Magnetic-Visual Sensor Fusion-based Dense 3D Reconstruction and Localization for Endoscopic Capsule Robots

Mehmet Turan[1], Yasin Almalioglu[2], Evin Pinar Ornek[3], Helder Araujo[4], Mehmet Fatih Yanik [5], and Metin Sitti[6]

***Abstract*— Reliable and real-time 3D reconstruction and localization functionality is a crucial prerequisite for the navigation of actively controlled capsule endoscopic robots as an emerging, minimally invasive diagnostic and therapeutic technology for use in the gastrointestinal (GI) tract. In this study, we propose a fully dense, non-rigidly deformable, strictly real-time, intraoperative map fusion approach for actively controlled endoscopic capsule robot applications which combines magnetic and vision-based localization, with non-rigid deformations based frame-to-model map fusion. The performance of the proposed method is evaluated using four different ex-vivo porcine stomach models. Across different trajectories of varying speed and complexity, and four different endoscopic cameras, the root mean square surface reconstruction errors vary from 1.58 to 2.17 cm.**

I. Introduction

Gastrointestinal diseases are the primary diagnosis for about 28 million patient visits per year in the United States[1]. In many cases, endoscopy is an effective diagnostic and therapeutic tool, and as a result about 7 million upper and 11.5 million lower endoscopies are carried out each year in the U.S. [2]. Wireless capsule endoscopy (WCE), introduced in 2000 by Given Imaging Ltd., has revolutionized patient care by enabling inspection of regions of the GI tract that are inaccessible with traditional endoscopes, and also by reducing the pain associated with traditional endoscopy [3]. Going beyond passive inspection, researchers are striving to create capsules that perform active locomotion and intervention [4]. With the integration of further functionalities, e.g. remote control, biopsy, and embedded therapeutic modules, WCE can become a key technology for GI diagnosis and treatment in near future.

Several research groups have recently proposed active, remotely controllable robotic capsule endoscope prototypes equipped with additional operational functionalities, such as highly localized drug delivery, biopsy, and other medical functions [5]–[15]. To facilitate effective navigation and intervention, the robot must be accurately localized and must also accurately perceive the surrounding tissues as demonstrated in Fig. 2. Three-dimensional intraoperative SLAM algorithms will therefore be an indispensable component of future active capsule systems. Several localization methods have been proposed for robotic capsule endoscopes such as fluoroscopy [16], ultrasonic imaging [17], positron emission tomography (PET) [16], magnetic resonance imaging (MRI) [16], radio transmitter based techniques, and magnetic field-based techniques [18]. It has been proposed that combinations of sensors, such as RF range estimation and visual odometry, may improve the estimation accuracy [19]. Moreover, solutions that incorporate vision are attractive because a camera is already present on capsule endoscopes, and vision algorithms have been widely applied for robotic localization and map reconstruction.

Feature-based SLAM methods have been applied on endoscopic type of image sequences in the past e.g [6], [8]–[11], [20]–[23]. As improvements to accomodate the flexibility of the GI tract, [24] suggested a motion compensation model to deal with peristaltic motions, whereas [25] proposed a learning algorithm to deal with them. [26] adapted parallel tracking and mapping techniques to a stereo-endoscope to obtain reconstructed 3D maps that were denser when compared to monoscopic camera methods. [27] has applied

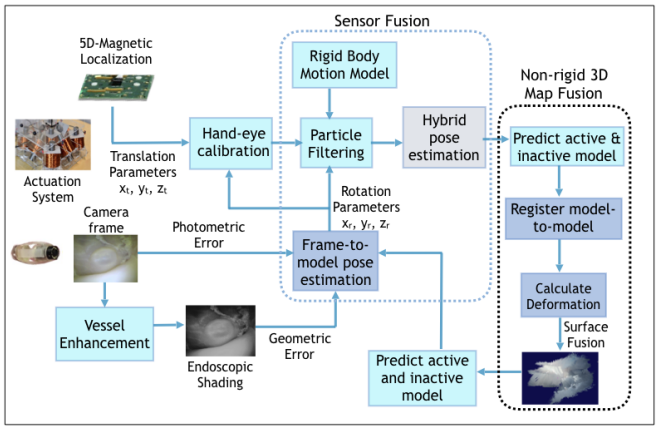


Fig. 1: System overview including 5-DoF magnetic localization, 6-DoF visual joint photometric-geometric frame-to-model pose optimization, inter-sensor calibration, particle filtering based sensor fusion, non-rigid deformations based frame-to-model map fusion.

[1]Mehmet Turan is with the Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Germany turan@is.mpg.de

[2]Yasin Almalioglu is with the Computer Science Department, University of Oxford, Oxford, UK yasin.almalioglu@cs.ox.ac.uk

[3]Evin Pinar Ornek is with the Informatics Department, Technical University of Muenich, Germany evin.oernek@tum.de

[4]Helder Araujo is with the Institute for Systems and Robotics, University of Coimbra, Portugal helder@isr.uc.pt

[5]M. Fatih Yanik is with the Department of Information Technology and Electrical Engineering, Zurich, Switzerland yanik@ethz.ch

[6]Metin Sitti is with the Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Germany sitti@is.mpg.de

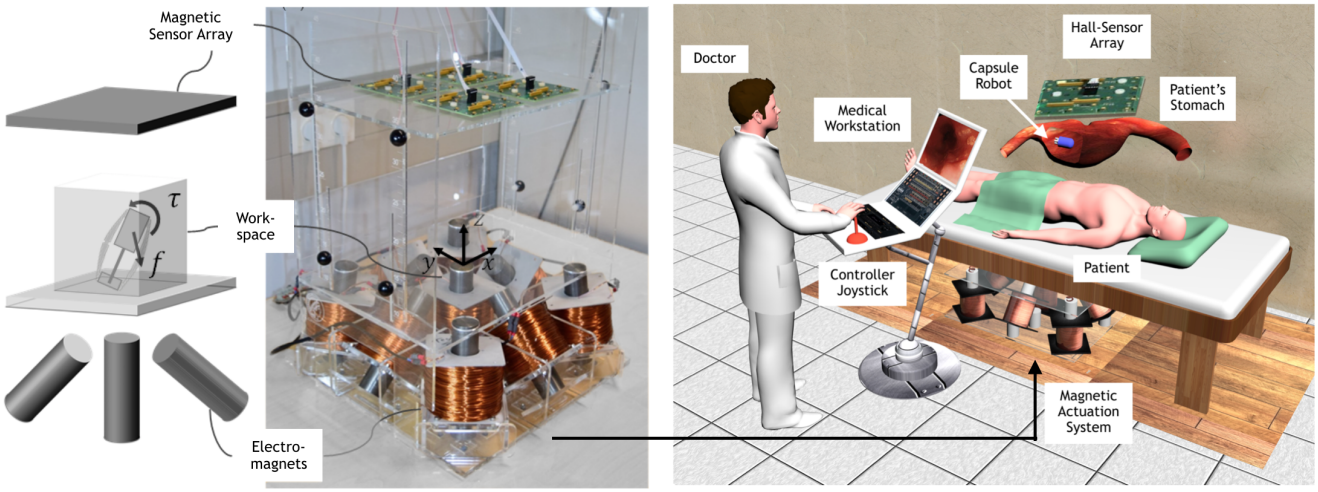


Fig. 2: Demonstration of the active endoscopic capsule robot operation using MASCE (Magnetically actuated soft capsule endoscope) designed for disease detection, drug delivery and biopsy-like operations in the upper GI-tract. MASCE is composed of a RGB camera, a permanent magnet, an empty space for drug chamber and a biopsy tool. Electromagnetic coils based actuation unit below the patient table exerts forces and torques to execute the desired motion. Medician in real-time using the live video stream onto the medical workstation and the controller joystick to manevour the endoscopic capsule to the desired position/orientation.

ORB features to track the camera and proposed a method to densify the reconstructed 3D map, but pose estimation and map reconstruction are still not accurate enough. All of these methods can fail to produce accurate results in cases of low texture areas, motion blur, specular highlights, and sensor noise – all of which are typically present during endoscopy. In this study, we propose a non-rigidly deformable RGB Depth fusion method for endoscopic capsule robot, which combines magnetic localization and visual pose estimation using particle filtering (see Fig. 1). We evaluate the proposed system in terms of surface mapping and capsule localization, using four different ex-vivo porcine stomachs.

II. System Overview and Analysis

The system architecture of the method is depicted in Figure 1. Alternating between localization and mapping, our approach performs frame-to-model 3D map reconstruction in real-time. Below we summarize key steps of the proposed system:

- Perform offline inter-sensor calibration between magnetic hall sensor array and capsule camera system;
- Estimate 3D position of the endoscopic capsule robot using magnetic localization;
- Estimate 3D rotation of the endoscopic capsule robot using visual pose optimization;
- Fuse magnetic position and visual rotation information using particle filtering and 6-DoF rigid body motion model;
- Perform non-rigid frame-to-model map registration making use of hybrid magneto-visual pose information and deformation constraints defined by the graph equations;

III. Method

A. Magnetic Localization System

Our 5-DoF magnetic localization system is designed for the position and orientation estimation of untethered mesoscale magnetic robots [18]. The system uses an external magnetic sensor system and electromagnets for the localization of the magnetic capsule robot. A 2D-Hall-effect sensor array measures the component of the magnetic field from the permanent magnet inside the capsule robot at several locations outside of the robotic workspace. Additionally, a computer-controlled magnetic coil array consisting of nine electromagnets generates the magnetic field for actuation. The core idea of our localization technique is the separation of the capsule's magnetic field component from the actuator's magnetic field component. For that purpose, the actuator's magnetic field is subtracted from the magnetic field data which is acquired by a Hall-effect sensor array. The magnetic localization system estimates a 5-DoF pose, which includes 3D translation and rotation about two axes. (From the magnetic localization information, our system only uses the 3D position parameters and the scale information). Figures 2 and 3 demonstrate the magnetic actuation and localization setup.

B. Visual Localization

1) Multi-scale vessel enhancement and depth image creation: Endoscopic images have mostly homogeneous and poorly textured areas. To prepare the camera frames for input into the ElasticFusion pipeline, our framework starts with a vessel enhancement operation inspired from [29]. Our approach enhances blood vessels by analyzing the multiscale second order local structure of an image. First, we extract

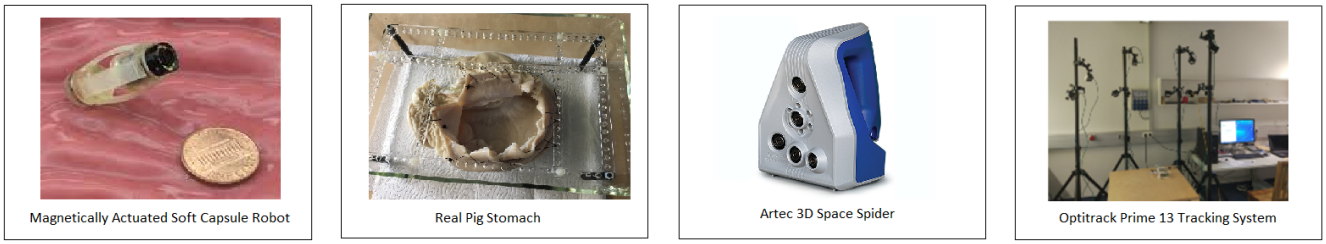


Fig. 3: Illustration of the experimental setup. MASCE is a magnetically actuated robotic capsule endoscope prototype which has a ringmagnet on the body. An electromagnetic coil array consisting of nine coils is used for the actuation of the MASCE. An opened and oiled porcine stomach simulator is used to represent human stomach. Artec 3D scanner is used for ground truth map estimation. OptiTrack system consisting of eight infrared cameras is employed for the ground truth pose estimation.

the Hessian matrix :

$$H = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{bmatrix} \tag{1}$$

where I is the input image, and I_{xx}, I_{xy}, I_{yx}, I_{yy} the second order derivatives, respectively. Secondly, eigenvalues $|\lambda_1| \leq |\lambda_2|$ and principal directions u_1, u_2 of the Hessian matrix are extracted. The eigenvalues and principal directions are then ordered and analyzed to decide whether the region belongs to a vessel. To identify vessels in different scales and sizes, multiple scales are created by convolving the input image and the final output is taken as the maximum of the vessel filtered image across all scales. Figure 4 shows input RGB images, vessel detection and vessel enhancement results for four different frames.

To create depth from input RGB data, we implemented a real-time version of the perspective shape from shading under realistic conditions [30] by reformulating the complex inverse problem into a highly parallelized non-linear optimization problem, which we solve efficiently using GPU programming and a Gauss-Newton solver. Figure 4 shows samples of input RGB images and depth images created from them.

2) Joint photometric-geometric pose estimation: We make use of direct surfel map fusion approach. The core algorithm is inspired by and modified from the ElasticFusion method originally described by Whelan et al. [28], which uses a dense map and non-rigid model deformation to account for changing environments. In parallel to Whelan et al. [28], our approach performs joint volumetric-photometric alignment, frame-to-model predictive tracking, and dense model-to-model loop closure with non-rigid space deformations.

The vision-based localization system operates on the principle of optimizing both relative photometric and geometric pose errors between consecutive frames. The camera pose of the endoscopic capsule robot is described by a transformation matrix $\mathbf{P}_t$:

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ 0_{1\times 3} & 1 \end{bmatrix} \in \mathbb{SE}_3. \tag{2}$$

Given the depth image $\mathscr{D}$, the 3D back-projection of a point $\mathbf{u}$ is defined as $\mathbf{p}(\mathbf{u},\mathscr{D}) = \mathbf{K}^{-1}\boldsymbol{u}d(\mathbf{u})$, where $\mathbf{K}$ is the camera intrinsics matrix and $\boldsymbol{u}$ is the homogeneous form of $\mathbf{u}$. Geometric pose estimation is performed by minimizing the energy cost function E_{icp} between the current depth frame, $\mathscr{D}_t^l$, and the active depth model, $\hat{\mathscr{D}}_{t-1}^a$:

$$E_{icp} = \sum_k ((\mathbf{v}^k - \exp(\hat{\xi})\mathbf{T}\mathbf{v}_k^t)\cdot\mathbf{n}^k)^2 \tag{3}$$

where $\mathbf{v}_t^k$ is the back-projection of the k-th vertex in $\mathscr{D}_t^l$, $\mathbf{v}^k$ and $\mathbf{n}^k$ are the corresponding vertex and normal from the previous frame. $\mathbf{T}$ is the estimated transformation from the previous to the current robot pose and $\exp(\hat{\xi})$ is the exponential mapping function from Lie algebra $\mathfrak{se}_3$ to Lie group $\mathbb{SE}_3$, which represents small changes The photometric pose ξ between the current surfel-based reconstructed RGB image $\mathscr{C}_t^l$ and the active RGB model $\hat{\mathscr{C}}_{t-1}^a$ is determined by minimizing the photometric energy cost function:

$$E_{rgb} = \sum_{\mathbf{u}\in\Omega} \left(I(\mathbf{u},\mathscr{C}_t^l) - I(\pi(\mathbf{K}\exp(\hat{\xi})\mathbf{T}\mathbf{p}(\mathbf{u},\mathscr{D}_t^l)),\hat{\mathscr{C}}_{t-1}^a) \right)^2 \tag{4}$$

where as above $\mathbf{T}$ is the estimated transformation from previous to the current camera pose.

The joint photometric-geometric pose optimization is defined by the cost function:

$$E_{\text{track}} = E_{\text{icp}} + w_{\text{rgb}}E_{\text{rgb}}, \tag{5}$$

with $w_{\text{rgb}} = 0.13$, which was determined experimentally for our datasets. For the minimization of this cost function in real-time, the Gauss-Newton method is employed. At each iteration of the method, the transformation $\mathbf{T}$ is updated as $\mathbf{T} \rightarrow \exp(\hat{\xi})\mathbf{T}$.

C. Relative pose of magnetic and visual localization systems

To relate the magnetic actuation and localization system with the proposed vision system, the relative pose has to be estimated, which can be done using rigid motion model and the constraint of the rigid transformation between the magnetic localization system, magnetic actuation system and the camera. To estimate the relative pose, we assumed a value for the missing rotational DoF in the magnetic sensor data and used an approach based on the method described in [32]. Several motions were performed, and using the estimates of the relative pose (between consecutive positions), the rigid transformation between the two coordinate systems was

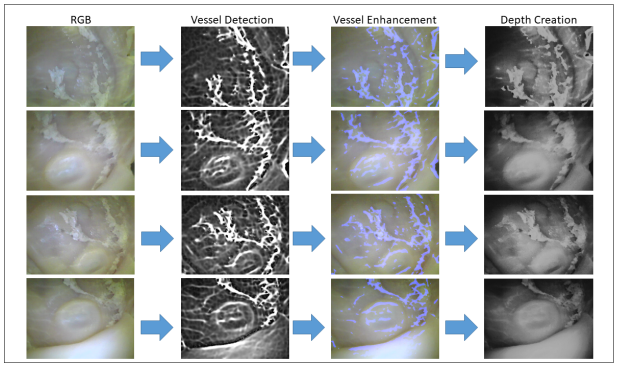


Fig. 4: For a given RGB frame, we extract the Hessian matrix and derive its eigenvalues and principal directions to detect the vessel. We convolve the input frame and final output to create multiple scale representations to identify the different vessels. After enhancement of vessel detected frame, we use shape from shading to create depth map. Qualitative results for sample frames are illustrated in the figure. Here, the dataset of our samples are collected in our experimental setup from an ex-vivo real pig stomach.

estimated. The use of several motions allowed the estimation of the uncertainty in the parameters.

D. Particle Filtering based Magneto-Visual Sensor Fusion

We developed a particle filtering based sensor fusion method for endoscopic capsule robots which is inspired by and modifed from [31]. As motion model, we use a rigid body motion model (3D rotation and 3D translation) assuming a constant velocity which is fairly obeyed during incremental motions of magnetically actuated endoscopic capsule robots. The proposed fusion approach estimates the 3D translation using the measurements from the magnetic sensor, including the scale factor, and the 3D rotation using visual localization.

The state $\mathbf{x_t}$ composes the 6-DoF pose for the capsule robot, which is assumed to propagate in time according to a transition model:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{v}_t) \tag{6}$$

where f is a non-linear state transition function and $\mathbf{v_t}$ is white noise. t is the index of a time sequence, $t \in \{1,2,3,...\}$. Observations of the pose are produced by n sensors $\mathbf{z}_{k,t}(k = 1,...,n)$ in general, where the probability distribution $p(\mathbf{z}_{k,t}|\mathbf{x}_t)$ is known for each sensor. We estimate the 6-DoF pose states relying on latent (hidden) variables by using the Bayesian filtering approach. The hidden variables of sensor states are denoted as $s_{k,t}$, which we call switch variables, where $s_{k,t} \in \{0,...,d_k\}$ for $k = 1,...,n$. d_k is the number of possible observation models, e.g., failure and nominal sensor states. The observation model for $\mathbf{z}_{k,t}$ can be described as:

$$\mathbf{z}_{k,t} = h_{k,s_{k,t},t}(\mathbf{x}_t) + \mathbf{w}_{k,s_{k,t},t} \tag{7}$$

where $h_{k,s_{k,t},t}(\mathbf{x}_t)$ is the non-linear observation function and $\mathbf{w}_{k,s_{k,t},t}$ is the observation noise. The prior probability for the switch parameter $s_{k,t}$ being in a given state j, is denoted as $\alpha_{k,j,t}$ and it is the probability for each sensor to be in a given state:

$$Pr(s_{k,t} = j) = \alpha_{k,j,t}, \quad 0 \leq j \leq d_k \tag{8}$$

where $\alpha_{k,j,t} \geq 0$ and $\sum_{j=0}^{d_k} \alpha_{k,j,t} = 1$ with a Markov evolution property. The objective posterior density function $p(\mathbf{x}_{0:t}, \mathbf{s}_{1:t}, \alpha_{0:t}|\mathbf{z}_{1:t})$ and the marginal posterior probability $p(\mathbf{x}_t|\mathbf{z}_{1:t})$, in general, cannot be determined in a closed form due to its complex shape. However, sequential Monte Carlo methods (*particle filters*) provide a numerical approximation of the posterior density function with a set of samples (*particles*) weighted by the kinematics and observation models.

E. Scene Reconstruction

For scene reconstruction, we use surfels. Each surfel has a position, normal, color, weight, radius, initialization timestamp and last updated timestamp. We also define a deformation graph consisting of a set of nodes and edges to detect non-rigid deformations throughout the frame sequence. Each node $\mathscr{G}^n$ has a timestamp $\mathscr{G}^n_{t_0}$, a position $\mathscr{G}^n_g \in \mathbb{R}^3$ and a set of neighboring nodes $\mathscr{N}(\mathscr{G}^n)$. The directed edges of the graph are neighbors of each node. A graph is connected up to a neighbor count k such that $\forall n, |\mathscr{N}(\mathscr{G}^n)| = k$. Each node also stores an affine transformation in the form of a 3×3 matrix $\mathscr{G}^n_{\mathbf{R}}$ and a 3×1 vector $\mathscr{G}^n_{\mathbf{t}}$. When deforming a surface, the $\mathscr{G}^n_{\mathbf{R}}$ and $\mathscr{G}^n_{\mathbf{t}}$ parameters of each node are optimized according to surface constraints. In order to apply a deformation graph to the surface, each surfel $\mathscr{M}^s$ identifies a set of influencing nodes in the graph $\mathscr{I}(\mathscr{M}^s, \mathscr{G})$. The deformed position of a surfel is given by:

$$\hat{\mathscr{M}}^s_{\mathbf{p}} = \phi(\mathscr{M}^s) = \sum_{n \in \mathscr{I}(\mathscr{M}^s, \mathscr{G})} w^n(\mathscr{M}^s)[\mathscr{G}^n_{\mathbf{R}}(\mathscr{M}^s_{\mathbf{p}} - \mathscr{G}^n_{\mathbf{g}}) + \mathscr{G}^n_{\mathbf{g}} + \mathscr{G}^n_{\mathbf{t}}] \tag{9}$$

while the deformed normal of a surfel is given by:

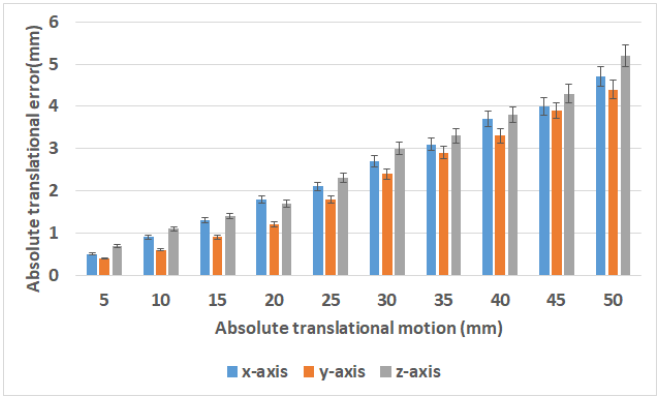
$$\hat{\mathscr{M}}^s_{\mathbf{p}} = \sum_{n \in \mathscr{I}(\mathscr{M}^s, \mathscr{G})} w^n(\mathscr{M}^s) \mathscr{G}^{n-1^T}_{\mathbf{R}} \mathscr{M}^s_{\mathbf{n}}, \tag{10}$$

where $w^n(\mathscr{M}^s)$ is a scalar representing the influence of $\mathscr{G}^n$ on surfel $\mathscr{M}^s$, summing to a total of 1 when $n = k$:

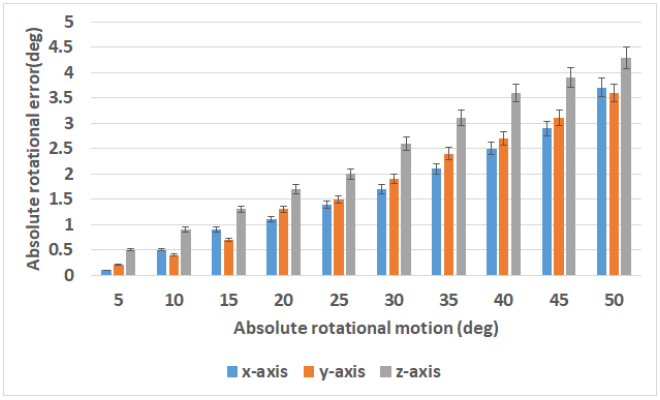
$$w^n(\mathscr{M}^s) = (1 - ||\mathscr{M}^s_{\mathbf{p}} - \mathscr{G}^n_{\mathbf{g}}||_2 / d_{\max})^2. \tag{11}$$

Here, $d_{\max}$ is the Euclidean distance to the $k+1$-nearest node of M^s.

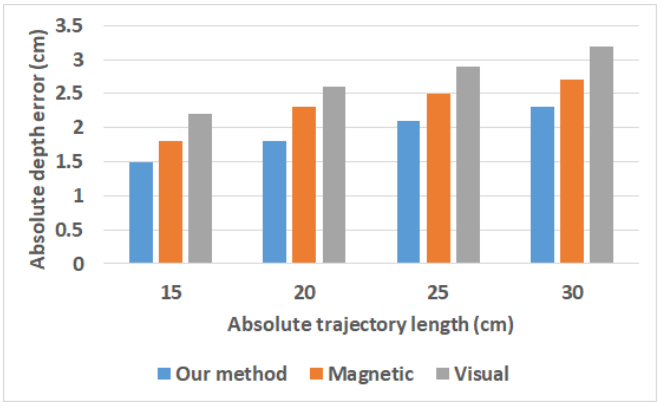
To ensure a globally consistent surface reconstruction, the framework closes loops with the existing map as those areas are revisited. This loop closure is performed by fusing reactivated parts of the inactive model into the active model and simultaneously deactivating surfels which have not appeared for a period of time.



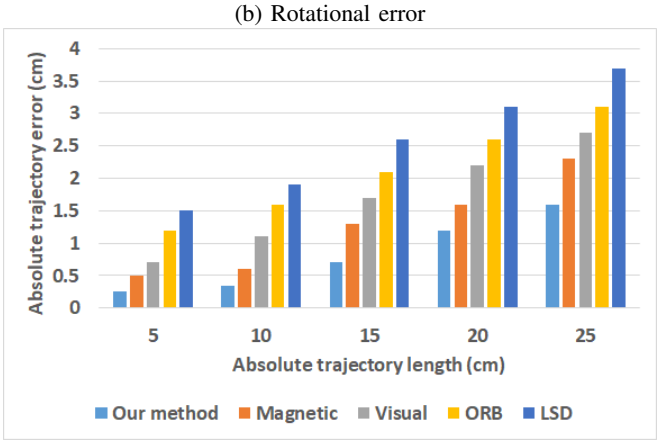
(a) Translational error



(b) Rotational error



(c) Depth error



(d) Trajectory error

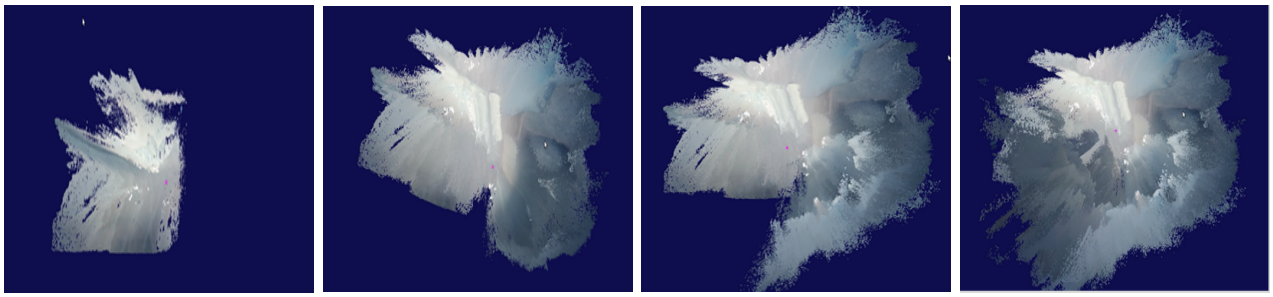
Fig. 5: Figure (a) and Figure (b) demonstrates translational and rotational errors of x, y, z axes for our proposed method. The translational motion of 5 mm results in around 0.5 mm drift on average for x,y,z, whereas a 5 degree rotational motion results in 0.5 degree error maximum. The absolute depth error results for magnetic localization, visual localization and our method is illustrated in (c). It can be observed that our method outperforms the others in depth estimation for different trajectory lengths. In (d), we compare the trajectory errors of magnetic localization, visual localization, ORB SLAM, LSD SLAM and our method. For each of different trajectory lengths, our method outperforms the localization methods that use only visual or magnetic sensors and SLAM methods. For example, in a trajectory with 20 cm, our method estimates with a 1.25 cm error, whereas the error of magnetic localization is 1.6, visual localization is 2.1, ORB SLAM is 2.6, and LSD SLAM is 3.

IV. EXPERIMENTS AND RESULTS

We evaluate the performance of our system both quantitatively and qualitatively in terms of surface reconstruction, trajectory estimation and computational performance. Figure 3 illustrates our experimental setup. Four different endoscopic cameras were used to capture endoscopic capsule videos which were mounted on our magnetically activated soft capsule endoscope (MASCE) systems. The dataset was recorded on four different open non-rigid porcine stomach. Ground truth 3D reconstructions of stomachs were acquired by scanning with a high-quality 3D scanner Artec Space Spider. These 3D scans served as the gold standard for the evaluations of the 3D map reconstruction. To obtain the ground truth for 6-DoF camera pose, an OptiTrack motion tracking system consisting of eight infrared cameras was utilized. A total of 15 minutes of stomach videos were recorded containing over $10K$ frames. Some sample frames of the dataset are shown in Fig. 4 for visual reference.

A. *Surface reconstruction and trajectory estimation*

We used the map benchmarking technique proposed by [33] for the evaluation of the map reconstruction and ATE [34] for trajectory comparisons. Since iterative closest point algorithm (ICP) is a non-convex procedure highly dependent on a good initialization, we first manually align reference and estimated point cloud by picking six corresponding point pairs between both point clouds. Using these six manually picked corresponding point pairs, the transformation matrix is estimated which minimizes square sum difference between aligned and reference cloud. As a next step, ICP is applied between manually aligned cloud pair to fine-tune the alignment. The termination criteria for ICP iterations is an RMSE difference of 0.001 cm between consecutive iterations. We use Euclidean distances between aligned and reference cloud points to calculate the RMSE for depth. Surface reconstruc-



(a) 10 frames (b) 100 frames (c) 300 frames (d) 500 frames

Fig. 6: Reconstructed 3D map of a porcine non-rigid stomach simulator for total number of 10, 100, 300 and 500 frames, respectively. The illustrations are complementary to surface reconstruction errors given in Fig. 5d. It is observable that the proposed method reconstructs 3D organ surface precisely.

tion errors are compared with the magnetic localization-based and visual localization-based surface reconstruction errors in Fig. 5c. Results indicate that the proposed method reconstructs 3D organ surface very precisely outperforming both methods. Table I shows the reconstruction error metrics for full trajectory lengths and four different porcine stomachs including mean, median, standard deviation, minimum and maximum error. Sample 3D reconstructed maps for different lengths of frame sequences (10, 100, 300, 500 frames) are shown in Fig. 6, for visual reference.

Figures 5a and 5b demonstrate absolute translational and rotational errors for our method, magnetic sensor-based localization and vision-based localization. Observation shows that proposed hybrid approach outperforms both sensor types clearly in terms of translational and rotational motion estimation. A translational motion of 5 mm results in a drift of around 0.5 mm on average for x,y,z axes, whereas a 5 degree rotational motion results in a maximum error of 0.5 degree. Figure 5 shows the absolute trajectory errors acquired by our method, compared to ORB SLAM [34], LSD SLAM [35], magnetic sensor-based and visual sensor-based localization. Results again indicate, that the proposed hybrid method outperforms other methods. For example, in a trajectory of 20 cm length, our method estimates with an error of 1.25 cm, whereas magnetic localization, visual localization, ORB and LSD SLAM estimate with an error of 1.6 cm, 2.1 cm, 2.6 cm, and 3 cm, respectively.

TABLE I: Reconstruction results for different stomach sequences.

Error (cm)	St0	St1	St2	St3
Mean	1.81	1.97	1.58	2.17
Median	1.69	1.55	1.38	1.98
Std.	1.94	2.67	1.73	2.32
Min	0.00	0.00	0.00	0.00
Max	3.4	4.2	3.1	4.5

B. Computational Performance

To analyze the computational performance of the system, we observed the average frame processing time across the videos. The test platform was a desktop PC with an Intel Xeon E5-1660v3-CPU at 3.00 GHz, 8 cores, 32GB of RAM and an NVIDIA Quadro K1200 GPU with 4GB of memory. The execution time of the system is depended on the number of surfels in the map, with an overall average of 45 ms per frame scaling to a peak average of 52 ms implying a worst case processing frequency of 19 Hz.

V. CONCLUSION

In this paper, we have presented a magnetic-RGB Depth fusion based 3D reconstruction and localization method for endoscopic capsule robots. Our system makes use of dense reconstruction in combination with particle filter based fusion of magnetic and visual localization. The proposed system is able to produce a highly accurate 3D map of the explored inner organ tissue and is able to stay close to the ground truth endoscopic capsule robot trajectory even for challenging robot trajectories. Even though the proposed system shows surface reconstruction errors varying between 1.58 to 2.17 cm, critical medical operations such as biopsy, drug delivery etc. require sub-millimeter precision which was not achievable with our framework. To reach sub-millimeter precisions and increase the robustness of the system, we intend to extend our work into stereo capsule endoscopy. Moreover, we plan to perform *in vivo* testings to validate the accuracy and robustness of the proposed approach in real medical conditions.

REFERENCES

[1] National Center for Health Statistics, "National ambulatory medical care survey: 2014 state and national summary tables," U.S. Centers for Disease Control and Prevension.

[2] A. F. Peery, E. S. Dellon, J. Lund, S. D. Crockett, C. E. McGowan, W. J. Bulsiewicz, L. M. Gangarosa, M. T. Thiny, K. Stizenberg, D. R. Morgan, *et al.*, "Burden of gastrointestinal disease in the united states: 2012 update," *Gastroenterology*, vol. 143, no. 5, pp. 1179–1187, 2012.

[3] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–418, 2000.

[4] A. Moglia, A. Menciassi, M. O. Schurr, and P. Dario, "Wireless capsule endoscopy: from diagnostic devices to multipurpose robotic systems," *Biomedical microdevices*, vol. 9, no. 2, pp. 235–243, 2007.

[5] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, and E. Diller, "Biomedical applications of untethered mobile milli/microrobots," *Proceedings of the IEEE*, vol. 103, no. 2, pp. 205–224, 2015.

[6] M. Turan, Y. Almalioglu, H. Gilbert, A. E. Sari, U. Soylu, and M. Sitti, "Endo-vmfusenet: Deep visual-magnetic sensor fusion approach for uncalibrated, unsynchronized and asymmetric endoscopic capsule robot localization data," *CoRR*, vol. abs/1709.06041, 2017. [Online]. Available: http://arxiv.org/abs/1709.06041

[7] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots," *Neurocomputing*, vol. 275, pp. 1861 – 1870, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121731665X

[8] M. Turan, Y. Almalioglu, H. Gilbert, H. Araújo, T. Cemgil, and M. Sitti, "Endosensorfusion: Particle filtering-based multi-sensory data fusion with switching state-space model for endoscopic capsule robots," *CoRR*, vol. abs/1709.03401, 2017. [Online]. Available: http://arxiv.org/abs/1709.03401

[9] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "A non-rigid map fusion-based direct slam method for endoscopic capsule robots," *International Journal of Intelligent Robotics and Applications*, vol. 1, no. 4, pp. 399–409, Dec 2017. [Online]. Available: https://doi.org/10.1007/s41315-017-0036-4

[10] M. Turan, Y. Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, and M. Sitti, "Sparse-then-dense alignment-based 3d map reconstruction method for endoscopic capsule robots," *Machine Vision and Applications*, vol. 29, no. 2, pp. 345–359, Feb 2018. [Online]. Available: https://doi.org/10.1007/s00138-017-0905-8

[11] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots," *arXiv preprint arXiv:1708.06822*, 2017.

[12] M. Turan, Y. Almalioglu, E. Konukoglu, and M. Sitti, "A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots," *CoRR*, vol. abs/1705.05435, 2017. [Online]. Available: http://arxiv.org/abs/1705.05435

[13] M. Turan, Y. Y. Pilavci, R. Jamiruddin, H. Araújo, E. Konukoglu, and M. Sitti, "A fully dense and globally consistent 3d map reconstruction approach for GI tract to enhance therapeutic relevance of the endoscopic capsule robot," *CoRR*, vol. abs/1705.06524, 2017. [Online]. Available: http://arxiv.org/abs/1705.06524

[14] M. Turan, A. Abdullah, R. Jamiruddin, H. Araújo, E. Konukoglu, and M. Sitti, "Six degree-of-freedom localization of endoscopic capsule robots using recurrent neural networks embedded into a convolutional neural network," *CoRR*, vol. abs/1705.06196, 2017. [Online]. Available: http://arxiv.org/abs/1705.06196

[15] M. Turan, Y. Almalioglu, H. Araújo, E. Konukoglu, and M. Sitti, "A non-rigid map fusion-based rgb-depth SLAM method for endoscopic capsule robots," *CoRR*, vol. abs/1705.05444, 2017. [Online]. Available: http://arxiv.org/abs/1705.05444

[16] T. D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2387–2399, 2012.

[17] S. Yim and M. Sitti, "3-d localization method for a magnetically actuated soft capsule endoscope and its applications," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1139–1151, 2013.

[18] D. Son, S. Yim, and M. Sitti, "A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 708–716, 2016.

[19] Y. Geng and K. Pahlavan, "On the accuracy of rf and image processing based hybrid localization for wireless capsule endoscopy," in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*, 2015, pp. 452–457.

[20] P. Mountney and G.-Z. Yang, "Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 1184–1187.

[21] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE transactions on medical imaging*, vol. 33, no. 1, pp. 135–146, 2014.

[22] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 275–282.

[23] L. Liu, C. Hu, W. Cai, and M. Q.-H. Meng, "Capsule endoscope localization based on computer vision technique," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 3711–3714.

[24] P. Mountney and G.-Z. Yang, "Motion compensated slam for image guided surgery," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pp. 496–504, 2010.

[25] P. Mountney, D. Stoyanov, A. Davison, and G.-Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2006, pp. 347–354.

[26] B. Lin, A. Johnson, X. Qian, J. Sanchez, and Y. Sun, "Simultaneous tracking, 3d reconstruction and deforming point detection for stereoscope guided surgery," in *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*. Springer, 2013, pp. 35–44.

[27] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. Montiel, "Orbslam-based endoscope tracking and 3d reconstruction," *arXiv preprint arXiv:1608.08149*, 2016.

[28] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "Elasticfusion: Real-time dense slam and light source estimation," *The International Journal of Robotics Research*, pp. 1697–1716, 2016.

[29] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 1998, pp. 130–137.

[30] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Metric depth recovery from monocular images using shape-from-shading and specularities," *IEEE International Conference on Image Processing (ICIP)*, 2012.

[31] F. Caron, M. Davy, E. Duflos, and P. Vanheeghe, "Particle filtering for multisensor data fusion with switching observation models: Application to land vehicle positioning," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2703–2719, 2007.

[32] P. Lébraly, E. Royer, O. Ait-Aider, and M. Dhome, "Calibration of non-overlapping cameras - application to vision-based robotics," in *Proc. BMVC*, 2010, pp. 10.1–12, doi:10.5244/C.24.10.

[33] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *Robotics and automation (ICRA), 2014 IEEE international conference on*. IEEE, 2014, pp. 1524–1531.

[34] R. Mur-Artal, J. Montiel, and J. D. Tardós, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[35] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.