2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

ArthroSLAM: multi-sensor robust visual localization for minimally invasive orthopedic surgery

Andres Marmol, Peter Corke, Thierry Peynot

Abstract*— Minimally invasive arthroscopic surgery is a very challenging procedure that requires the manipulation of instruments in limited intraarticular space using distorted and sometimes uninformative images. Localizing the arthroscope reliably and at all times w.r.t. surrounding tissue is of fundamental importance to prevent unintended injury to patients. However, even highly-trained surgeons can struggle to localize the arthroscope using poor image feedback. In this paper, we propose and demonstrate for the first time a visual Simultaneous Localisation and Mapping (SLAM) system, termed *ArthroSLAM*, capable of robustly and reliably localizing an arthroscope inside a human knee joint. The proposed system fuses the information obtained from the arthroscope, an external camera mounted on an arthroscope holder, and the odometry of a robotic arm manipulating the scope, in an Extended Kalman Filter framework. Also for the first time, we implement five alternative strategies for localization and compare them to our method in a realistic setup with a human cadaver knee joint. *ArthroSLAM* is shown to outperform the alternative strategies under various challenging conditions, localizing reliably and at all times with a mean Relative Pose Error of up to 1.4mm and 0.7°. Additional experiments conducted with degraded odometry data also validate the robustness of the method. An initial evaluation of the sparse map of a knee section computed by our method exhibits good morphological agreement. All results suggest that *ArthroSLAM* is a viable component for the robotic orthopedic surgical assistant of the future.

I. INTRODUCTION

Arthroscopy is a Minimally Invasive Procedure (MIP) used for the diagnosis and treatment of joint disorders. A continuous stream of images is captured through an arthroscope, which allows surgeons to visualize the inside of the joint and to manipulate instruments with respect to anatomical regions of interest. Tissues are tightly arranged and in the best cases clearance only reaches up to 7mm [1]. The limited intraarticular space makes this manipulation challenging and error prone, often resulting in unintended collisions of the arthroscope with the surrounding tissue [2]. We believe that a vision-guided robotic arm will eventually relieve the surgeon from manipulating the scope, thus minimizing the chances of collisions. Our research aims to develop a vision-based orthopedic surgical assistant that can continuously and reliably localize the arthroscope with millimeter accuracy, and estimate a 3D map of the joint.

Instrument localization in the surgical context has been demonstrated using external fiducials such as Electromagnetic (EM) or optical trackers. However, these approaches are typically limited by field interferences, probe placement and line-of-sight occlusions. Furthermore, localization is absolute w.r.t. an external frame of reference and no additional information can be inferred about the surrounding tissues.

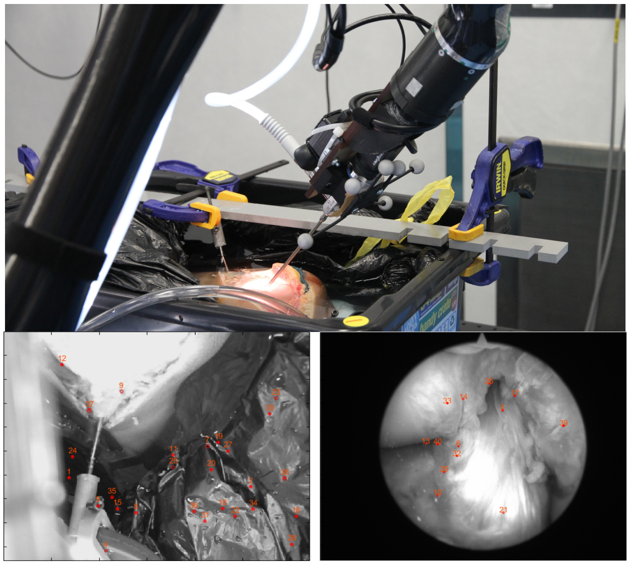

Fig. 1: (Top) Robotic surgical assistant imaging a human limb. Two visual sensors are rigidly attached to a robotic arm: the arthroscope and an external camera provide non-overlapping views of the knee anatomy. Bottom: Keypoint detection in the images acquired from the external camera (left) and arthroscope (right).

Alternatively, monocular Structure-from-Motion (SfM) or Simultaneous Localisation and Mapping (SLAM) techniques can be used with current surgical setups with very limited adjustments required, if any. Besides estimating the camera trajectories, a key advantage of these techniques lies in the fact they also provide either sparse or dense 3D reconstructions of neighboring tissues [3]. Delayed initializations and/or metric priors are commonly required for proper conditioning and scale recovery due to the bearing-only nature of the sensor [4], [5]. Arthroscopy has unique and challenging imaging conditions, which might explain the lack of successful demonstration of visual SLAM in the literature. Limited space, narrow views and textureless tissues combine to reduce the image content's saliency. Furthermore, blood and floating tissue can prevent the acquisition of useful images altogether.

SLAM literature has repeatedly shown that the fusion, or combination, of data from several sensor modalities is

This research was partly funded by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).

All authors ({andres.marmolvelez, peter.corke, t.peynot}@qut.edu.au), are with the ARC Centre of Excellence for Robotic Vision, Queensland University of Technology, Brisbane, QLD, 4000, Australia.

978-1-5386-8093-3/18/$31.00 ©2018 IEEE

a recommended strategy to resolve the scale ambiguity and increase the overall performance and reliability of a monocular localization system in challenging environments [4], [6]. Therefore, in view of the challenges faced in arthroscopy, in this paper we propose a visual SLAM system that makes use of a second camera imaging the exterior of the patient's body, in addition to the arthroscope. We also exploit the odometry of the robotic arm holding the arthroscope and reconstruct trajectories and map *at scale*. An image of our proposed prototype surgical assistant is shown in Fig. 1. In our proposed strategy, *ArthroSLAM*, a kinematic model is used to predict the location of the scope held by the robotic arm. An Extended Kalman Filter (EKF) refines the estimated scope's location by fusing external and intra-articular non-overlapping observations gathered by the two cameras.

The paper provides a comprehensive experimental validation using video sequences of a human joint. *ArthroSLAM* is shown to robustly and reliably estimate six trajectories of common exploratory motions that were recorded from a knee arthroscopy mockup performed on a cadaver. We also propose and implement five partial fusion localization strategies for the first time in the arthroscopy context, and compare them against *ArthroSLAM*.

In addition, we demonstrate the robustness of the proposed approach when provided with odometry data of various accuracy levels as well as with trajectories with artificially-induced noise. In all experiments *ArthroSLAM* localizes uninterruptedly *at scale* during run time, with minimum drift. It does not require pre- or post-process stages for initialization or scale recovery and it is also robust to temporary uninformative content in the arthroscopic images. Our strategy localizes the arthroscope reliably with a mean Relative Pose Error (RPE) of up to 1.4mm and 0.7°, demonstrating its feasibility for intraarticular navigation. Furthermore, an initial evaluation of the sparse 3D map generated by our method suggests good agreement with the joint's morphology without scale ambiguity. To the best of the authors' knowledge this work is the first successful demonstration of robust visual SLAM in arthroscopy under realistic conditions, i.e. close to the conditions experienced in actual surgery.

The remainder of the paper is organized as follows: Section II reviews existing localization strategies from related medical domains. Sections III and IV describe relevant localization methods for arthroscopy including our proposed approach. The methodology to conduct and evaluate the experiments is discussed in Section V, while the results are analyzed in Section VI. We conclude the paper in Section VII with future research directions.

II. LITERATURE REVIEW

Robust scope localization is of great importance during orthopedic procedures in order to ensure patients' safety and prompt recovery. Current navigation solutions resort to expensive passive or active tracking devices that have to be added to the normal surgical workflow. Most commonly accepted is the use of reflective markers attached to the distal end of the instruments [7]. The markers are tracked with special infrared cameras typically mounted on tripods or large trussed rigs. In this motion capture system it is critical that the line of sight between markers and cameras remains unobstructed. Alternatively, active fiducials, such as EM probes, can be tracked inside the body, though in a small volume close to a transducer. The probe's footprint, its isolation from other EM sources and the safe fixation to the instrument are also common issues when considering these tracking solutions [8]. Irrespective of their operation principle, both of these tracking devices do not provide any information about the surrounding tissues.

Interestingly, the introduction of robotic platforms to the operating room has yet to tap on the full potential of the robots' own sensory data for localization. Currently, surgical robots are vastly utilized in hospitals as slave tools under the direct control of the surgeon [9]. Current trends seek to exploit the robot's proprioception actively, such as demonstrated by few semi-autonomous surgical robotic systems in neuro- and urologic surgery [9], [10].

Earlier work demonstrated the feasibility of applying off-the-shelf SfM and monocular SLAM algorithms to MIP images. The application of feature-based methods in the context of abdominal and cardiac procedures [11] has been shown to recover the scope's trajectory as well as a sparse 3D representation of the surrounding tissue (*up to scale*).

Recently, feature-based and direct methods have shown notable progress in outdoor scenes [12], yet performance is brittle when deployed in-vivo. MIP images, in particular arthroscopic images, include: 1) limited baseline motions, 2) non-rigid, smooth and homogeneous tissues, 3) unreliable or insufficient salient image content. Under these conditions, the use of a single monocular sensor would inevitably lead to delayed, discontinuous or unreliable estimates. Furthermore, the use of a bearing-only sensor typically requires careful initialization and an offline scale-recovery procedure. Altogether, the usefulness of the method's estimates would be heavily compromised.

Sensor fusion strategies have been commonly utilized in the literature to obtain improved information content from distinct and noise-corrupted sensors' signals. In the medical context, [6] demonstrated the fusion of both magnetic and visual sensory data for the robust localization of a wireless capsule robot during gastrointestinal procedures. In turn, [13] combined optical tracking data with electromagnetic sensors to localize a needle's tip under deflection. Earlier work by [14] described the fusion of magnetic and inertial data for endoscopic localization.

Fundamentally, the aforementioned work treated localization as a recursive Bayesian estimation problem. In doing so, a filter estimates recursively the probability distribution that represents the object's location and refines it with noisy/uncertain sensor measurements. To the best of our knowledge there is no literature describing either sensor fusion or localization techniques in the context of minimally invasive orthopedic procedures.

In this work we demonstrate for the first time visual SLAM with an arthroscope for minimally invasive orthopedic procedures. Information from the robot's odometry and a second visual sensor are fused with arthroscopic footage using an EKF. This fusion allows for the estimation of both camera poses *and* map estimates *at scale* during run time, without requiring initialization or scale-recovery stages. The additional sensors add reliability to our SLAM system by counteracting the limited amount of salient information as typically seen in arthroscopic images.

III. ArthroSLAM

Our approach is motivated by the need for a localization solution that can: 1) provide trajectory and map estimates *at scale* without requiring initialization or offline scale recovery stages, 2) operate continuously even under temporarily uninformative sensor observations and, 3) localize reliably with appropriate accuracy in the human joint. This section describes our strategy to combine sensor information in order to address the above requirements. Images from the inside of a human joint are already available as part of normal arthroscopy workflow. Nonetheless, single camera-based localization is unreliable under challenging imaging conditions and cannot recover the scale. In order to overcome the previous limitations, we propose the use of an inexpensive external camera looking at the patients' limb as well as odometry data typically available in surgical robots. We rigidly attach the external camera to a customized holder installed at the final link of a robotic arm (see Fig. 3). The camera observes the *exterior* of the patient's leg and common surgical equipment. This view has no visual overlap with the images of the arthroscopic camera but it is typically much clearer and richer in texture (see Fig. 1). Although potentially helpful, no artificial objects need to be introduced in the scene. We further utilize the odometry of a robotic arm, not only to resolve for the scale ambiguity but also to compensate for localization drift common to visual odometry methods.

Formally, we utilize an EKF formulation to solve for the arthroscope's localization as a function of the robotic arm's motion commands (u) and the visual sensors' observations (z). The filter estimates the true pose of the scope (x) using two discrete-time nonlinear models. A process model (f) as shown in Equation (1) is used to simplify the real robot dynamics. The process noise v, is included to account for modeling inaccuracies. A second model, as seen in Equation (2), is used to explain the sensor measurements as a function of the robot's true pose and sensor noise, w. In both models, k indicates the time step.

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k) \tag{1}$$

$$\mathbf{z}_k = h(\mathbf{x}_k, \mathbf{w}_k) \tag{2}$$

In our system the process model is governed by the kinematics of a 6 Degrees of Freedom (DoF) robot arm. A nonlinear Denavit-Hartenberg model allows us to compute the robot's end effector pose given the angular position of each of the revolute joints. In practice, mechanical discrepancies, backslash and the joint actuators' accuracy will introduce uncertainty in the pose estimate. The robot is driven through an uncalibrated inverse kinetic solver in open loop. We store the pose of the robot as a 7-element state vector (x), with the rotation encoded in a quaternion.

Measurements of the environment are brought into the system by assuming a pin-hole projective model for both monocular cameras. We also account for lens barrel distortion, which is significant for the arthroscopic camera. Sensor measurements correspond to salient keypoints extracted from gray-scale images using Scale-Invariant Feature Transform (SIFT). The choice of features and matching parameters has been informed by a previous study on arthroscopic tissue by the authors [15]. Outlier sensor readings most commonly arise from moving tissue, floating debris or false matching in between the low-saliency arthroscopic images. These outliers are naturally rejected by the filter as they represent sensor motions incompatible with the robot odometry and/or the external camera's observations.

We significantly enhanced a MATLAB toolbox [16] to implement our strategy. In particular we expanded the toolbox to process timestamped streams of images using our own feature detection and matching algorithm with active search.

IV. ALTERNATIVE FUSION STRATEGIES

This section describes alternative strategies to *partially* combine the sensory data. We use the term partial fusion to highlight the fact that not all available sensor data are used concurrently by these methods. In Section VI we compare our method to these strategies. Although all strategies were tuned carefully to maximize their performance, we acknowledge that a detailed sensitivity analysis is outside the scope of the paper. We refer the reader to Fig. 2 for a high-level depiction of the strategies. The external camera and the arthroscope poses are referred to as $\mathrm{P}_{\mathrm{ext}}$ and $\mathrm{P}_{\mathrm{art}}$ respectively. The transformation matrix T between the camera centers is obtained from calibration and assumed constant.

- **Pure odometry (PureOdo.)**: This approach relies only on the odometry data given by the robot. While the method could potentially localize the arthroscope, it would not offer any information about the surrounding tissues in the joint.
- **Pure arthroscopic SfM (ArthroSfM)**: A classical SfM pipeline [5] is carried out using only the intraarticular images to estimate the poses of $\mathrm{P}_{\mathrm{art}}$. The strategy proceeds as follows: 2D correspondences between consecutive arthroscopic images are used to find the Fundamental matrix (F) and infer the relative camera poses. Multiple-view triangulation is used to generate 3D points. A Bundle Adjustment (BA) algorithm refines both camera poses and 3D point estimates by minimizing the reprojection error. At least seven correspondences must be established between adjacent frames to allow for the computation of F. An M-estimator Sample Consensus (MSAC) algorithm is used to reject incorrect correspondences. The trajectory is estimated up-to-scale but Singular Value Decomposition (SVD) can be used to recover the scale from a best-fit with the robot's odometry.
- **Dual camera SfM (DualSfM)**: For this strategy, the SfM estimation process described for method *ArthroSfM* is also applied here to the *external* camera's images. The estimated camera poses, $\mathrm{P}_{\mathrm{ext}}$, can be scaled and transformed in the

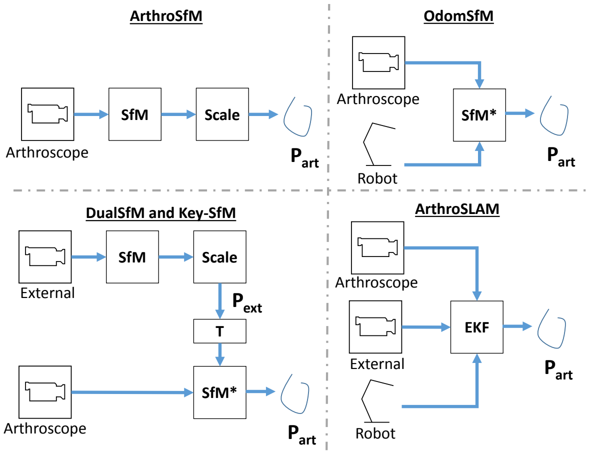

Fig. 2: Overview of five different fusion strategies using combinations of three types of data sources (scope images, external images and robot odometry). $\boxed{\text{SfM*}}$ and $\boxed{\text{SfM}}$: SfM with and without pose priors respectively. $\boxed{\text{Scale}}$: Offline scale-recovery stage. $\boxed{\text{T}}$: Rigid transformation. P_{art} and P_{ext} are the estimated camera poses *at scale*. Strategy *Key-SfM* uses the same topology as *DualSfM* but with a subset of frames.

arthroscope's coordinate frame to provide an approximation of P_{art}. These camera poses, along with any 2D-2D arthroscopic correspondences, are used to initialize a BA optimization. The refined poses describe the arthroscope's trajectory without scale ambiguity.

- **Keyframe-based arthro SfM (Key-SfM)**: In this strategy features are tracked through the image sequences acquired with the external camera, however, the poses P_{ext} are estimated and optimized only for a small subset of *keyframes*. This is known to improve the epipolar geometry conditioning and we used ORBSLAM2 [12] to compute such keyframes. The method proceeds as described for strategy *DualSfM* using only these keyframes.
- **Prelocalized arthro SfM (OdoSfM)**: The robot odometry provides an initial estimate of the arthroscope's pose (P_{art}). Contrary to the strategy of *ArthroSfM*, the fundamental matrix arising from arthroscopic correspondences is used for outlier rejection and not for relative pose estimation. Thanks to the pre-initialization with the robot odometry, the estimated trajectory is already scaled.

V. EXPERIMENTAL SETUP

A. Surgical assistant platform

Our robotic assistant features a *JacoV2* 6-DoF robotic manipulator with a customized holder to which both the arthroscope and the external camera are rigidly attached. The system is depicted in Fig. 1.

A FLIR 1288×964px *Blackfly* U3-13S2C camera is attached to a commercial Stryker *Ideal Eyes* 0502-704-030 arthroscope. We refer to these two elements as the *arthrocam*. Scopes of similar specifications (4mm diameter, 30° oblique optics, 115° wide angle) are common across manufacturers and specialties (shoulder and ankle arthroscopy). The external camera is a 640×480px Logitech *HD270* commercial webcam. The camera was positioned to look slightly above the arthroscope tip providing views of the patient's limb and the surgical setup (see Fig. 1).

The cameras were calibrated offline using checkerboard patterns [17]. The *arthrocam* was calibrated underwater using a sealed high-quality printed checkerboard of 0.54cm^2 area. This procedure is necessary to account for the water's diffraction as present during irrigated arthroscopy. The calibration pixel errors were $\{0.7394, 0.7263\}$px and $\{0.1299, 0.1117\}$px for the *arthrocam* and the external camera respectively. Cameras with better optical properties are available in the arthroscopy and consumer market, e.g., HD/4K. However, in this paper we demonstrate the feasibility of our system with low-grade and economic alternatives and expect improved performance with better equipment.

B. Ground truth systems

An Optitrack *Flex13* motion capture system was used to track the customized arthroscope holder with a rig of 6 infrared-reflective markers. The system's 8 infrared cameras provided ground truth poses with an accuracy of 0.2mm.

The hand-eye calibration toolbox by [18] provided the relative transformation from the Optitrack markers' rig to the *arthrocam* and the external camera (${}^{\text{opt}}\text{T}_{\text{art}}$ and ${}^{\text{opt}}\text{T}_{\text{ext}}$ respectively). The calibration residuals were $\{0.5177\text{mm}, 0.0132°\}$ and $\{1.0056\text{mm}, 0.2750°\}$ for the *arthrocam* and the external camera respectively.

The relative transformation between cameras (T) can be inferred from these two transformations. Figure 3 illustrates the location of the aforementioned coordinate frames. Hand-eye transformations are assumed rigid in this work. This assumption holds true if the customized arthroscopic tool does not bend during the procedure. Section V-D describes how the experimental setup ensures that no reaction forces oppose the arthroscope's motion.

An Artec *Eva* 3D scanner was used to scan the internal tissue with up to 0.1mm accuracy. The resulting point cloud was used as reference to assess *ArthroSLAM*'s map in Section VI-D. Due to a lack of anchoring points between the Optitrack and Artec systems, the point cloud could not be properly aligned in the Optitrack frame of reference.

C. Evaluating the arthroscope localization

During surgical procedures the relative distance from the scope to the tissues is of particular interest. However, in this paper we assess the estimated trajectories and the maps separately due to technical limitations of the ground truth setup. We focus particularly on robust localization but still assess the consistency of our *ArthroSLAM*'s map later in Section VI-D. It is worth noting that in cases where preoperative imaging of the anatomy (i.e. a map) is available [19], robust localization results, such as the one obtained with our method, are relevant.

The motion capture system provides ground truth for the localization evaluation. Since the arthroscope's optical center cannot be fitted with reflective markers during realistic ex-vivo experiments, we use a third reference frame, P_{opt}, whose poses can be accurately tracked.

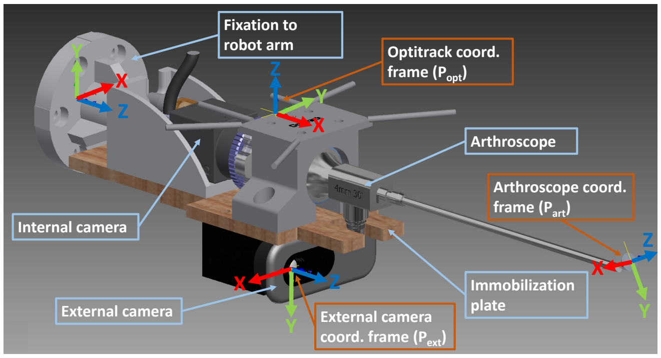
Fixation to robot arm
Optitrack coord. frame (P_{opt})
Arthroscope
Arthroscope coord. frame (P_{art})
Internal camera
Immobilization plate
External camera
External camera coord. frame (P_{ext})
X
Y
Z

Fig. 3: Customized arthroscopic tool: main components (in light blue) and coordinate frames of interest (in orange).

The arthroscope pose estimated with any of the SfM/SLAM strategies can be used to infer the pose of the Optitrack coordinate frame using the appropriate hand-eye transformation. We refer to the estimated poses as $\tilde{P}_{opt}$. The resulting trajectory is then compared with the motion-capture ground truth. We evaluate both the Absolute Trajectory Error (ATE) and RPE [20] to quantify both absolute localization and drift errors. All estimations are run three times and we report the mean of the errors.

D. Dataset

Six video sequences were recorded during a mock-up arthroscopy with a harvested leg from a cadaver[1]. Attention was given to emulate the so-called fulcrum effect, which eliminates two of the scope's degrees of freedom: The scope's rod is constrained by the incision point and can only be panned, tilted, rolled and inserted/retracted. It should be noted that given the scope's oblique optics, simply rolling the *arthrocam* provides distinct views of the knee's interior without displacing the scope's tip.

- 'HalfRoll' (HR): Semicircular rotation about the scope's rod (Duration: ≈18 seconds). Robot's single joint rotation for about 170°.
- 'FullRoll' (FR): Circular rotation about the scope's rod (≈44s). Robot's single joint rotation for about 380°. Note that rotations occur about the scope's rod axis and not about its optical axis which is oblique at 30°.
- 'Tilt' (Ti): Exploratory tilt motion (Duration: ≈28s, Displacement: ≈134mm). Robot's multiple joints motion.
- 'Insert/Retract' (IR): Scope's displacement along the rod axis (≈26s, ≈62mm). Robot's multiple joints motion.
- 'PivotsA' (PA): Exploratory pan/tilt motions (≈17s, ≈108mm). Robot's multiple joints motion. An example of this trajectory is shown later in Fig. 5.
- 'PivotsB' (PB): Rapid exploratory pan/tilt motions (≈19s, ≈143mm). Robot's multiple joints motion.

Robot odometry in the rolling sequences was fairly accurate (mean RPE below 1mm and 1°) as the control of a single joint could be done directly in the robot's joint space. Other sequences were less accurate (up to 2mm error): they involved not only all robot joints but also included modeling inaccuracies as part of the robot's inverse kinematics solution. The robot's angular positioning error remains below 1° across all sequences.

All sequences were acquired in an open knee as illustrated in Fig. 1. This setup prevented the direct contact between the arthroscope's rod and the skin, thus ensuring that the hand-eye transformations, in particular ${}^{opt}T_{art}$, remained unchanged during the experiments. The specimen was submerged in water to emulate the effect of irrigation in the image quality: floating debris, non rigid fibrilations and specular reflections are thus present in the images. The knee was immobilized to minimize the effect of strong tissue deformation in the evaluation. Immobilization of the open knee also made it possible to scan the tissue with a structured light scanner and obtain a dense 3D model of the internal structures.

We validated experimentally that the *arthrocam*'s image brightness did not change radically between test cases with closed and open knees. Although counter-intuitive at first, this result is understandable as the arthroscope's illumination is provided through the same rod that gathers the images. During close range imaging, even if the knee is dissected, the arthroscope's light is the dominant light source. Figure 4 depicts two example images contained in the sequences.

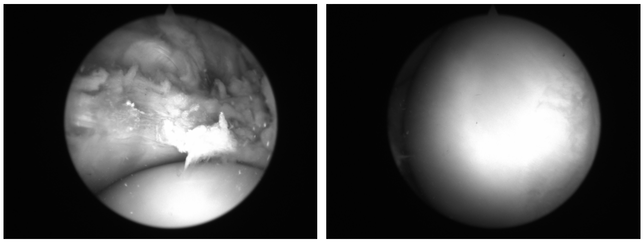

Fig. 4: (Left, from sequence 'HalfRoll'): Arthroscopic image with floating tissues, tissue fibrilation and strong reflection. (Right, from sequence 'FullRoll'): Arthroscopic image of smooth and homogeneous cartilage surface.

VI. EXPERIMENTAL RESULTS

In this section we first validate the benefit of including several sensors for increased localization performance. Next, we evaluate our *ArthroSLAM* strategy and compare it with other partial fusion strategies to estimate the pose of the arthroscope. We further analyze our method's robustness by adding synthetic noise to the robot's odometry. We conclude this section by presenting a preliminary assessment of the reconstructed sparse map provided by our method.

A. ArthroSLAM fusion validation

Table I presents the estimation results of different variants of our approach. To facilitate comparison, we include first the results of the *PureOdo* strategy followed

[1] Data acquisition was approved by the Australian National Health and Medical Research Council (NHMRC) - Registered Committee Number EC00171 under Approval Number 1400000856.

by three *ArthroSLAM* variants: *Odo+Ext*, *Odo+Art* and *Odo+Ext+Art*. In these variants, the EKF fuses the robot's odometry with the visual data from the external camera, the arthroscope or both sensors respectively. The RPE is computed and we show the translational (e_t) and rotational (e_R) errors of the estimated poses. Each experiment is run thrice and we report the mean of the errors. For simplicity we omit the results on sequences 'FullRoll' and 'PivotsA', which exhibit similar trends captured by the other sequences.

TABLE I: Mean RPE for several sensor fusion variants. e_t in mm., e_R in deg. Green cells show the best two results per sequence.

	HR		Ti		IR		PB	
	e_t	e_R	e_t	e_R	e_t	e_R	e_t	e_R
PureOdo	0.4	0.5	0.9	0.1	0.4	0.1	2.1	0.3
Odo+Ext	0.4	0.4	0.5	0.1	0.5	0.1	1.7	0.3
Odo+Art	0.4	0.5	1.0	0.1	0.9	0.1	3.6	0.3
Odo+Art+Ext	0.4	0.4	0.5	0.1	0.4	0.1	1.4	0.3

Comparing to pure odometry localization, the use of a single visual sensor led to similar or slightly worse localization performance. Conversely, the use of both sensors always led to the same or improved localization results (greeen cells in Table I). Sequence 'PivotsB' results also showed that variant *Odo+Art* is prone to higher errors as the redundancy of the external camera is not available to help discarding outliers in the scope's images (floating tissues or false matches). Consequently, and for the remainder of the experimental results, we use the *Odo+Ext+Art* variant of our approach.

B. Arthroscope localization

Table II collects the results of our method along with other partial fusion strategies across three of the six sequences. As before, we use the RPE to report both the mean translational (e_t) and rotational (e_R) errors of the estimated poses. Maximum errors were included to quantify deviations for which the scope would collide with the tissue. We also computed translational errors using the ATE, but observed a similar trend in the results as with the RPE metric. In the interest of simplicity we omit ATE results from the table. Column *%Img* indicates what percentage of the available images were actually used in the estimation.

A common scenario in arthroscopy is the lack of images with salient content. For example in sequence 'FullRoll' a smooth condyle surface occupied most of the camera's field of view preventing the detection of keypoints (see Fig. 4). Furthermore, blood, floating debris and imaging artifacts can lead to incorrect feature matching. These conditions heavily impact strategies that rely solely on the arthroscopic visual feedback, such as *ArthroSfM*.

Strategy *ArthroSfM* was observed to have brittle performance: translational and rotational errors were above 12mm and 2° in three of the sequences ('HalfRoll', 'FullRoll', 'Insert/Retract') while the estimation failed altogether for the remaining three cases ('Tilt', 'PivotsA', 'PivotsB'). Frame-to-frame correspondences could only be established for the full duration of sequence 'HalfRoll' and 40% of the remaining two sequences. The lack of sufficient correspondences immediately implied a localization loss for this strategy.

TABLE II: Mean and maximum RPE localization results across representative sequences. e_t in mm., e_R in deg. Green cells show the two best results per sequence. N/A indicates failed estimation. *ArthroSLAM* results are in bold.

Seq.	Strategy	Mean	Max.	Mean	Max.	%Img
		e_t	e_t	e_R	e_R	
FR	PureOdo.	0.6	3.5	0.7	3.1	-
	ArthroSfM	18.6	53.7	4.8	13.7	38.69
	DualSfM	67.3	164.3	39.1	66.3	100
	Key-SfM	3.6	7.7	1.7	3.5	2.39
	OdoSfM	0.9	13.5	0.8	3.0	100
	ArthroSLAM	**0.5**	**2.8**	**0.7**	**2.6**	**100**
IR	PureOdo.	0.4	1.2	0.1	0.2	-
	ArthroSfM	12.7	43.5	3.4	11.7	40.00
	DualSfM	31.3	85.2	0.8	2.2	100
	Key-SfM	N/A	N/A	N/A	N/A	N/A
	OdoSfM	0.5	2.0	0.1	0.3	100
	ArthroSLAM	**0.4**	**1.6**	**0.1**	**0.2**	**100**
PB	PureOdo.	2.1	5.5	0.3	0.8	-
	ArthroSfM	N/A	N/A	N/A	N/A	N/A
	DualSfM	19.7	49.7	2.5	6.0	100
	Key-SfM	1.0	1.9	0.2	0.3	0.67
	OdoSfM	2.2	5.9	0.3	0.7	100
	ArthroSLAM	**1.4**	**3.8**	**0.3**	**0.8**	**100**

In **strategy *DualSfM*** the use of an external camera allows for better keypoint association between frames. Only in one sequence ('PivotsA') is there an insufficient number of correspondences at about 80% of the sequence's duration. Despite this increase in reliability, the accuracy is the worst of all strategies. The accuracy drop is even more pronounced in the rolling sequences with drift of as much as 67mm and 40°. These results can be explained in the light of the epipolar geometry theory. In the end-effector of our robot, the external camera optical axis is nearly parallel to the arthroscope's rod axis. This arrangement can cause that the points observed by the external camera appear as lying on a plane at infinity. This can lead to ambiguous solutions in the fundamental matrix estimation [21] and ultimately cause the estimation to drift strongly. An alternative to prevent this ambiguity is the use of a delayed initialization based on keyframes for which the epipolar geometry is well-defined.

For **strategy *Key-SfM*** we no longer observed a divergence in the estimation and in fact most of the results could be ranked as the second or third best. It was observed that the pose estimates were short-lived, using only few frames of the available images (up to %Img=2%). The reduced baseline resulting from the fulcrum effect reduced the number of suitable key-frame candidates. The results on such a small subset were inherently biased. As an example, sequence 'PivotsB' exhibited small errors of about 1mm and 0.2° since the metrics were computed over just three neighboring keyframes. It was also found that the external camera's keyframes did not necessarily have associated arthroscopic images that were well-suited for feature matching. This resulted in an estimation largely dependent on the external camera without sufficient information to build a map of the internal anatomy.

The pre-initialization of the camera poses using odometry information allowed **strategy *OdoSfM*** to estimate all trajectories completely, even when it was not possible to establish a continuous tracking of arthroscopic features. The performance was virtually identical to that of strategy *PureOdo*, except for sequence ‘FullRoll’. The results indicate that early introduction of the robot’s odometry can: 1) prevent strong drift in the estimates due to epipolar degeneracies and 2) continue the estimation while the arthroscopic images are uninformative. We also observed that the sole information provided by the arthroscope was not sufficient to increase the localization reliability.

As previously discussed in Section VI-A, **strategy *ArthroSLAM*** either delivers the same or improved localization results when compared with plain odometry. As the motion increases in complexity, our strategy reduces strongly the translational error with trivial impact on the rotational error. The accuracy improved by 20% to 50% in all pivoting sequences when our strategy is compared to *PureOdo*. Overall, our approach is shown to localize the arthroscope with errors up to 1.4mm and 0.7° in average across all evaluated sequences. The maximum errors were observed to reach up to 3.8mm and 2.6° in sequences ‘PivotsB’ and ‘FullRoll’ respectively. Examples of the trajectories can be seen in Fig. 5.

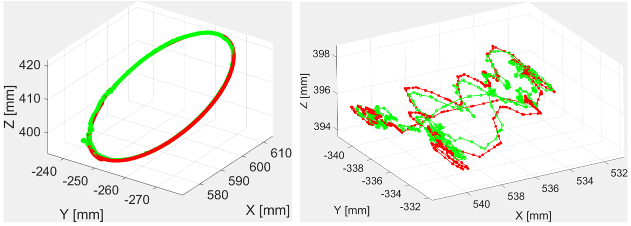

Fig. 5: Ground truth (green) and *ArthroSLAM* estimated trajectory (red) for sequences ‘FullRoll’ and ‘PivotsA’ resp.

Pose estimates are provided uninterruptedly for all image frames. The odometry and the external camera allow for the estimation to succeed even if no salient arthroscopic content is present. Furthermore, the robot’s odometry prevents strong estimation drift under degenerate epipolar conditions. In all cases, *ArthroSLAM* localizes with the smallest mean error, however, the maximum error may still need to be reduced before safe intraarticular navigation can be attempted. Table III provides a high-level overview of the experimental results in view of the system requirements: **robust** to uninformative arthroscopic images, **rel**iable localization in the joint and immediate at-**scale** trajectory and **map** estimates. The lack of visual feedback makes strategy *PureOdo* impervious to any disturbances in the arhtroscopic images, but also unable to provide a map. Section VI-D will later illustrate relative 3D information of the tissues that our method can offer as opposed to this strategy. *Key-SfM* results span a very limited subset of frames making long-term localization unreliable. Strategy *OdoSfM* provides robust, at-scale estimates, but it lacks the added reliability provided by the external camera.

TABLE III: Fulfillment of system requirements per strategy: **Robust**ness, **Rel**iability, immediate **Scale** recovery and **Map** estimation. *ArthroSLAM* results are in bold.

	Robust.	Rel.	Scale	Map
PureOdo.	✓	✓	✓	✗
ArthroSfM	✗	✗	✗	✓
DualSfM	✓	✗	✗	✓
Key-SfM	✓	✗	✗	✗
OdoSfM	✓	✓	✓	✓
ArthroSLAM	✔	✔✔	✔	✔

C. Noise and uncertainty analysis

In order to further confirm *ArthroSLAM*’s performance for cases with unreliable odometry, we proceeded to synthetically worsen the rolling trajectories. Zero-mean Gaussian noise was added to each of the three translational and three rotational increments along a trajectory. Table IV summarizes the results for medium (σ_G of 0.05mm and 0.1°) and high (σ_G of 0.1mm and 0.2°) odometry disturbances. We show the mean and standard deviation of the errors for three runs.

TABLE IV: Sensitivity analysis for rolling sequences (HR, FR) with added medium (med) and high (high) noise. e_t in mm., e_R in deg. Green cells show the best results per sequence. *ArthroSLAM* results are in bold.

Seq.	Strategy	e_t	σ_t	e_R	σ_R
HR-Med	PureOdo.	1.07	-	0.59	-
	ArthroSLAM	**0.70**	**0.12**	**0.20**	**0.05**
HR-High	PureOdo.	1.90	-	1.14	-
	ArthroSLAM	**1.18**	**0.31**	**0.98**	**0.41**
FR-Med	PureOdo.	0.97	-	0.52	-
	ArthroSLAM	**0.77**	**0.25**	**0.31**	**0.07**
FR-High	PureOdo.	1.76	-	0.83	-
	ArthroSLAM	**1.17**	**0.24**	**0.40**	**0.22**

Overall our method is shown to improve the localization accuracy in both cases with medium and high levels of noise added to the odometry. In particular translational and rotational error reductions of up to 35% and 65% respectively can be observed when compared to the *PureOdo.* strategy.

D. Reconstructed map’s assessment

Our open knee arthroscopy setup also allowed us to record ground truth of the intra-articular surfaces. Due to technical limitations the 3D scanner data could not be accurately aligned with the frame of reference used in the localization experiments. As such, our ability to quantitatively evaluate the tissue 3D reconstruction, i.e. the SLAM map, is reduced.

As an alternative validation, we selected three map points close to regions of anatomic interest. We then asked an orthopedic surgeon to select the corresponding points in the 3D model. The correspondences can be observed in Fig. 6. We labeled the points ({A, B, C}) and used them to compute distances between anatomic regions: Distance ‘BC’ approximates the distance between lateral and medial femoral condyles, while distances ‘AB’ and ‘AC’ can be used to approximate the depth of the femoral notch.

We measured the aforementioned distances in both the 3D model and the SLAM map. The distances, in millimeters, are

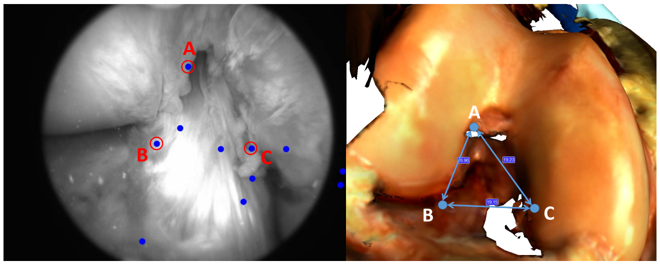

Fig. 6: (Left): Selected map points of interest. (Right): Equivalent point locations in the 3D dense model. The points could describe the morphology of the femoral head.

presented in the format {3D model, SLAM map}: BC{19.2, 20.0}, AB{17.0, 17.9}, AC{19.2, 19.3}. The results show that there is good agreement among the three distances of interest. This would suggest that the map estimated by our method does correctly overcome scale and projective ambiguities.

VII. Conclusion

Minimally invasive arthroscopic procedures are extremely challenging, yet surgical robots have the potential to alleviate the surgeon's workload by performing certain manipulation tasks. Before turning this into a reality, surgical robots need robust localization methods for safe navigation inside the human body. This paper proposed and demonstrated *ArthroSLAM*, the first system for robust SLAM with an arthroscopic camera in a human joint. To overcome the limited and unreliable information that can be extracted from arthroscopic images, the proposed approach exploits an external camera and a robot arm's odometry within an EKF framework. The paper presented the first implementation and evaluation of several alternative localization strategies in the context of minimally invasive orthopedic procedures. Results of realistic ex-vivo experiments demonstrated that our system can continuously estimate the location of the arthroscope *at scale* with up to 1.4mm and 0.7° mean RPE. The system is robust to common challenging imaging conditions: degraded quality due to irrigation, floating debris and fibrillations, as well as the presence of homogenous and smooth tissues. We validated the robustness of our approach by using odometry with several levels of noise both from synthetic and real data. A preliminary assessment also suggested that the 3D map computed with our method could be used in measuring anatomical areas of interest in the knee joint.

In future work, we will improve our setup to better evaluate the quality and accuracy of the SLAM map and, in particular, evaluate the relative positioning of the scope w.r.t. the tissues. We will explore improvements in the robot model and control (i.e. kinematic calibration), as well as determining the optimal placement of the external camera to reduce the maximum localization errors. Finally, in this paper we assumed the arthroscope to be perfectly rigid. Future work would investigate how to take into account the deflection of the scope that arises from contact with the skin.

Acknowledgment

We thank Prof. Ross Crawford (MD), Prof. Jonathan Roberts and Dr. Anjali Jaiprakash for supporting this research. We also thank Dr. Joan Solà, Dr. Teresa Vidal-Calleja, Dr. Anders Eriksson and Dr. Fumio Sasakawa (MD) for fruitful discussions and recommendations.

References

[1] J. Dacre, D. Scott, J. Da Silva, G. Welsh, and E. Huskisson, "Joint space in radiologically normal knees," *Rheumatology*, vol. 30, no. 6, 1991.

[2] A. Jaiprakash *et al.*, "Orthopaedic surgeon attitudes towards current limitations and the potential for robotic and technological innovation in arthroscopic surgery," *J. of Orthopaedic Surg.*, 2017.

[3] B. Lin, Y. Sun, X. Qian, D. Goldgof, R. Gitlin, and Y. You, "Videobased 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey," *Int. J. of Med. Robot. and Comput. Assisted Surg.*, 2015.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Rob.*, vol. 32, no. 6, 2016.

[5] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[6] M. Turan, Y. Almalioglu, H. Gilbertd, H. Araujoe, E. Konukoglub, and M. Sittia, "Magnetic-Visual Sensor Fusion based Medical SLAM for Endoscopic Capsule Robot," *arXiv preprint 1705.06196*, 2017.

[7] Y. Hayashi, K. Misawa, M. Oda, D. J. Hawkes, and K. Mori, "Clinical application of a surgical navigation system based on virtual laparoscopy in laparoscopic gastrectomy for gastric cancer," *Int. J. of Comput. Assisted Radiology and Surg.*, vol. 11, no. 5, 2016.

[8] J. Hummel, M. Figl, W. Birkfellner, M. R. Bax, R. Shahidi, C. R. Maurer Jr., and H. Bergmann, "Evaluation of a new electromagnetic tracking system using a standardized assessment protocol," *Physics in Medicine and Biology*, vol. 51, no. 10, 2006.

[9] M. Hoeckelmann, I. J. Rudas, P. Fiorini, F. Kirchner, and T. Haidegger, "Current capabilities and development potential in surgical robotics," *Int. J. of Advanced Robot. Systems*, vol. 12, no. 5, 2015.

[10] M. Yip and N. Das, "Robot autonomy for surgery," *arXiv preprint 1707.03080*, 2017.

[11] P. Mountney, D. Stoyanov, and G. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Mag.*, vol. 27, no. 4, 2010.

[12] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, 2015.

[13] B. Jiang, W. Gao, D. F. Kacher, T. C. Lee, and J. Jayender, "Kalman Filter Based Data Fusion for Needle Deflection Estimation Using Optical-EM Sensor," in *Med. Image Computing and Comput. Assisted Intervention*, 2016.

[14] H. Ren, D. Rank, M. Merdes, J. Stallkamp, and P. Kazanzides, "Multisensor data fusion in an integrated tracking system for endoscopic surgery," *IEEE Trans. Inform. Technol. Biomed.*, vol. 16, no. 1, 2012.

[15] A. Marmol, T. Peynot, A. Eriksson, A. Jaiprakash, J. Roberts, and R. Crawford, "Evaluation of keypoint detectors and descriptors in arthroscopic images for feature-based matching applications," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, 2017.

[16] J. Sola, D. Marquez, J.-M. Codol, and T. Vidal-Calleja, "An EKF-SLAM toolbox for MATLAB," http://homepages.laas.fr/jsola/JoanSola/eng/toolbox.html, 2009.

[17] J.-Y. Bouguet, "Camera calibration toolbox for Matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/, 2002.

[18] C. Wengert, "Hand-eye calibration add-on for the Matlab camera calibration toolbox," 2016.

[19] C. Hennersperger, B. Fuerst, S. Virga, O. Zettinig, B. Frisch, T. Neff, and N. Navab, "Towards MRI-Based Autonomous Robotic US Acquisitions: A First Feasibility Study," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, 2017.

[20] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," *IEEE Int. Conf. on Intelligent Robots and Systems*, 2012.

[21] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *Int. J. of Comput. Vision*, vol. 27, no. 2, 1998.