to now, a relationship between the feature vector $\mathbf{s}$ that characterizes the IM-induced instrument motion and the corresponding joint space velocity has been obtained, which is powerful for initiating vision-based trajectory regulation to reveal the hand-eye transformation.

B. Vision-Based Adaptive Controller

Next, we design an adaptive controller to online regulate the spatial property of the instrument tip trajectory by tracking its rigid-body motion behavior. First, we consider an arbitrary 3D point lying on the shaft center of the instrument which is within the field-of-view of the camera denoted in homogeneous form $\mathbf{p} \in \mathbb{R}^4$

$$\mathbf{p} = \begin{bmatrix} \gamma\mathbf{x}(\mathbf{q}_s)^\intercal & 1 \end{bmatrix}^\intercal \tag{8}$$

where γ is an arbitrary positive scalar that renders $\mathbf{p}$ virtually constrained to the instrument shaft for our geometric modelling. Then its projected 2D point on the camera image can be computed as follows:

$$\begin{bmatrix} \mathbf{y}^\intercal & 1 \end{bmatrix}^\intercal = \frac{\gamma}{{}^c z(\mathbf{q}_s)}\mathbf{K}\mathbf{T}^{-1}\mathbf{p}(\mathbf{q}_s) \tag{9}$$

with $\mathbf{K} \in \mathbb{R}^{3\times4}$ the known camera intrinsic matrix. $\mathbf{T} \in \mathbb{R}^{4\times4}$ is the unknown hand-eye transformation matrix with

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix}. \tag{10}$$

The term ${}^c z(\mathbf{q}_s) = \mathbf{r}_3^\intercal\mathbf{x}(\mathbf{q}_s) + t_3$ in (9) denotes the depth to the camera, with $\mathbf{r}_3$ and t_3 being the third row in $\mathbf{R} \in \mathbb{R}^{3\times3}$ and the third element in $\mathbf{t} \in \mathbb{R}^3$, respectively. To characterize the visual feedback, we define a 2D vector $\mathbf{m}_0 \in \mathbb{R}^2$ which denotes orientation of the projected centerline l_0 of the instrument shaft in the image upon a detectable $\mathbf{y}$. As the instrument moves from a pre-set configuration, the orientation of the new centerline l is denoted by $\mathbf{m} \in \mathbb{R}^2$. Here we derive their relative distance via vector projection

$$d = \mathbf{m}^\intercal\mathbf{m}_{0_\perp} \tag{11}$$

where $\mathbf{m}_{0_\perp}$ depicts a unit vector perpendicular to $\mathbf{m}_0$. We then differentiate (11) to obtain the following relationship

$$\dot{d} = \mathbf{m}_{0_\perp}^\intercal\dot{\mathbf{y}} = \eta(\mathbf{q}_s)\dot{\mathbf{q}}_s \tag{12}$$

where $\eta(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ maps the joint velocity subject to the change of 2D line distance in the image, with

$$\eta(\cdot) = \frac{\lambda({}^c z(\mathbf{q}_s) - \mathbf{x}\mathbf{r}_3^\intercal)}{{}^c z^2(\mathbf{q}_s)}\mathbf{m}_{0_\perp}^\intercal\mathbf{K}\mathbf{T}^{-1}\mathbf{J}_s(\mathbf{q}_s)\dot{\mathbf{q}}_s. \tag{13}$$

Combining (6) and (12) results in the following form:

$$\dot{d} = \underbrace{\dot{\phi}\frac{\partial d}{\partial \mathbf{q}_s}(\mathbf{q}_s)\mathbf{M}(\mathbf{q}_s, \mathbf{s})}_{\mathbf{A}(\cdot)} \begin{bmatrix} \zeta & 1 \end{bmatrix}^\intercal \tag{14}$$

from which $\dot{\theta} = \zeta\dot{\phi}$, $\eta(\cdot) = \partial d(\mathbf{q}_s)/\partial \mathbf{q}_s$, and $\mathbf{A}(\cdot) \in \mathbb{R}^{1\times 2}$ is the overall interaction matrix. Importantly, this shows that a given visual data input $\dot{d}$ could be satisfied via (14) by initiating a corresponding $\dot{\theta}$ regardless of $\dot{\phi}$, which is precisely known to generate IM-induced motions. As $\mathbf{T}$ occurs as a factored form in (13), the relationship (14) could be further arranged into the following with respect to unknown constant terms:

$$\dot{d} = \mathbf{W}(\mathbf{q}_s, \dot{\mathbf{q}}_s, \mathbf{s}, \dot{\mathbf{s}})\mathbf{a} \tag{15}$$

where $\mathbf{W}(\cdot) \in \mathbb{R}^{1\times 2}$ is a regressor matrix constructed solely by online-measurable data, the vector $\mathbf{a} \in \mathbb{R}^l$ contains the constant eye-hand transformation. Giving an estimate of $\mathbf{a}$ by $\hat{\mathbf{a}}$ leads to the following estimation error:

$$e_d = \hat{\dot{d}} - \dot{d} = \mathbf{W}(\hat{\mathbf{a}} - \mathbf{a}) \tag{16}$$

To stabilize the error e_d to zero, we implement the following updating rule to $\hat{\mathbf{a}}$ based on continuous monitoring of e_d:

$$\frac{d}{dt}\hat{\mathbf{a}} = -\mathbf{\Gamma}\mathbf{W}^\intercal e_d, \tag{17}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{l\times 2}$. Then the asymptotic convergence of $\Delta\mathbf{a} = \hat{\mathbf{a}} - \mathbf{a}$ can be guaranteed by considering the Lyapunov-like quadratic function $V = \frac{1}{2}(\Delta\mathbf{a})^\intercal(\Delta\mathbf{a})$ such that $\dot{V} = -e_d\mathbf{W}\mathbf{\Gamma}^\intercal\mathbf{\Gamma}\mathbf{W}^\intercal e_d \leq 0$ which proves the stability.

Remark 1: The convergent performance of $\hat{\mathbf{a}}$ does not necessarily indicate an accurate estimate of the comprising hand-eye transformation but only contributes to a stabilized control system to determine IFPs (to appear in Section IV-C).

Now we derive how the input of visual feedback $\dot{d}$ regulates the spatial property $\dot{\theta}$ of the IFP π subject to IM-induced trajectory. We propose the following controller:

$$\mathbf{u} = -\kappa\hat{\mathbf{A}}^+\frac{\partial Q}{\partial d} \tag{18}$$

where $\hat{\mathbf{A}}^+$ is the pseudoinverse and κ is a positive scalar, $\mathbf{u} = [u_1\ 1]^\intercal$ with $u_1 = \zeta$ is the control input applied to (14) to stably minimize d to 0 over time using $Q = \frac{1}{2}d^2$ as a cost function [37], such that $\mathbf{m}$ aligns with $\mathbf{m}_0$ which further stabilizes θ despite the change of ϕ. A constant scalar ϵ is used to detect the settlement of d over a time period T to deal with the feedback noise. The spatial property θ of the IFP is then determined once the following inequality holds

$$\int_{t_0}^{t_0+T} |\dot{\theta}(t)|dt < \epsilon, \quad t \in [t_0,\ t_0 + T]. \tag{19}$$

C. Computation of Hand-Eye Transformation

In this subsection, we compute the hand-eye transformation based on the settled IFPs. Denote the stabilized θ by θ_d, the leading 2D projection of the IFP in the image plane is precisely the centerline l_0. The normal vector of the settled IFP π in the camera frame can be described under both the camera frame and the robot base frame as follows:[2]

$${}^c\mathbf{n}_\pi = \frac{\mathbf{c}_{l_1} \times \mathbf{c}_{l_2}}{||\mathbf{c}_{l_1} \times \mathbf{c}_{l_2}||}, \quad {}^b\mathbf{n}_\pi = \frac{\mathbf{u}(\theta_d) \times \mathbf{v}_t}{||\mathbf{u}(\theta_d) \times \mathbf{v}_t||} \tag{20}$$

where $\mathbf{c}_{l_1}$ and $\mathbf{c}_{l_2}$ are computed from

$$\mathbf{c}_{l_1} = f\mathbf{K}^{-1}\begin{bmatrix}\mathbf{y}_{l_1}^\intercal & 1\end{bmatrix}^\intercal, \quad \mathbf{c}_{l_2} = f\mathbf{K}^{-1}\begin{bmatrix}\mathbf{y}_{l_2}^\intercal & 1\end{bmatrix}^\intercal \tag{21}$$

using two arbitrary 2D image points $\mathbf{y}_{l_1}$ and $\mathbf{y}_{l_2}$ located on the known projected centerline l_0 in the image, f is the focal length. Recovery of rotation matrix $\mathbf{R}$ is based on the observation of corresponding vectors pairs under the two coordinate frames. The solution to this problem can be referred to [38] which uses least-square method to compute the $\mathbf{R}$ (without iterations) in angle-axis representation, by collecting down to only two pairs of ${}^c\mathbf{n}_\pi$ and ${}^b\mathbf{n}_\pi$. Knowing two settled IFPs (π_1 and π_2) with different projected centerlines in the image suffices the estimation of $\mathbf{R}$, which is viable upon two initial configurations ($\mathbf{q}_{s_{0_1}}$ and $\mathbf{q}_{s_{0_2}}$) to obtain θ_{d_1} and θ_{d_2} using adaptive regulation, respectively.

The last step of our algorithm is to recover the position term $\mathbf{t}$. Although it is able to recover the transformation $\mathbf{T}$ via three vector pairs in [39], we seek to use online visual feedback of the instrument tip to decouple position estimation from orientation based on a known rotation matrix $\mathbf{R}$. Consider two robot configurations $\mathbf{q}_1 = [\mathbf{q}_s^\intercal\ q_{3_1}\ q_4\ \dots\ q_6]^\intercal$ and $\mathbf{q}_2 = [\mathbf{q}_s^\intercal\ q_{3_2}\ q_4\ \dots\ q_6]^\intercal$ which lead to two positions $\mathbf{x}_1(\mathbf{q}_1)$ and $\mathbf{x}_2(\mathbf{q}_2)$. Their in-between distance is then known and acts as a "virtual marker" in 3D Cartesian space whose 2D projected points are denoted by $\mathbf{y}_1$ and $\mathbf{y}_2$. Then, one can fully describe the 6-DoF pose via $\mathbf{R}$ and two scalar parameters $\hat{z}_1$ and $\hat{z}_2$ (instead of three due to the constraint from image feedback). The genuine pose of the instrument could be recovered by obtaining using online regulation of $\hat{z}_1$ and $\hat{z}_2$ upon image-based errors $\mathbf{y}_1$ and $\mathbf{y}_2$ (refer to our previous work [26] for detailed implementation). Knowing $\mathbf{x}_1(\mathbf{q}_1)$ and $\mathbf{x}_2(\mathbf{q}_2)$ from robot kinematics, the term $\mathbf{t}$ is thus determined as well. Such estimation process takes place in low-dimensional (3-DoF) space using image feedback, which could be efficient and accurate for vision-based instrument manipulation.

V. Results

A. Simulations

Simulations are conducted from which the ideal ground truth of hand-eye transformation could be retrieved to evaluate the performance of our algorithm. We use the Virtual Robot Experimentation Platform (V-REP) running with remote API interfacing the Matlab R2017a (MathWorks Inc) as our virtual robot-camera platform. The model of the dVRK introduced by [40] is adopted and further imported to the V-REP. To simulate the clinical set-up, the relative distance between the RCMs of the Endoscopic Camera Manipulator (ECM) and the Patient-Side

[2]The computed vectors might have two directions as the instrument might be inserted from the left or right side of the image. The ambiguity is solvable upon deliberate instrument motions to observe the movement of $\mathbf{m}$.

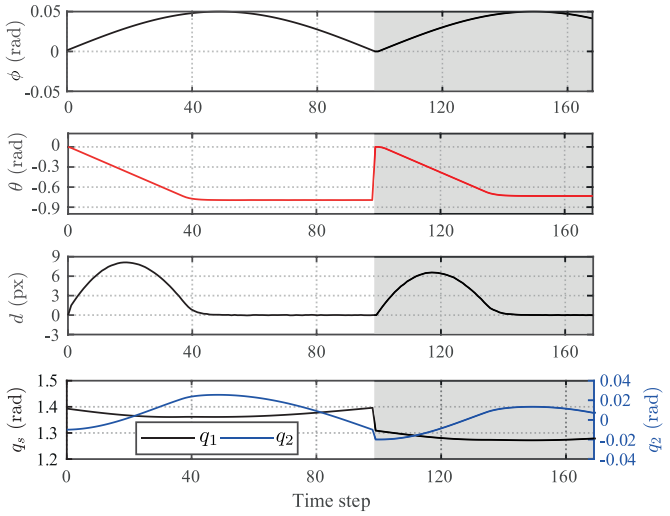

Fig. 2. The change of 2D distance d, spatial property of the IFP θ, and first 2-DoF joint motions of the robotic instrument over time. The time step $t \in \{0, 97\}$ and $t \in \{98, 172\}$ (shaded area) denote the process of regulating IFP upon the first/second initial robot configuration, respectively.

Manipulator (PSM) is set to ∼100 mm, which is the adopted distance for the two entry points in robot-assisted laparoscopy in clinical practice [41]. The depth of the instrument tip with respect to the optic center of the camera along its optic center is around 50–150 mm. We select the DeBakey Forceps (DBF) as the target instrument for our calibration whose kinematics data is precisely known.

A monocular virtual camera continuously observes the instrument with a pin-hole projection model. As the calculations in (6) and (14) require $\phi(t)$ to be a C^1 function. In the simulations and experiments, we define

$$\phi(t) = a\sin\omega t. \tag{22}$$

where $a = 0.05, \omega = \pi/100$ to generate a pendulum resembling trajectory with respect to its initial configuration for IM-based control.

We first show a single calibration process to demonstrate the performance of our approach. The threshold ϵ is set to 0.02, with $\hat{\mathbf{t}}_0 = [0\ 0\ 0]^\intercal$, $\theta_0 = 0$ and $|\dot{\theta}|$ being saturated to 0.002. The camera viewing angle is set to 60° (as in RMIS). Fig. 2 illustrates the evolution of IFP upon IM-induced trajectory regulation. As two IFPs are collected from two different initial robot configurations, there exists two times of regulation process. In each process, the parameter θ converges subject to the visual feedback d minimized to zero. Once the first IFP is considered determined via (19), the robot moves to a second robot configuration $\mathbf{q}_{s_{0_2}}$ (with random deviation) for the another process, which explains the sudden change of $\mathbf{q}_{s_{0_2}}$. The evolution of $\mathbf{q}_s$ indicates small joint motion ranges with $\Delta q_1 = 0.03$ rad, $\Delta q_2 = 0.17$ rad, which require a minimal workspace. The value d_0 starts from zero since $\phi_{t_0} = 0$ during each estimation step. Fig. 3 also demonstrates the regulation process of IFP upon IM, which is depicted as the projection of the instrument tip on $\mathcal{F}_{2_0}$ over time. The set-up is identical to it in Fig. 2, while the IFP's spatial property is adjusted from different $|\Delta\theta| = |\theta_0 - \theta_d|$ in

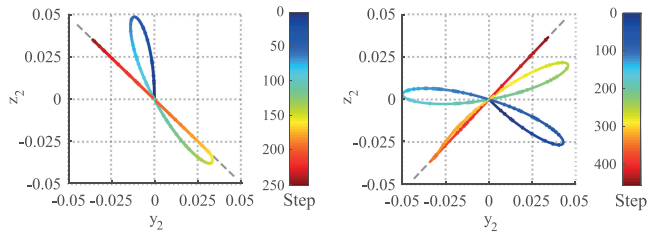

Fig. 3. Demonstration of the instrument tip motions relative to frame $\mathcal{F}_2$ subject to the varying θ from vision-based adjustment. The converged line (in dash gray line) indicates the projected settled IFP on plane $y_{2_0}O_{2_0}z_{2_0}$. The color indicates the corresponding time step under different tip positions.

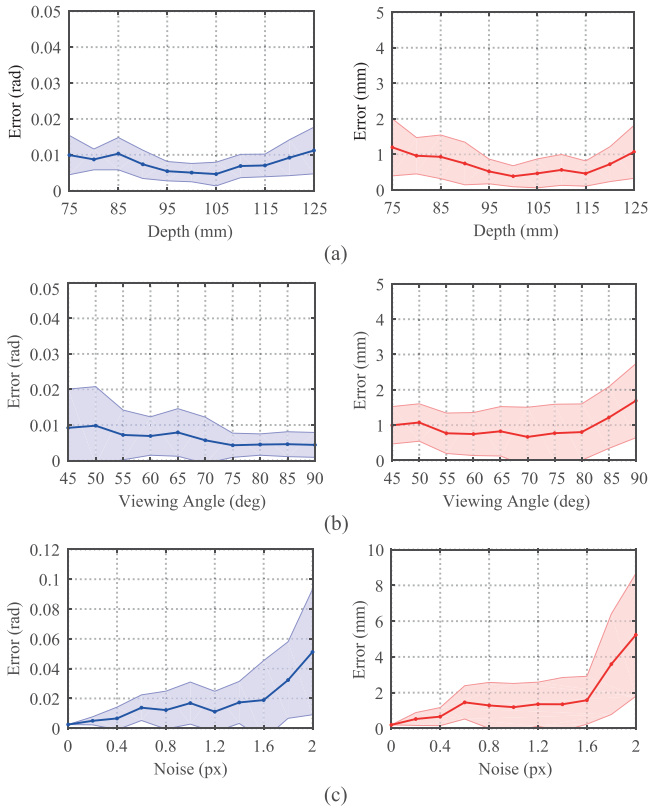

Fig. 4. Rotation error (left) and position error (right) of the hand-eye calibration results by using different visual depths toward the instrument (a), different viewing angles of the robot (b), and different noises applied (c).

Fig. 3 where $|\Delta\theta| = 0.8$ in Fig. 3(a) and $|\Delta\theta| = 1.6$ in Fig. 3(b). Convergence of θ is reached in both cases.

As there exists theoretical error in our modelling, in the simulation, we have also conducted comparative analysis to investigate the calibration accuracy upon different set-ups. We first adopt the previous set-up by fixing the viewing angle of the camera to 60° and adjusts the camera position along its depth with distance from the instrument body set between 75 mm–125 mm. Then, we fix the camera distance from the instrument to 100 mm and changes the camera viewing angle between 45°–90°. Ten trials are tested for each identical set-up. Fig. 4(a) illustrates the rotation error (unit in rad) and position error (unit in mm) of the calibration result upon changing the depth. Smaller depth tends to increase the errors since the magnified pixel-wise

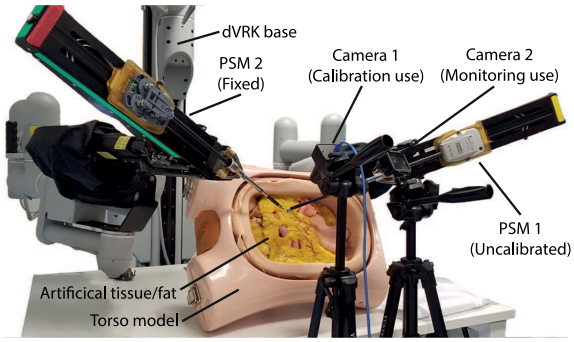

Fig. 5. The experimental set-up using the dVRK.

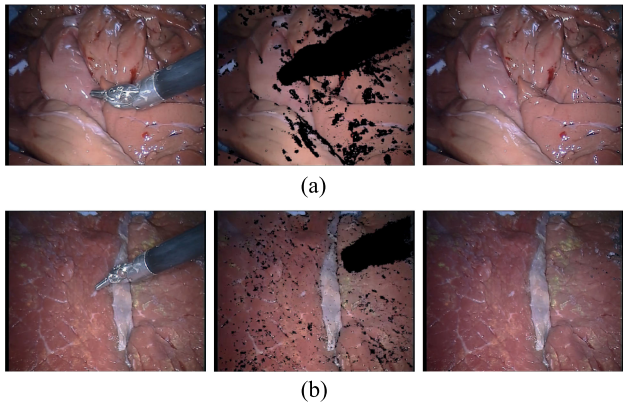

Fig. 6. Background extraction results for tool detection using two raw videos from the *EndoVis'15* dataset. The left images are the snapshots of raw image inputs where the instrument partially occludes the background. The middle images show the intermediate process (black pixels as undetermined at the moment). The right ones are the generated background images.

difference between the projected instrument shaft centerline and the one of the 2D instrument region in the image. The error distributions under different positions fluctuate due to the random selection of $\mathbf{q}_{s_0}$. In Fig. 4(b), the errors are not consistent as the instrument centerline could be detected more accurately as the projected region becomes more "slender" compared to close-up observations. Among most situations, our algorithm has reached $<$0.01 rotation error and $<$1 mm position error.

To evaluate calibration robustness, we generate detection noise to the visual data to study the leading results. The noise is uniformly generated during the data collection process with its magnitude ranging between 0–2 pixels, which affects the tracking of both 2D centerline position and the instrument tip position in the image. Fig. 4(c) shows the performance from which ten trials are considered for each level of noise. The errors arise with lager noise applied, as our method depends on the online visual feedback to regulate e_d, $\mathbf{e}_1$ and $\mathbf{e}_2$. However, an error of $<$0.015 rad and $<$1.2 mm can be reached given the noise $<$1.2 pixels.

B. Experiments

We use the dVRK with two robot manipulators, namely PSM 1 and PSM 2, to conduct our experimental study as the robot platform. The cisst/SAW libraries and dVRK ROS MATLAB wrapper are used to communicate between the upper-level controller (on an average desktop PC with Intel i7 CPU + 8 GB RAM) and the robot actuation. The PSM 1 is to be calibrated using our algorithm which is equipped with a DBF whose forward kinematics is known. A ProGrasp Forceps (PGF) is mounted on PSM 2 for needle handover. An industrial monocular camera (namely Camera 1) is connected via USB to the upper-level PC to capture visual feedback with 640 $\times$ 480 pixels resolution at 30 fps. The camera is pre-calibrated using Zhang's method [42] with a 0.12 pixel of mean back-projection error from 20 single images capturing a 4-mm 7 $\times$ 6 calibration grid. A human torso model is placed behind the instrument as a static intra-corporeal background appeared in RMIS. The complete set-up is shown in Fig. 5.

An image processing algorithm is developed to detect the instrument from the image. Based on the IM-induced motions, the instrument change its projected region in the image, which partially occlude different parts of the background upon movement. We vote for each pixel over time whether its RGB pixel intensities has changed dramatically within last certain frames via a pre-set threshold. Those pixels are permanently categorized as "background," until the proportion of such pixels in the image has surpassed a predefined threshold (set to 99.95% in our case). We demonstrate the algorithm feasibility by applying to video sequences from *EndoVis'15* MICCAI Challenge dataset[3] as shown in Fig. 6, from which a realistic background can be generated to ease the instrument detection phase based on background subtraction. Note that developing an advanced image processing algorithm is not the primary concern of this work, while the problem could also be solvable by motion-based or learning-based segmentation approaches.

To evaluate the performance of our algorithm, we compare with the Tsai's method [9] by analyzing the deviation of the instrument tip's 2D back-projection on the image plane from its visually-appeared position. The lateral and longitudinal calibration error using Tsai's method (upon a 20-group data collection with a 4-mm 7 $\times$ 6 calibration grid) is $<$0.05 mm and $<$0.12 mm, respectively. Note that this is not the ground truth as the tendon-driven design of the robot joints could lead to much lower accuracy in task positioning as in [23], and is only used for comparative analysis. We set the threshold ϵ of $|\theta(t)|$ to be 0.03 with $T = 20$, and tuning gain $\kappa = 0.05$. The initial elements in $\hat{\mathbf{a}}_0$ and the gain elements Γ are set to 1 and all 0.002, respectively. Fig. 7 demonstrates the dual-stage calibration process of l towards l_0 as well as $\hat{\mathbf{y}}_1$ towards $\mathbf{y}_1$, both using adaptive regulation. To analyze the calibration performance, we initiate two different robot configurations C-I and C-II, from which the robot follows a pre-defined 3D circular trajectory in C-I and a manually manipulated one in C-II. Fig. 8 illustrates the back-projection snapshots of the joint positions with solely the calibrated kinematic data using the two methods. Three example 3D trajectories and the back-projection errors are shown in Fig. 9. Note that the shapes of the reconstructed 3D trajectories using two methods resemble each other which implies similar calibrated results of rotation matrix, apart from the $>$10 mm

[3][Online]. Available: Source: https://github.com/surgical-vision/EndoVis PoseAnnotation

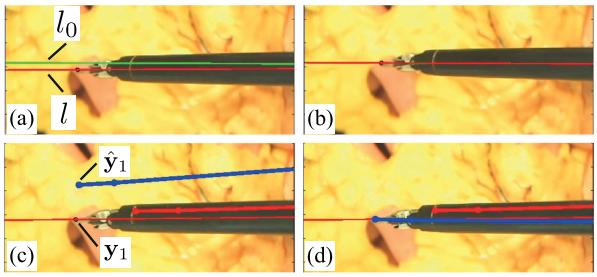

Fig. 7. Demonstration of the calibration process. (a) & (b): Orientation estimation process: Regulating l (green) to be aligned to l_0 (red). (c) & (d) Position estimation process: Regulating $\hat{\mathbf{y}}_1$ to $\mathbf{y}_1$ (instrument back-projections are shown using our method in blue and Tsai's method in red).

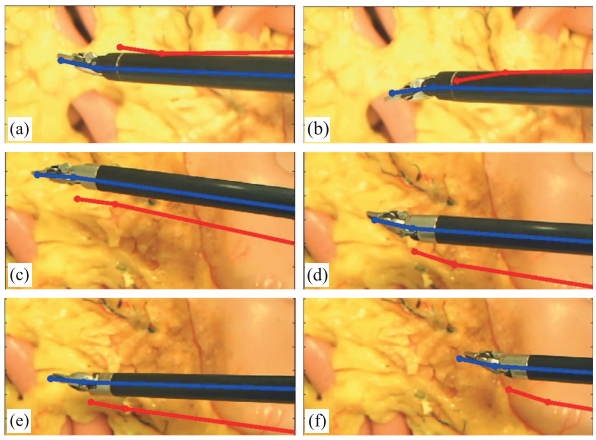

Fig. 8. Back-projection of kinematic data on the image computed from eye-to-hand calibration results (red as the Tsai's method and blue as ours). Note that results from (a)/(b) and (c)/(d)/(e)/(f) are with configuration C-I and C-II, respectively.

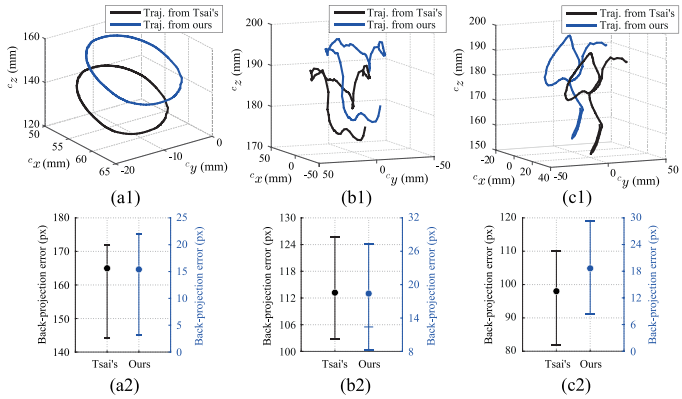

Fig. 9. Three example 3D instrument tip trajectories computed from solely the kinematic data and their projected 2D errors, with (a1)/(a2) using a pre-defined trajectory, (b1)/(b2) and (c1)/(c2) the (random) manual trajectories.

translation difference. Meanwhile, our method exhibits better calibration accuracy compared to it using Tsai's method, as smaller 2D back-projection errors (from >100 px to ∼20 px) are recorded over time in all three 3D trajectories. The ranges of error distribution remain similar which might result from the intrinsic robotic positioning inaccuracy. This indicates the importance of online visual adjustment in our method to reduce robot tracking errors within local movements.

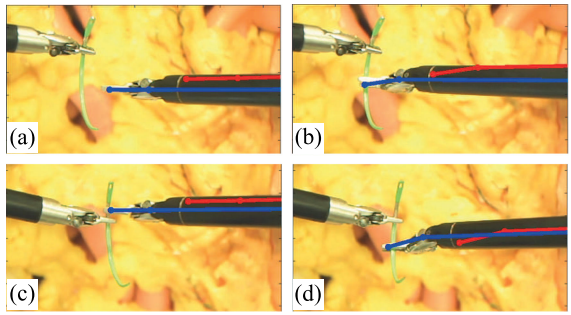

Fig. 10. Dual-arm needle handover with a fixed left arm and a calibrated right arm using our method (Case 1: (a)/(b); Case 2: (c)/(d)). The instrument back-projection from our approach (in blue) is accurate to complete needle re-grasping while using Tsai's method (in red) tends to fail.

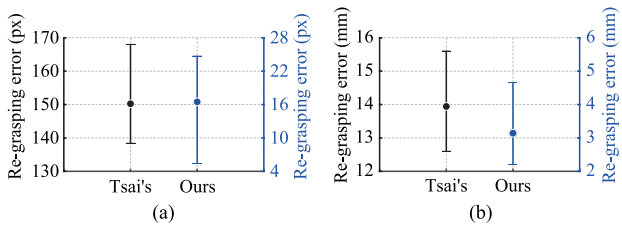

Fig. 11. (a) The measured 3D positioning errors between the instrument tip and the pre-defined re-grasping point. (b) The back-projected errors of the instrument tip in the 2D image plane.

C. Case Study: Dual-Arm Needle Handover

We finally conduct case study to simulate suturing needle handover in RMIS to further evaluate the calibration accuracy for task-relevant instrument positioning. To manually define a 3D re-grasping point on the needle, we add a second camera for stereo vision with a ∼120 mm baseline (mean back-projection error of 0.24 pixel upon 20 pairs of collected images with the same checkerboard. The 3D distance between the center of the instrument's open jaws and the re-grasping point is manually measured as the positioning error. We show two re-grasping cases using different robot configurations and needle poses, with each running ten trials of independent calibration. Fig. 10 illustrates the dual-arm needle re-grasping process, and Fig. 11 shows the calculated 2D errors in (a) and measured 3D errors in (b) of the instrument tip, with the mean 13.9/3.2 mm 3D positioning error and 150.1/16.4 px back-projection error within ten trials using Tsai's/our method, respectively. In both processes, the back-projected tip errors upon our method do not exceed 20 px. The instrument positioning is accurate enough for needle re-grasping due to the use of online visual feedback.

VI. Conclusion

A novel autonomous hand-eye calibration method for robotic instrument using a fixed monocular camera is proposed in this paper. The method directly leverages the instrument's online rigid-body motion behavior via IM to reveal additional sensory information instead of applying external calibration objects or the exact CAD model. By proposing the IFP, two groups of data suffice the calibration process within limited workspace.

The accuracy is competitive to the state-of-the-art results with $\sim$3 mm 3D instrument positioning error in the case study and $\sim$20 px back-projection error. While lens distortions of a medical endoscope might affect the calibration accuracy, the raw image could be rectified via pre-calibrating the distortion parameters such that the pinhole camera model still reasonably holds. It could be potentially used for intra-corporeal hand-eye calibration in small workspace after pre-operative set-up, or for robots with long kinematic chain to deal with inaccurate models.

This work so far focuses on feasibility study. In the future, we will use a stereo medical endoscope under a more realistic scenario to test the performance of our algorithm. We also seek to evaluate its performance in complex workspace with robot-to-robot calibration for dual-arm autonomous surgery.

REFERENCES

[1] M. Yip and N. Das, "Robot autonomy for surgery," p. 1, 2017, *arXiv:1707.03080*.

[2] K. A. Nichols and A. M. Okamura, "Methods to segment hard inclusions in soft tissue during autonomous robotic palpation," *IEEE Trans. Robot.*, vol. 31, no. 2, pp. 344–354, Apr. 2015.

[3] S. Voros, J.-A. Long, and P. Cinquin, "Automatic localization of laparoscopic instruments for the visual servoing of an endoscopic camera holder," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention.*, 2006, pp. 535–542.

[4] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Sci. Trans. Med.*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[5] S. Sen, A. Garg, D. V. Gealy, S. McKinley, Y. Jen, and K. Goldberg, "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2016, pp. 4178–4185.

[6] F. Zhong, Y. Wang, Z. Wang, and Y.-H. Liu, "Dual-arm robotic needle insertion with active tissue deformation for autonomous suturing," *IEEE Rob. Autom. Lett.*, vol. 4, no. 3, pp. 2669–2676, Jul. 2019.

[7] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: A review of the literature," *Med. Image Anal.*, vol. 35, pp. 633–654, 2017.

[8] Z. Zhang, L. Zhang, and G.-Z. Yang, "A computationally efficient method for hand–eye calibration," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 10, pp. 1775–1787, 2017.

[9] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Trans. Robot. Autom.*, vol. 5, no. 3, pp. 345–358, Jun. 1989.

[10] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robot. Res.*, vol. 18, no. 3, pp. 286–298, 1999.

[11] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer Handbook of Robotics*. Berlin, Germany: Springer, 2016, pp. 1657–1684.

[12] C. Doignon and M. de Mathelin, "A degenerate conic-based method for a direct fitting and 3-D pose of cylinders with a single perspective view," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 4220–4225.

[13] C. Doignon, F. Nageotte, B. Maurin, and A. Krupa, "Model-based 3-D pose estimation and feature tracking for robot assisted surgery with medical imaging," in *Proc. Worshop, IEEE Int. Conf. Robot. Autom.*, 2007, pp. 1–10.

[14] R. Wolf, J. Duchateau, P. Cinquin, and S. Voros, "3D tracking of laparoscopic instruments using statistical and geometric modeling," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2011, pp. 203–210.

[15] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1050–1058, Apr. 2013.

[16] R. Hao, O. Özgüner, and M. C. Çavuşoğlu, "Vision-based surgical tool pose estimation for the da vinci robotic surgical system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2018, pp. 1298–1305.

[17] M. Allan, S. Ourselin, D. J. Hawkes, J. D. Kelly, and D. Stoyanov, "3-D pose estimation of articulated instruments in robotic minimally invasive surgery," *IEEE Trans. Med. Imaging*, vol. 37, no. 5, pp. 1204–1213, May 2018.

[18] N. Rieke *et al.*, "Surgical tool tracking and pose estimation in retinal microsurgery," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 266–273.

[19] T. Kurmann *et al.*, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2017, pp. 505–513.

[20] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1542–1549, Jul. 2017.

[21] X. Du *et al.*, "Articulated multi-instrument 2-d pose estimation using fully convolutional networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1276–1287, May 2018.

[22] I. Laina *et al.*, "Concurrent segmentation and localization for tracking of surgical instruments," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2017, pp. 664–672.

[23] A. Reiter, P. K. Allen, and T. Zhao, "Appearance learning for 3d tracking of robotic surgical tools," *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 342–356, 2014.

[24] M. Allan *et al.*, "Image based surgical instrument pose estimation with multi-class labelling and optical flow," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 331–338.

[25] M. Ye, L. Zhang, S. Giannarou, and G.-Z. Yang, "Real-time 3d tracking of articulated tools for robotic surgery," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2016, pp. 386–394.

[26] D. Navarro-Alarcon *et al.*, "Robust image-based computation of the 3d position of rcm instruments and its application to image-guided manipulation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 4115–4121.

[27] F. Mourgues *et al.*, "Flexible calibration of actuated stereoscopic endoscope for overlay in robot assisted surgery," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2002, pp. 25–34.

[28] J. Schmidt, F. Vogt, and H. Niemann, "Robust hand–eye calibration of an endoscopic surgery robot using dual quaternions," in *Proc. Joint Pattern Recognit. Symp.*, 2003, pp. 548–556.

[29] A. Malti and J. P. Barreto, "Robust hand-eye calibration for computer aided medical endoscopy," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2010, pp. 5543–5549.

[30] K. Pachtrachai, F. Vasconcelos, G. Dwyer, S. Hailes, and D. Stoyanov, "Hand-eye calibration with a remote centre of motion," *IEEE Rob. Autom. Lett.*, vol. 4, no. 4, pp. 3121–3128, Oct. 2019.

[31] K. Pachtrachai *et al.*, "Adjoint transformation algorithm for hand–eye calibration with applications in robotic assisted surgery," *Ann. Biomed. Eng.*, vol. 46, no. 10, pp. 1606–1620, 2018.

[32] K. Pachtrachai, M. Allan, V. Pawar, S. Hailes, and D. Stoyanov, "Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2016, pp. 2485–2491.

[33] Z. Wang *et al.*, "Vision-based calibration of dual rcm-based robot arms in human-robot collaborative minimally invasive surgery," *IEEE Rob. Autom. Lett.*, vol. 3, no. 2, pp. 672–679, Apr. 2017.

[34] J. Bohg *et al.*, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Trans. Robot.*, vol. 33, no. 6, pp. 1273–1291, Dec. 2017.

[35] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz, "Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 5003–5010.

[36] D. Katz and O. Brock, "Manipulating articulated objects with interactive perception," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2008, pp. 272–277.

[37] J.-J. E. Slotine *et al.*, *Applied Nonlinear Control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[38] K. Halvorsen, M. Lesser, and A. Lundberg, "A new method for estimating the axis of rotation and the center of rotation," *J. Biomechanics*, vol. 32, no. 11, pp. 1221–1227, 1999.

[39] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 5, pp. 698–700, Sep. 1987.

[40] G. A. Fontanelli, M. Selvaggio, M. Ferro, F. Ficuciello, M. Vendittelli, and B. Siciliano, "A v-rep simulator for the da vinci research kit robotic platform," in *Proc. Prof. IEEE Int. Conf. Biomed. Rob. Biomechatronics.*, 2018, pp. 1056–1061.

[41] P. Escobar and T. Falcone, *Atlas of Single-Port, Laparoscopic, and Robotic Surgery*. Berlin, Germany: Springer, 2014.

[42] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.