Towards Autonomous Control of Surgical Instruments using Adaptive-Fusion Tracking and Robot Self-Calibration

Chiyu Wang*,[1], João Cartucho*,[1], Daniel Elson[1], Ara Darzi[1] and Stamatia Giannarou[1]

Abstract— **The ability to track surgical instruments in real-time is crucial for autonomous Robotic Assisted Surgery (RAS). Recently, the fusion of visual and kinematic data has been proposed to track surgical instruments. However, these methods assume that both sensors are equally reliable, and cannot successfully handle cases where there are significant perturbations in one of the sensors' data. In this paper, we address this problem by proposing an enhanced fusion-based method. The main advantage of our method is that it can adjust fusion weights to adapt to sensor perturbations and failures. Another problem is that before performing an autonomous task, these robots have to be repetitively recalibrated by a human for each new patient to estimate the transformations between the different robotic arms. To address this problem, we propose a self-calibration algorithm that empowers the robot to autonomously calibrate the transformations by itself in the beginning of the surgery. We applied our fusion and self-calibration algorithms for autonomous ultrasound tissue scanning and we showed that the robot achieved stable ultrasound imaging when using our method. Our performance evaluation shows that our proposed method outperforms the state-of-art both in normal and challenging situations.**

I. INTRODUCTION

Recent advances in Robotic Assisted Surgery (RAS) have enabled autonomous execution of surgical tasks [1], [2]. A typical autonomous task is tissue scanning, which is based on Ultrasound (US) [3], [4], Optical Coherence Tomography (OCT) [5] or Electric Bio-Impedance (EBI) [6]. Autonomous tissue scanning, reduces surgical workload, since the robot does the imaging characterization of the tissue autonomously. Other autonomous tasks include suturing, debridement, and others. For all these autonomous tasks it is crucial that the robot estimates accurately the surgical instrument's pose. There are two main approaches for estimating instrument pose during surgery, either by using vision only, or by fusing vision and kinematics. Vision-based methods mainly fall into two categories depending on whether they use natural-features [7], extracted from the instruments directly, or artificial-features, extracted from fiducial markers [8], [9]. As a supplement of vision, kinematic information is also available in surgical robots, and can be used along with visual data. These fusion methods typically use recursive filters, such as Kalman Filter and its extensions [3], [8], [10].

The above fusion methods assume that both visual and kinematic sensors are always equally reliable in tracking the instruments. Hence, equal and constant weights of both sensors are used to fuse visual and kinematic data. In practice, the accuracy of visual information deteriorates when the instruments move fast, causing motion blur, or when the instrument is far from the camera and visual details (features) on the instruments cannot be detected. On the other hand, inaccuracies in the calibrations and noise in the robot joint readings introduce errors to the kinematic data. However, the existing methods do not have a mechanism to deal with above challenging conditions such as noise in the data, poor illumination, and motion blur. Another limitation of existing EKF-based fusion methods is that they assume that the EKF noise covariance matrices are pre-defined and fixed. However, the noise covariance matrices are crucial for the good performance of the EKF, and should be updated at each step to avoid divergence of the filter [11].

When fusing visual and kinematic data, the visual input is calculated in the camera coordinate frame while the kinematic input is in the robot's base coordinate frame. Therefore, calibrating the transformation from the robot's base to the camera's coordinate frame, ${}^{C}T_{B}$, is crucial. This transformation is not fixed and needs to be estimated every time the surgeon adjusts the robot to a new patient. The problem is that for a surgical robot to become truly autonomous it would also need to self-calibrate ${}^{C}T_{B}$. This would enable the robot to autonomously operate, without requiring human assistance during the calibration process.

In this paper, the above challenges are addressed. First, an enhanced visual and kinematics fusion method, based on the Extended Kalman Filter (EKF), is proposed. This method advances state-of-the-art by using fuzzy logic to attribute weights to the vision and to the kinematic input separately. The sensor measurement closer to the EKF prediction, is more likely to be the most accurate one. Hence, when the visual input is more accurate than the kinematic input, a higher weight is attributed to the vision and vice versa. These weights are updated automatically during tracking. Therefore, even under challenging situations, such as when the instrument is occluded, our fusion method adapts automatically the weights which are used to accurately estimate the instrument's pose. In addition, to boost the performance of the EKF, our noise covariance matrices are adaptively tuned during tracking. Finally, to enable full autonomy of the robot, we propose a self-calibration routine for a surgical robot to perform at the beginning of the surgery. Here, we apply our fusion and self-calibration methods for autonomous tissue scanning.The contributions are:

1) A novel method for adaptive fusion of visual and kinematic information which dynamically assigns higher

*These authors contributed equally to the work.

Chiyu Wang: c.wang20@imperial.ac.uk
João Cartucho: j.cartucho19@imperial.ac.uk
[1]The Hamlyn Centre for Robotic Surgery, Imperial College London, London SW7 2AZ, UK

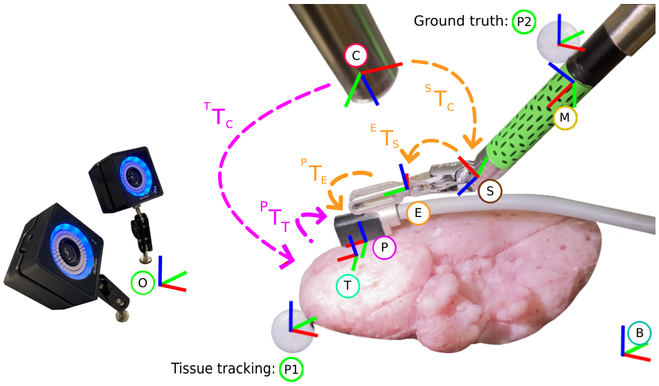

Fig. 1: Coordinate frames: left endoscopic camera (C), ultrasound probe (P), target scanning point (T), cylindrical marker (M), and robot's base (B), shaft (S), end effector (E).

weight to the more reliable sensor.

2) For the first time, self-calibration is introduced in robotic surgery to estimate ${}^{C}T_{B}$ automatically.

II. Methodology

A. Framework overview

Our proposed framework, illustrated in Fig.1, has been applied to autonomous scanning of tissue, using an endoscopic Ultrasound (US) probe controlled by the da Vinci® robotic platform. The tissue scanning framework can be mathematically described by the minimization of the difference between the desired probe pose w.r.t camera (${}^{P^*}T_C = {}^{P^*}T_{T^*} \times {}^{T^*}T_C$) and the current probe pose w.r.t camera (${}^{P}T_C = {}^{P}T_E \times {}^{E}T_S \times {}^{S}T_C$), which is calculated from vision and kinematics measurements. ${}^{P}T_E$ is a rigid transformation that is estimated prior to the tissue scanning as explained in Sec.III-B. ${}^{E}T_S$ is the transformation between shaft and end effector of Patient Side Manipulator (PSM), which can be calculated by joint values and forward kinematics. ${}^{S}T_C$ is the pose of the shaft w.r.t the camera.

Our main goal is to control the robot to be able to follow in real-time a target scanning point located on the tissue surface with pose w.r.t the camera denoted as ${}^{T^*}T_C$. ${}^{P^*}T_{T^*}$ is the desired probe pose w.r.t tissue target point. To achieve this scanning goal, ${}^{P^*}T_{T^*}$ should be kept equal to the identity matrix during the task. To minimize the difference between ${}^{P}T_C$ and ${}^{P^*}T_C$, ${}^{E^*}T_B$ is used to control the PSM to the desired end effector pose. ${}^{E^*}T_B$ can be calculated as:

$$\begin{aligned} {}^{E^*}T_B &= {}^{E^*}T_{S^*} \times {}^{S^*}T_S \times {}^{S}T_B \\ &= {}^{P^*}T_{E^*}^{-1} \times {}^{P^*}T_{T^*} \times {}^{T^*}T_C \times {}^{S}T_C^{-1} \times {}^{S}T_B \end{aligned} \quad (1)$$

In our framework, ${}^{S}T_C$ can be accurately estimated by fusing vision (${}^{S}T_C^{v} = {}^{S}T_M \times {}^{M}T_C$) and kinematics (${}^{S}T_C^{k} = {}^{S}T_B \times ({}^{C}T_B)^{-1}$) information. ${}^{S}T_M$ is the marker to shaft transformation, which is obtained from the calibration before any surgery, as explained in Sec.II-C. ${}^{M}T_C$ is the marker pose w.r.t the camera which is estimated using a state-of-the-art cylindrical marker which encodes a series of binary codes [9]. ${}^{C}T_B$, is camera pose w.r.t PSM base, which is automatically estimated using our self-calibration as explained in Sec.II-D.

B. Adaptive fusion-based surgical instrument tracking

The Extended Kalman Filter (EKF) is used to fuse ${}^{S}T_C^{v}$ and ${}^{S}T_C^{k}$ defined in Sec. II-A, to accurately estimate the pose of the surgical instrument.

1) EKF: The pose of an instrument is composed of two parts: the location (x, y, z) and the orientation in quaternions (q_w, q_x, q_y, q_z). Linear velocities (v_x, v_y, v_z) and angular velocities $(\omega_x, \omega_y, \omega_z)$ captured from the dVRK are used to establish a non-linear fusion model. Our model is based on the assumption of constant linear and angular velocities. A Time-Discrete EKF has been designed where the state prediction process is defined as $x_{k|k-1} = f(x_{k-1|k-1})$. The prediction function $f(x)$ is defined as:

$$\begin{pmatrix} l(k) \\ q(k) \\ q_w(k) \\ v(k) \\ \omega(k) \end{pmatrix} = f(x_{k-1}) = \begin{pmatrix} l(k-1) + \Delta t \times v(k-1) \\ q(k-1) \times (1 + \frac{\Delta t}{2} \times \omega(k-1)) \\ q_w(k-1) \\ v(k-1) \\ \omega(k-1) \end{pmatrix} \quad (2)$$

where, $l = [x, y, z]^T$, $q = [q_x, q_y, q_z]^T$, $v = [v_x, v_y, v_z]^T$, $\omega = [\omega_x, \omega_y, \omega_z]^T$. Δt is the time difference between consecutive timestamps. $q(k)$ and $q_w(k)$ are updated using the 1st term of Taylor series approximation, to reduce computation time. The prediction covariance matrix, $P_{k|k-1}$, can be estimated by using the state transition matrix, F_k, and the process noise covariance matrix, Q_k, as $P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k$ where, Q_k represents the reliability of the prediction, as explained in section II-B.3. F_k, is the Jacobian matrix of $f(x)$:

$$F_k = \frac{\partial f}{\partial x} = \begin{pmatrix} F_{k11} & F_{k12} \\ F_{k21} & F_{k22} \end{pmatrix} \quad (3)$$

where, $F_{k11} = diag(1, 1, 1, 1 + \frac{\Delta t}{2}\omega_x, 1 + \frac{\Delta t}{2}\omega_y, 1 + \frac{\Delta t}{2}\omega_z, 1)$, $F_{k22} = diag(1, 1, 1, 1, 1, 1)$, F_{k12} is a 7×6 matrix defined as $[diag(\Delta t, \Delta t, \Delta t, \frac{\Delta t}{2}q_x, \frac{\Delta t}{2}q_y, \frac{\Delta t}{2}q_z); [0, 0, 0, 0, 0, 0]]$, F_{k21} is a 6×7 zero matrix.

The synchronized visual and kinematic data uses the same prediction process, but different update processes. Hence, different noise covariance matrices are used to update them separately. For either vision or kinematics, the update process is defined as:

$$r_k = z_k - h(x_{k|k-1}) \quad (4)$$

$$S_k = H_k P_{k|k-1} H_k^T + R_k \quad (5)$$

$$K_k = P_{k|k-1} H_k^T (S_k)^{-1} \quad (6)$$

$$x_{k|k} = x_{k|k-1} + K_k \times r_k \quad (7)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \quad (8)$$

where, z_k is the measurement. The measurement function is defined as $h(x) = x$. H_k is the Jacobian matrix of $h(x)$, which is the identity matrix in our EKF. R_k is the measurement noise covariance matrix. r_k is the residual between the pose estimation and measurement. S_k is the residual covariance

TABLE I: Membership function of weight

	Residual of k/v	Relative weight of k/v
Fuzzy set	Membership function	Membership function
Z	(0;0;0.325)	(0;0;0.125)
S	(0.25;0.35;0.45)	(0.025;0.175;0.325))
M	(0.375;0.5;0.625)	(0.25;0.5;0.75)
L	(0.55;0.625;0.75)	(0.625;0.775;0.925)
VL	(0.675;0.75;0.75)	(0.875;0.925;0.925)

TABLE II: Fuzzy rules of weight for vision & (kinematics)

r_v r_k	Z	S	M	L	VL
Z	M(M)	M(M)	L(S)	L(S)	VL(Z)
S	M(M)	M(M)	M(M)	L(S)	L(S)
M	S(L)	M(M)	M(M)	M(M)	L(S)
L	S(L)	S(L)	M(M)	M(M)	M(M)
VL	Z(VL)	S(L)	S(L)	M(M)	M(M)

TABLE III: Membership function for the covariance matrix

DoM	Membership	adjQ/adjR	Membership
Fuzzy set	function	Fuzzy set	function
(S	(0;0;0.75)	D)	(0;0;0.9)
(E	(0.5;0.75;5)	M)	(0.75;0.9;1.25;1.5))
(L	(2.5;10;10)	I)	(1.25;2.0;2.0)

matrix and K_k is the Kalman gain. $x_{k|k}$ and $P_{k|k}$ are the updated state and covariance matrices, respectively. Separate EKFs are used for vision and kinematics resulting in the $x_{k|k,v}$ and $x_{k|k,k}$ updated states, respectively.

2) Adaptive weight: After getting the updated states from vision-based EKF ($x_{k|k,v}$) and kinematics-based EKF ($x_{k|k,k}$), the weights $weight_v$ and $weight_k$ are used to fuse them as:

$$x_{k|k,fusion} = weight_k \times x_{k|k,k} + weight_v \times x_{k|k,v} \tag{9}$$

For efficient data fusion, an adaptive weight scheme has been designed in our work to assign higher weight to the sensor which provides more accurate instrument pose estimation. To determine which sensor is more accurate, a fuzzy logic algorithm has been adopted which takes the residual information (r_k, r_v) estimated in Eq. (4) as input and outputs relative weights (w_k, w_v). The fuzzy logic algorithm uses the membership functions in Table I to calculate the membership of a residual to one of the five fuzzy sets namely, Zero (Z), Small positive (S), Middle positive (M), Large positive (L) and Very Large positive (VL) [12]. The fuzzy rules in Table II are used to estimate the membership of a relative weight given the membership of a residual pair. The Centroid Defuzzification method [13] is used to calculate the relative weights using the above membership. When there is no vision input, the relative weight will be set as $w_k = 1, w_v = 0$. The relative weights are normalized to absolute weights as:

$$[weight_k, weight_v] = [\frac{w_k}{w_k + w_v}, \frac{w_v}{w_k + w_v}] \tag{10}$$

3) Adaptive noise covariance matrix: The measurement noise matrix R is hard to define since it depends on the sensor and the measurement process model. A solution for defining R is to use the difference between the theoretical residual covariance matrix S_k in Eq. (5) and its actual value $\hat{C}_{rk}$ to constantly update R. This difference can be described by the Degree of Match (DoM) as [14]:

$$DoM = \frac{S_k}{\hat{C}_{rk}} = \frac{S_k}{\frac{1}{N}\sum_{i=k-N+1}^{k} r_i r_i^T} \tag{11}$$

where, in $\hat{C}_{rk}$ the residual r_i can be calculated from Eq. (4) and N is the number of timestamps that are used to calculate $\hat{C}_{rk}$. Ideally, S_k and $\hat{C}_{rk}$ are the same, and $DoM \approx 1$.

In this work, a fuzzy logic algorithm is used to estimate $adjR$ to tune R to satisfy the adaption rules below and the membership function in Table III.

1) If $DoM < 1$, then $S_k < \hat{C}_{rk}$, so R should increase: (S, D)
2) If $DoM = 1$, then $S_k = \hat{C}_{rk}$, so R is not changed: (E, M)
3) If $DoM > 1$, then $S_k > \hat{C}_{rk}$, so R should decrease, (L, I)

DoM has three fuzzy sets: Small (S), Equal (E) and Large (L), and $adjR$ has three fuzzy sets: Increase (I), Maintain (M) and Decrease (D). The current measurement covariance matrix R_k is updated as $R_{k+1} = adjR \times R_k$, so that the discrepancy between S_k and $\hat{C}_{rk}$ in the next timestamp can be mitigated.

The other noise covariance matrix that needs to be estimated is the process noise matrix Q which represents the uncertainty of our kinematics model. This uncertainty results from unknown linear (a) and angular acceleration (α). Similar to the estimation of $adjR$, Eq. (11) is used to estimate the DoM and subsequently $adjQ$. The S_k in Eq. (11) can be calculated as $S_k = H_k(F_k P_{k|k} F_k^T + Q_k)H_k^T + R_{k+1}$ where, R_{k+1} is the measurement covariance matrix of next timestamp defined above.

The update of matrix Q needs to be divided into two parts Q_{tr} and Q_{rt}, for the translation and rotation, respectively. For each translation state element (x, y, z) and its corresponding linear velocity, the translation covariance matrix Q_{tr} can be derived by:

$$\begin{aligned} d_{tr} &= \begin{bmatrix} \frac{1}{2}a\Delta t^2 \\ a\Delta t \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix} a \\ Q_{tr} &= E\left[d_{tr} d_{tr}^T\right] = \sigma_a \begin{bmatrix} \frac{1}{4}\Delta t^4 & \frac{1}{2}\Delta t^3 \\ \frac{1}{2}\Delta t^3 & \Delta t^2 \end{bmatrix} \end{aligned} \tag{12}$$

where, d_{tr} is the uncertainty of our kinematics model. a is linear acceleration. For each rotation state element (qx, qy, qz) and its corresponding angular velocity, the rotation covariance matrix Q_{tr} can be derived by:

$$\begin{aligned} d_{rt} &= \begin{bmatrix} \frac{1}{4}q\alpha\Delta t^2 \\ \alpha\Delta t \end{bmatrix} = \begin{bmatrix} \frac{1}{4}q\Delta t^2 \\ \Delta t \end{bmatrix} \alpha = \\ Q_{rt} &= E\left[d_{rt} d_{rt}^T\right] = \sigma_\alpha \begin{bmatrix} \frac{1}{16}q^2\Delta t^4 & \frac{1}{4}q\Delta t^3 \\ \frac{1}{4}q\Delta t^3 & \Delta t^2 \end{bmatrix} \end{aligned} \tag{13}$$

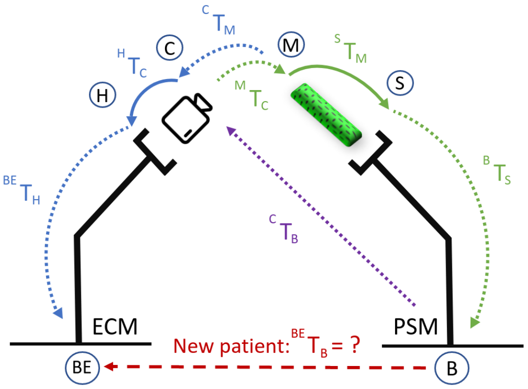

Fig. 2: Robotic arms. The fixed transformations are represented with solid lines, the ever-changing transformations with doted lines, and the transformation that can change but is maintained fixed during a surgery (${}^{BE}T_B$) is shown in a dashed red line.

where, d_{rt} is the uncertainty of our kinematics model. α is angular acceleration. q is orientation in quaternions format. σ_a and σ_α are scale factors which represent the uncertainty of the linear and angular acceleration. They are challenging to measure directly from the sensors. Hence, they are estimated as $\sigma(k) = adjQ \times \sigma(k-1)$.

C. Calibration before any surgery

As illustrated in Fig.2, both the transformation from the marker to the shaft (${}^S T_M$) and from the camera to its holder (${}^H T_C$) are fixed. These transformations will always remain fixed since the marker is rigidly attached to the surgical instrument and likewise, the camera is rigidly attached to the ECM camera holder. Therefore, ${}^S T_M$ and ${}^H T_C$ can be determined before any surgery, and are used during the autonomous calibration (Sec. II-D and II-E). Both these transformations are determined using a standard hand-eye calibration method $A_iX = XB_i$ [15]. When $X = {}^S T_M$, then $A_i = {}^B T_S$ and $B_i = {}^M T_C$, and the Endoscopic Camera Manipulator (ECM) remains static during this calibration. When $X = {}^H T_C$, then $A_i = {}^{BE} T_H$ and $B_i = {}^C T_M$, and the PSM remains static during this calibration.

D. Autonomous calibration during surgery

The main goal of the robot self-calibration is to determine the transformation from the PSM's base to the ECM's base (${}^{BE}T_B$), as illustrated in Fig.2. This transformation is changed for each new patient, since there is no available kinematic information in the set-up joints, which are the joints between B and BE. It is calculated as:

$$ {}^{BE}T_B = {}^{BE}T_H \times {}^H T_C \times {}^C T_B \tag{14}$$

where, ${}^{BE}T_H$ is accurately measured using ECM kinematics, ${}^H T_C$ is determined before surgery (Sec. II-C), and ${}^C T_B$ is determined automatically by the robot as follows:

1) First, the surgeon adjusts the set-up joints and inserts the surgical instruments inside the new patient. The surgeon is instructed to position the tip of the shaft in the centre of the endoscopic view, at least 5 $[cm]$ away from the tissue. This placement defines a safety volume for the surgical instrument to move inside during the self-calibration. To aid the surgeon's perception of distances, the safety volume around the shaft is shown to the surgeon in the endoscopic image. After step 1, the self-calibration is initiated;
2) The robot captures a new image and, using all the images captured so far, performs a robot-world hand-eye calibration [16] $A_iX = ZB_i$ (where $A_i = {}^B T_S$, $X = {}^S T^*{}_M$, $Z = ({}^C T_B)^{-1}$, and $B_i = {}^C T_M$), to estimate ${}^C T_B$;
3) The solution of ${}^C T_B$ is validated by comparing $X = {}^S T^*{}_M$ with the ground-truth ${}^S T_M$, which was determined before surgery (Sec. II-C). If ${}^S T^*{}_M \approx {}^S T_M$, i.e. if their relative translation is smaller than $1[mm]$ and their relative rotation smaller that $1[°]$, then the solution is accepted, and the self-calibration is stopped. Otherwise, step 4 is initiated:
4) The robot moves the surgical instrument to a new position, inside the safety volume, that is at least $0.5[mm]$ away from the previous positions, to avoid repetitions. Since a surgical instrument always enters through a fixed entry point in the patient, we are only concerned about avoiding collisions between the tip of the instrument and the tissue. Therefore, a safety volume is defined by the surgeon in step 1 and the surgical instrument always remains inside the safety volume. After moving the surgical instrument, the robot goes back to step 2. Note that the ECM arm remains static during the entire self-calibration process.

At the end of the self-calibration, we have an accurate estimation of ${}^C T_B$ and therefore ${}^B T_{BE}$ can be calculated using Eq. (14). Once ${}^B T_{BE}$ is known, the transformation from the camera to the PSM base (${}^C T_B$) can always be estimated, even when the camera moves to a new position, which happens frequently in surgery.

$$ {}^C T_B = ({}^H T_C)^{-1} \times ({}^{BE}T_H)^{-1} \times {}^{BE}T_B \tag{15}$$

As mentioned in the introduction, ${}^C T_B$ is the fundamental transformation for autonomous robots, since the robot perceives the surgical scene using the endoscopic camera (C) as reference, but operates using as reference the PSM's base (B).

E. Autonomous calibration during surgery using a single image

We made a separate experiment, to test whether it is possible to estimate this transformation ${}^C T_B$ using a single image. The advantage of using a single image is that it would be faster than the method in proposed in Sec. II-D. This single image captures the surgical instrument after being inserted in the new patient by the surgeon. The motivation for this experiment was that if it was possible to estimate ${}^C T_B$ using a single image, then ${}^B T_{BE}$ could also be estimated from a single image, using Eq. (14), without needing to move around the surgical instrument for the self-calibration. As explained in Sec.II-C, ${}^S T_M$ is accurately known and fixed. Hence, it can be used as pre-knowledge for the self-calibration. Using this fixed transformation (${}^{S}T_M$), and a single image-kinematic pair, ${}^{C}T_B^*$ could be estimated as:

$$ {}^{C}T_B^* = ({}^{C}T_M) \times ({}^{S}T_M)^{-1} \times ({}^{B}T_S)^{-1} \tag{16} $$

However, the resulting ${}^{C}T_B^*$ is a rough estimation since it is calculated using a single image-kinematic pair. At this stage, the accuracy of ${}^{C}T_B^*$ depends mostly on the accuracy of the visual method, which estimates ${}^{C}T_M$, since the other two transformations, ${}^{S}T_M$ and ${}^{B}T_S$, are accurate. Consequently, here we propose an optimization to refine ${}^{C}T_B^*$ without using ${}^{C}T_M$. This refinement is done by minimizing the pixel reprojection error between the detected feature points of the cylindrical marker, and the reprojected points of the same features but using kinematic information instead. The rough ${}^{C}T_B^*$ is used for initial value of the following cost function:

$$ \underset{{}^{C}T_B^s}{\arg\min} \sum_{i=1}^{m} \left\| \mathbf{P}({}^{C}T_B^s \times {}^{B}T_S \times {}^{S}T_M \times x3d_i) - x2d_i \right\|^2 \tag{17} $$

where, m is the total number of detected features in the marker; each i^{th} feature has a fixed 3D coordinate $x3d_i$, in the marker's coordinate frame, and a 2D pixel position $x2d_i$, in the endoscopic image; (${}^{B}T_S$) is measured directly from the PSM kinematics; $\mathbf{P}(\cdot)$ is the projection function of camera, which converts a 3D position in the camera's coordinate system into its corresponding 2D image pixel position, using the camera intrinsic and distortion parameters. ${}^{S}T_M$ is the transformation from the marker to the shaft, which is know before surgery Sec.II-C; ${}^{C}T_B^s$ consists of 6 parameters that need to be refined, i.e. three translation parameters (x,y,z) and three rotation vector (r_x,r_y,r_z) parameters. This refinement is achieved by minimizing the cost function in Eq. (17), using a Trust Region Reflective algorithm (TRF) [17], implemented using Python's Scipy [18] library.

III. Experiment and result

A. Experiment setup

Our proposed framework is based on the da Vinci® Research Kit [19]. This robotic platform allows us to control both an endoscopic camera, with a resolution of 720×576, and a 8[mm] surgical instrument. A state-of-the-art cylindrical marker [9] was wrapped around the surgical instrument which is used both in the proposed fusion and self-calibration methods. The ultrasound images were captured using a UTS-533 linear array Ultrasound Probe connected to a ProSound® Alpha 10 ultrasound machine. The robot was controlled by a computer using the Robot Operation System (ROS) Kinetic. This computer was equipped with an Intel® Core i5-3317U and 6 GB RAM. For collecting ground truth data and tracking tissue motion, an OptiTrack® System was used.

B. Tissue tracking

Depending on the autonomous surgical task, the target pose ${}^{T}T_C$ is defined accordingly. For autonomous tissue tracking, ${}^{T}T_C$ will typically correspond to a 3D points on the surface of the tissue that is tracked using stereo-vision data [3]. Here, since the focus of the paper is the tracking of the surgical instruments and not of the tissue, we used an external OptiTrack® system. Any soft-tissue tracking approach can be used instead. In our work, the tissue has been tracked using three OptiTrack® sphere markers rigidly attached to a kidney phantom, as illustrated in Fig.1 (P1). Therefore, a rigid transformation from the optical spheres to a point on the surface of the tissue can be used to set a target point ${}^{T}T_C$ to be scanned using the ultrasound probe. Ultrasound scanning requires control of an Ultrasound probe, which is rigidly attached to the PSM's end effector. Therefore, ${}^{P}T_E$ is estimated before the autonomous tissue scanning task using: ${}^{P}T_E = {}^{P}T_T \times {}^{T}T_C \times {}^{C}T_B \times {}^{E}T_B^{-1}$. According to Sec.II-A, ${}^{P}T_T$ is an identity matrix. ${}^{C}T_B$ is calculated from Sec.II-D. ${}^{E}T_B$ is the end effector pose w.r.t PSM base, which is directly measured from the PSM's kinematics.

C. Accuracy of data fusion

In our framework, the fusion node outputs the transformation ${}^{S}T_C$. To evaluate the fusion result, we used the OptiTrack® System to collect ground truth data. The OptiTrack® System has high accuracy with mean error of 0.044 [mm], and thus can be used to obtain the ground truth of ${}^{S}T_C$, given the rigid transformation between (S) and (P2), as illustrated in Fig.1.

To evaluate our proposed fusion method, we estimated the shaft pose in one normal situation and four challenging situations including (1) occlusion (out of FoV), (2) high instrument speed, (3) changing illumination and (4) a complex situation including all the previous challenges. We also compared our proposed fusion method with Zhang's fusion approach by using the framework in [8] to fuse our visual and kinematic data. In all five situations, translation and rotation errors shown in Table IV were calculated using the Euclidean distance of translation vectors and the Inner Product of Unit Quaternions, which shows that our method has a lower translation and rotation error in both normal and challenging situations, compared with method in [8]. This verifies that the proposed update of the EKF weights improves tracking accuracy.

Finally, to evaluate the performance of our fusion method in the presence of severe noise, artificial noise was added to the kinematic and vision inputs for all the situations described in Table. IV. In practice, we added random noise between -10[mm] to 10[mm] in x,y,z and from -0.01 to 0.01 in q_x,q_y,q_z of EKF state vector, which is mentioned in Sec. II-B.1 separately. Table.IV shows that our method has an overall higher accuracy compared with the-state-of-the-art [8]. Specifically, in a complex situation with kinematic noise, as shown in Fig.3, our fusion method was significantly more accurate. This is due to our method, assigning a higher weight to the more reliable sensor, which was the visual input in this case.

D. Self-calibration Accuracy

1) Self-calibration using multiple images: The set-up joints were moved five times to test the accuracy of the self-

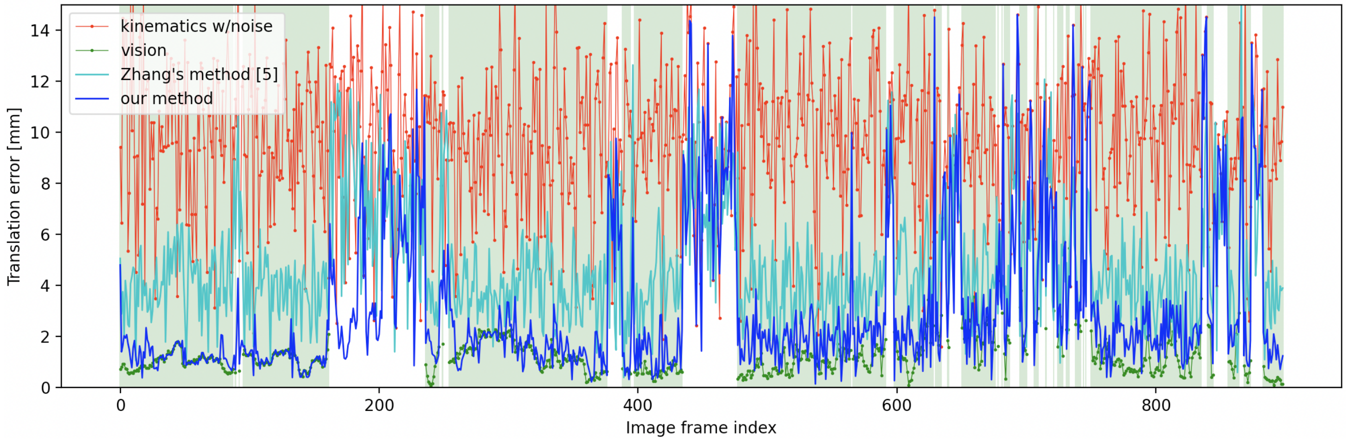

Fig. 3: Translation error of fusion at complex situation with random noise on kinematics input.

TABLE IV: Translation and (Rotation) error for fusion (Mean±Std) [mm] and (°). DR is the detection rate for the vision

	Raw experiment					Kinematics w/noise			Vision w/noise		
Situation	DR	Vision* [9]	Kinematics	Zhang's Fusion[8]	Our Fusion	Kinematics	Zhang's Fusion[8]	Our Fusion	Vision	Zhang's Fusion[8]	Our Fusion
Normal Speed	100	0.41±0.72 (0.51±0.40)	0.96±0.50 (0.79±0.43)	0.76±0.41 (0.75±0.57)	0.42±0.42 (0.47±0.57)	9.62±2.79 (5.31±1.93)	3.74±1.18 (4.35±3.81)	0.99±0.61 (1.70±0.62)	9.51±2.76 (2.58±2.47)	4.42±1.98 (2.21±1.77)	0.87±0.47 (0.84±0.55)
High Speed	87	2.21±0.85 (2.00±1.86)	1.81±0.86 (0.87±0.39)	2.05±0.83 (1.98±1.78)	1.91±0.87 (1.21±0.82)	9.90±3.12 (5.41±1.99)	4.43±2.00 (4.74±2.34)	2.26±1.95 (3.11±2.12)	9.85±2.85 (3.21±3.21)	4.44±2.02 (2.35±1.00)	2.16±0.87 (0.92±0.21)
Marker Occlusions	71	0.62±0.35 (0.67±0.52)	1.49±0.59 (0.91±0.51)	1.13±0.74 (0.80±0.56)	0.88±0.67 (0.78±1.08)	9.76±2.83 (5.13±1.89)	4.78±2.42 (4.54±2.48)	3.02±3.42 (2.36±1.79)	9.53±2.79 (2.59±2.97)	4.96±2.44 (2.12±2.35)	1.46±0.64 (1.08±0.73)
Changing Illumination	75	1.12±0.62 (0.47±0.15)	1.20±0.64 (0.30±0.19)	1.15±0.68 (0.35±0.46)	1.10±0.70 (0.32±0.28)	9.66±2.85 (5.28±1.82)	4.57±2.11 (3.56±1.45)	2.23±2.10 (2.47±1.79)	9.55±2.93 (2.68±2.61)	4.21±2.57 (2.08±1.34)	1.26±0.64 (1.12±0.31)
Complex Situation	75	1.18±0.67 (1.77±2.15)	1.48±0.67 (0.80±0.45)	1.36±0.81 (1.99±1.93)	1.22±0.60 (1.05±1.15)	9.57±2.75 (5.28±1.96)	4.86±2.33 (7.12±7.41)	2.90±2.75 (3.07±2.20)	9.67±2.84 (3.14±3.30)	3.37±1.48 (4.59±3.42)	1.53±0.70 (1.27±1.81)

calibration of ${}^{C}T_B = Z$, which was automatically estimated by the robot as described in Sec.II-D. To get each of the ground truth Z_{gt}, the ECM was held at a static position while the surgical instrument moved for a total of 300 different poses. When capturing data from each new pose, we stopped the surgical instrument for five seconds to avoid any motion blur on the captured images which would negatively affect the calibration. The average measured error between Z_{gt} and the prediction Z, was 0.99 $[mm]$ for translation and 0.47 [°] for rotation, indicating that the self-calibration achieves a very similar calibration result when compared to the one obtained with human assistance. This comparison was done by calculating the Euclidean distance for the translation error and the Inner Product of Unit Quaternions [20] for the rotation error.

2) Self-calibration using a single image: Similarly, each Z_{gt} was compared to the predictions made using a single image Z_{single}. In single image calibration, the average translation error is 2.83 $[mm]$ and the average rotation error was 1.81 [°].

E. Ultrasound Stability

For the ultrasound stability experiment, the ultrasound probe was first placed over the target pose. At the initial target pose after calibration, an ultrasound image was collected, which was used as template to compare subsequently collected ultrasound images. The logic is that if our framework is able to follow the target pose, and therefore compensate the motion of the tissue, the ultrasound image should look similar throughout the entire motion of the tissue. This similarity was measured by computing the Normalized Cross-Correlation (NCC) between the initial ultrasound image and each of the other ultrasound images. For this experiment, the tissue was moved up and down along the camera's Y-axis orientation. As shown in Fig.4, when using our framework (green line) the ultrasound image is stable, scoring an average NCC value of 96%. Without motion compensation (red line), the ultrasound images are not stable. Without motion compensation, the NCC score is high when the tissue is back to the original pose, due to the periodic motion, and the NCC is low otherwise. In the bottom of Fig.4, some ultrasound images are shown, corresponding to the initial ultrasound image and the subsequent ones with index from 100 to 800 with a step of 100 frames in between. With our motion compensation, it is clear that the ultrasound image is stable. On the contrary, without motion compensation the ultrasound images are not stable. When the ultrasound image is mostly black it means that the ultrasound probe is not in proper contact with the tissue.

IV. Conclusion and Discussion

Overall our fusion method was more consistent between the different challenging situations. The fusion results could have been improved by adding joint's acceleration measurements to the framework, since the constant velocity assumption would not be required. However, the current acceleration

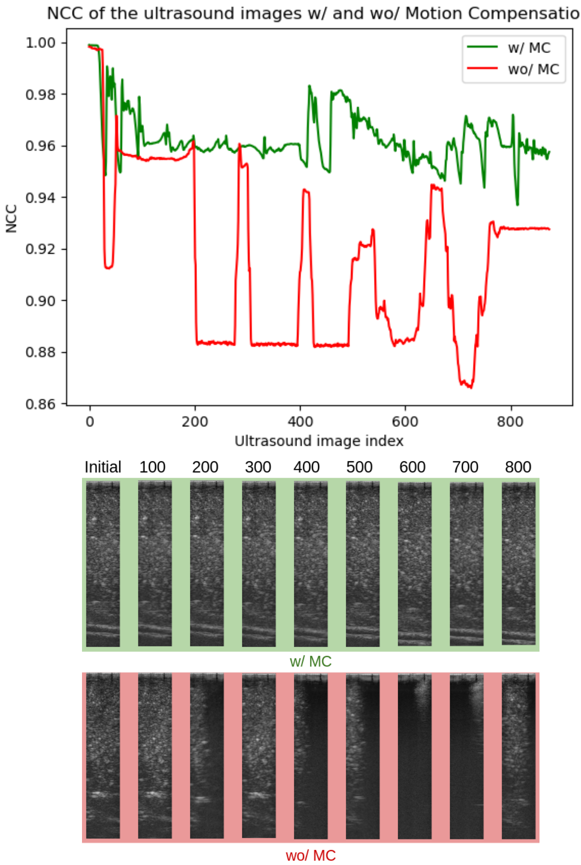

Fig. 4: Ultrasound stability with the tissue moving up and down to simulate a patient's breathing.

measurements provided by the dVRK are noisy. Regarding the choice of the fusion method, EKF is used in this work but other filters, such as the Unscented Kalman Filter (UKF) can be applied. Deep learning could be used to replace our fuzzy logic algorithm for weight estimation but this would require big training datasets. We plan on getting even more accurate results by using a endoscope with higher frame rate. A higher frame rate would improve the accuracy of the poses provided by the vision method since it would remove motion blur, and it would also increase the sampling frequency of our fusion method. Another improvement would have been to use both stereo images, instead of the left image only, but this would require accurate estimation of the transformation between the left and right images. However, since there is still a large usage of monocular endoscopic cameras in surgery, we have decided to make our method as generalisable as possible by using the left-camera images only. As future work, we plan on validating our framework on ex vivo and in vivo experiments.

References

[1] Varier, V.M., Rajamani, D.K., Goldfarb, N., Tavakkolmoghaddam, F., Munawar, A. and Fischer, G.S., 2020, August. Collaborative Suturing: A Reinforcement Learning Approach to Automate Hand-off Task in Suturing for Surgical Robots. In 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1380-1386). IEEE.

[2] Han, L., Wang, H., Liu, Z., Chen, W. and Zhang, X., 2020. Vision-based cutting control of deformable objects with surface tracking. IEEE/ASME Transactions on Mechatronics.

[3] Zhan, J., Cartucho, J. and Giannarou, S., 2020, May. Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11147-11154). IEEE.

[4] Loschak, P.M., Brattain, L.J. and Howe, R.D., 2016. Algorithms for automatically pointing ultrasound imaging catheters. IEEE Transactions on Robotics, 33(1), pp.81-91.

[5] Zhang, Z., Rosa, B., Caravaca-Mora, O., Zanne, P., Gora, M.J. and Nageotte, F., 2021. Image-guided control of an endoscopic robot for OCT path scanning. IEEE Robotics and Automation Letters, 6(3), pp.5881-5888.

[6] Penza, V., Cheng, Z., Koskinopoulou, M., Acemoglu, A., Caldwell, D.G. and Mattos, L.S., 2021. Vision-guided autonomous robotic electrical bio-impedance scanning system for abnormal tissue detection. IEEE Transactions on Medical Robotics and Bionics, 3(4), pp.866-877.

[7] Ye, M., Zhang, L., Giannarou, S. and Yang, G.Z., 2016, October. Real-time 3d tracking of articulated tools for robotic surgery. In International conference on medical image computing and computer-assisted intervention (pp. 386-394). Springer, Cham.

[8] Zhang, L., Ye, M., Giataganas, P., Hughes, M. and Yang, G.Z., 2017, May. Autonomous scanning for endomicroscopic mosaicing and 3D fusion. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3587-3593). IEEE.

[9] Cartucho, J., Wang, C., Huang, B., Elson, D., Darzi, A., and Giannarou, S., 2021. An Enhanced Marker Pattern that Achieves Improved Accuracy in Surgical Tool Tracking. In Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. Taylor & Francis.

[10] Baek, Y.M., Tanaka, S., Harada, K., Sugita, N., Morita, A., Sora, S. and Mitsuishi, M., 2012. Robust visual tracking of robotic forceps under a microscope using kinematic data fusion. IEEE/ASME Transactions on Mechatronics, 19(1), pp.278-288.

[11] Fitzgerald, R. J. Divergence of the Kalman filter. IEEE Trans. Automatic Control, AC-16 (6).pp. 736- 747. 1971.

[12] Sun, Y., Pan, B., Zou, S. and Fu, Y., 2020. Adaptive fusion-based autonomous laparoscope control for semi-autonomous surgery. Journal of medical systems, 44(1), pp.1-13.

[13] Wang, Y.M., 2009. Centroid defuzzification and the maximizing set and minimizing set ranking based on alpha level sets. Computers & Industrial Engineering, 57(1), pp.228-236.

[14] Al-Sharman, M.K., Emran, B.J., Jaradat, M.A., Najjaran, H., Al-Husari, R. and Zweiri, Y., 2018. Precision landing using an adaptive fuzzy multi-sensor data fusion architecture. Applied Soft Computing, 69, pp.149-164.

[15] Park, F.C. and Martin, B.J., 1994. Robot sensor calibration: solving AX= XB on the Euclidean group. IEEE Transactions on Robotics and Automation, 10(5), pp.717-721.

[16] Shah, M., 2013. Solving the robot-world/hand-eye calibration problem using the Kronecker product. Journal of Mechanisms and Robotics, 5(3), p.031007.

[17] Byrd, R.H., Schnabel, R.B. and Shultz, G.A., 1988. Approximate solution of the trust region problem by minimization over two-dimensional subspaces. Mathematical programming, 40(1), pp.247-263.

[18] Virtanen, P. et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, pp.261–272.

[19] Kazanzides, P., Chen, Z., Deguet, A., Fischer, G.S., Taylor, R.H. and DiMaio, S.P., 2014, May. An open-source research kit for the da Vinci$^{®}$ Surgical System. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 6434-6439). IEEE.

[20] Huynh, D.Q., 2009. Metrics for 3D rotations: Comparison and analysis. Journal of Mathematical Imaging and Vision, 35(2), pp.155-164.