DISS. ETH NO. 25108

Advanced Localization and Mapping Techniques for Endoscopic Capsule Robots

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MEHMET TURAN

Dipl. Eng., RWTH Aachen

born on 21 August 1985

citizen of Turkey

accepted on the recommendation of

Prof. Dr. Mehmet Fatih Yanik, examiner
Prof. Dr. Metin Sitti, co-examiner
Prof. Dr. Michael Black, co-examiner
Assistant Professor Mahmut Selman Sakar, co-examiner

2018

To my beloved wife Kübra, my lovely son, and my parents

Acknowledgement

Firstly, I would like to express my sincere gratitude to my Max Planck Institute advisor Prof. Dr. Metin Sitti and my academic ETH advisor Prof. Dr. M. Fatih Yanik for their continuous support of my Ph.D study, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. Besides my advisors, I would like to thank the rest of my thesis committee: Prof. Dr. Benjamin Grewe from ETH Zurich, Prof. Dr. Michael Black from Max Planck Institute for Intelligent Systems and Prof. Dr. Mahmut Selman Sakar from École Polytechnique Fédérale de Lausanne. Besides them, I would like to thank Prof. Dr. Helder Araujo from Coimbra University, Dr.Hunter Gilbert from Louisiana State University, Prof. Dr. Ender Konukoglu from ETH Zurich, my collaborators Yasin Almalioglu (M.Sc) from Oxford University, Donghoon Son (M.Sc) from Max Planck Institute for Intelligent Systems their continuous, patient discussions we have had in the last four years. I would also like to thank my interns Ali Eren Sari, Ufuk Soylu and Ipek Ganiyusufoglu for their support. I would like to thank Max Planck Society and Max Planck-ETH Center for Learning Systems for funding my researchLast but not the least, I would like to especially thank my wife Kübra who always supported and motivated me during my PhD study, and my parents for their unlimited support to me in my whole life in general.

Abstract

Endoscopic capsule robots are an emerging and exciting non-invasive medical device technology for comfortable inspection of the gastrointestinal tract organs, enabling additional therapeutic operations such as biopsy, targeted drug delivery, and surgical treatment. Unlike current passive capsule endoscopes used in hospitals, endoscopic capsule robots are actively steerable medical devices, which allow access to body regions, which were impossible to reach with standard hand-held endoscopes before. Biopsy, targeted drug delivery, and surgical treatment require fast and reliable feedback about robotic position and map representation, ideally with submillimeter precision. Thus, accurate and robust real-time localization and mapping are of significant importance for actively steerable endoscopic capsule robots. This dissertation mainly focuses on novel three-dimensional mapping and localization techniques for endoscopic capsule robots, which among other approaches, make heavy use of computer vision, deep learning and sensor fusion techniques. Unlike static hand-engineered algorithms existing in literature, the presented methods in this dissertation allow the medical device system to dynamically continue learning via streamed data from successive procedures and to adapt to the environmental variations among different patient organs using transfer learning and re-tuning techniques. Detailed quantitative and qualitative evaluations performed on oiled, non-rigid porcine stomachs and realistic soft surgical EsophagoGastroDuodenoscopy simulator show that proposed frameworks outperform state-of-the-art localization and mapping methods for both hand-held and capsule endoscopes.

Abstract

Anders als die aktuell in Krankenhäusern verwendeten passiven Kapsel-Endoskope sind endoskopische Kapsel-Roboter aktiv steuerbare Geräte, die es ermöglichen Körperregionen zu erreichen, die bisher mit den standardmäßigen Handheld-Endoskopen nicht erreicht werden konnten. Biopsie, gezielte Wirkstoffzufuhr und chirurgische Behandlungen verlangen schnelle und zuverlässige Angaben über die Position des Roboters sowie die dazu gehörige Kartendarstellungen (Mapping), am besten mit einer Sub-Millimeter-Genauigkeit. Daher sind eine akkurate und solide Echtzeit-Lokalisierung und –Mapping von entscheidender Bedeutung für den Einsatz von Kapsel-Robotern in der Endoskopie. Diese Dissertation behandelt im wesentlichen neue dreidimensionale Mapping- und Lokalisierungstechniken für endoskopische Kap- sel-Roboter, die neben anderen Ansätzen vor allem Techniken wie Computer Vision, Deep Learning und Sensor-Fusion einsetzen. Anders als die statischen, handprogrammierten Algorithmen in der Literatur, erlauben die Methoden in dieser Dissertation dem System dynamisch weiter zu lernen durch gestreamte Daten von sukzessiven Vorgängen und sich an die variierende Organumgebung bei verschiedenen Patienten anzupassen. Detaillierte quantitative und qualitative Experimenten und Analysen an eingeölten echte Schweinemägen und einem realistischer Softsimulator zeigen, dass die entwickelten Methoden die bisherigen Lokalisierungs- und Mapping-Methoden inder Literatur sowohl für Handheld-Endoskope als auch für Kapsel-Endoskope ubertreffen.

Preface

This cummulative dissertation consists of following papers:

List of papers:

1. **Mehmet Turan**, Yasin Almalioglu, Helder Araujo, Ender Konukoglu, and Metin Sitti. *A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots*, **International Journal of Intelligent Robotics and Applicazations**, December 2017, Volume 1, Issue 4, pp 399–409.

2. **Mehmet Turan**, Yasin Almalioglu, Evin Pinar Ornek, Helder Araujo, Mehmet Fatih Yanik, and Metin Sitti. *Magnetic-Visual Sensor Fusion-based Dense 3D Reconstruction and Localization for Endoscopic Capsule Robots*, **Submitted to IROS 2018.**

3. **Mehmet Turan**, Yasin Almalioglu, Helder Araujo, Ender Konukoglu, and Metin Sitti. *Deep EndoVO: A Recurrent Convolutional Neural Network (RCNN) based Visual Odometry Approach for Endoscopic Capsule Robots*, **Neurocomputing**, Volume 275, 31 January 2018, Pages 1861-1870.

4. **Mehmet Turan**, Nail Ibrahimli, Evin Pinar Ornek, Can Giracoglu, Yasin Almalioglu, Mehmet Fatih Yanik, and Metin Sitti. *Unsupervised Odometry and Depth Learning for Endoscopic Capsule Robots*, **Submitted to IROS 2018.**

5. **Mehmet Turan**, Yasin Almalioglu, Hunter Gilbert, Alp Eren Sari, Ufuk Soylu, and Metin Sitti. *Endo-VMFuseNet: Deep Visual-Magnetic Sensor Fusion Approach for Uncalibrated, Unsynchronized and Asymmetric Endoscopic Capsule Robot Localization Data*, **accepted for IEEE ICRA 2018.**

6. **Mehmet Turan**, Yasin Almalioglu, Hunter Gilbert, Helder Araujo, Taylan Cemgil, Metin Sitti. *EndoSensorFusion: Particle Filtering-Based Multisensory Data Fusion with Switching State-Space Model for Endoscopic Capsule Robots*, **accepted for IEEE ICRA 2018.**

7. **Mehmet Turan**, Yusuf Yigit Pilavci, Ipek Ganiyusufoglu, Helder Araujo, Ender Konukoglu, Metin Sitti. *Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots*, **Machine Vision and Applications**, Volume 29, Issue 2, pp 345–359.

8. Metin Sitti, Hakan Ceylan, Wenqi Hu, Joshua Giltinan, **Mehmet Turan**, Sehyuk Yim, Eric Diller. *Biomedical applications of untethered mobile milli- and microrobots*, **Proceedings of the IEEE**, February 2015, Volume 103, Issue 2, pp 205-224.

Introduction

Untethered medical robots of millimetric scale and below attract growing attention providing unprecedented direct access to the human body which is expected to have a great impact on health care and bioengineering applications in near future. Especially in the past decade, advances in microsensors, microelectronics, computational power and algorithms have enabled miniaturized and low-cost devices in a variety of high impact applications. Following these advances, untethered, pill-size and swallowable capsule endoscopes with on-board cameras and wireless image transmission device have been developed and used in hospitals for screening the gastrointestinal (GI) tract and diagnoses of diseases such as the inflammatory bowel disease, the ulcerative colitis, and the colorectal cancer.

Unlike standard endoscopy, endoscopic capsule robots are non-invasive, painless, and more appropriate to be employed for prolonged screening purposes. Moreover, they can access difficult body parts that were not possible to reach before with standard endoscopy (e.g. small intestines). Such advantages make pill-size capsule endoscopes a significant alternative screening method over standard endoscopy. However, current capsule endoscopes used in hospitals are passive devices locomoted by peristaltic motions of the inner organs. Control over the capsule's position, orientation, and functions would give the doctor a more precise access to targeted body parts and more intuitive (and less error-prone) diagnosis opportunity. Active motion control is, on the other hand, heavily dependent on the precise and reliable real-time pose estimation and mapping capability, which makes robot localization and mapping key functions for a successful endoscopic capsule robot operation. Many challenges posed by the GI tract such as self-repetitive texture, non-rigid organ deformations, random peristaltic motions, viscosity, specularities caused by the organ fluids, lack of distinctive feature points, low accuracy and small size onboard sensors equipped on the capsule robot are challenges to count on. Figure 1 demonstrates an active endoscopic capsule robot operation scenario, where the doctor performs the medical operation in real-time using the medical workstation and a joystick to maneuver the capsule robot. Electromagnetic coils based actuation unit below the patient table receives commands from the controller unit to exert forces and torques on the capsule robot. A 2D-Hall sensor array, placed on top of the patient's stomach, streams out the real-time position and orientation information of the robot.

This cumulative dissertation consists of following papers:

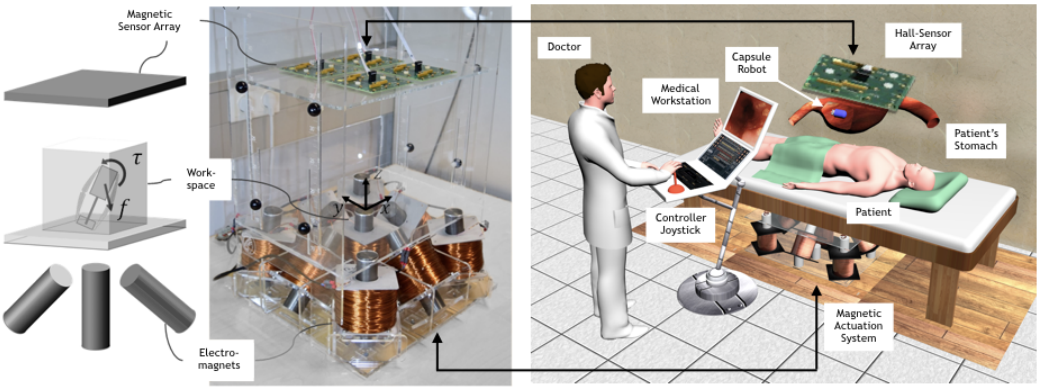


Figure 1: Demonstration of the active endoscopic capsule robot operation.

- A novel medical simultaneous localization and mapping (SLAM) technique which makes use of GPU accelerated non-rigid frame-to-model fusion, joint volumetric-photometric pose estimation and dense model-to-model loop closure techniques. Note that the presented method is only vision-based and does not need any extra sensor;

- A fully dense, non-rigidly deformable, strictly real-time, intraoperative map fusion approach for actively controlled endoscopic capsule robot applications, which combines benefits of magnetic and vision-based localization, with non-rigid deformations based frame-to-model map fusion;

- A supervised deep monocular visual odometry (VO) method for endoscopic capsule robots based on recurrent convolutional neural networks (RCNNs), where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used for feature extraction and inference of dynamics across the frames, respectively;

- An unsupervised deep localization and depth estimation approach for endoscopic capsule robots consisting of two simultaneously trained sub-networks, the first one assigned for depth estimation via encoder-decoder strategy, and the second assigned to regress the camera pose in 6-DoF. The model observes sequences of monocular images and aims to interpret them to estimate the camera motion and depth information in an end-to-end and unsupervised fashion directly from input pixels;

- A sequence-to-sequence deep sensor fusion approach for endoscopic capsule robot localization which has several important novelties and advantages over existing sensor fusion approaches: sensor data does not need to be synchronized, the method is agnostic to sensor type and dimensionality, and the neural network training procedure automatically performs the eye-in-hand calibration for each sensor, including those with reduced (less than 6 dimensional) information;

- A novel multi-sensor fusion algorithm based on switching state space models with particle filtering using the endoscopic capsule robot dynamics modelled

by recurrent neural networks (RNNs), which can handle sensor faults and non-linear motion models;

- A comprehensive medical 3D reconstruction method for endoscopic capsule robots, which is built in a modular fashion including preprocessing, keyframe selection, sparse-then-dense alignment-based pose estimation, bundle fusion, and shading-based 3D reconstruction;
- A comprehensive review of the current advances in biomedical untethered mobile milli- and microrobots with an emphasis on the potential impacts of such devices in the near future and existing and emerging challenges associated with medical operations performed via such minituarized robotic technologies.

International Journal of Intelligent Robotics and Applications (2017) 1:399–409
https://doi.org/10.1007/s41315-017-0036-4

REGULAR PAPER

A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots

Mehmet Turan[1,2] · Yasin Almalioglu[1,2] · Helder Araujo[3] · Ender Konukoglu[2] · Metin Sitti[1]

Received: 13 August 2017 / Accepted: 6 November 2017 / Published online: 24 November 2017

Abstract
Since the development of capsule endoscopy technology, medical device companies and research groups have made significant progress to turn passive capsule endoscopes into robotic active capsule endoscopes. However, the use of robotic capsules in endoscopy still has some challenges. One such challenge is the precise localization of the actively controlled robot in real-time. In this paper, we propose a non-rigid map fusion based direct simultaneous localization and mapping method for endoscopic capsule robots. The proposed method achieves high accuracy for extensive evaluations of pose estimation and map reconstruction performed on a non-rigid, realistic surgical EsophagoGastroDuodenoscopy Simulator and outperforms state-of-the art methods.

Keywords Endoscopic capsule robot · Dense direct medical SLAM · Non-rigid frame-to-model fusion

1 Introduction

In the past decade, advances in microsensors and microelectronics have enabled small, low cost devices in a variety of high impact applications. Following these advances, untethered pill-size, swallowable capsule endoscopes with an onboard camera and wireless image transmission device have been developed and used in hospitals for screening the gastrointestinal (GI) tract and diagnosing diseases such as the inflammatory bowel disease, the ulcerative colitis, and the colorectal cancer. Unlike standard endoscopy, endoscopic capsule robots are non-invasive, painless, and more appropriate to be employed for long-duration screening purposes. Moreover, they can access difficult body parts that were not possible to reach before with standard endoscopy (e.g., small intestines). Such advantages make pill-size capsule endoscopes a significant alternative screening method over standard endoscopy (Liao et al. 2010; Nakamura et al. 2008; Pan and Wang 2012; Than et al. 2012). However, current capsule endoscopes used in hospitals are passive devices controlled by peristaltic motions of the inner organs. The control over capsule's position, orientation, and functions would give the doctor a more precise reachability of targeted body parts and more intuitive and correct diagnosis opportunity. Several groups have recently proposed active, remotely controllable robotic capsule endoscope prototypes equipped with additional functionalities, such as local drug delivery, biopsy, and other medical functions (Sitti et al. 2015; Yim et al. 2013; Carpi et al. 2011; Keller et al. 2012; Mahoney et al. 2013; Yim et al. 2014). An active motion control is, on the other hand, heavily dependent on a precise and reliable real-time pose estimation capability, which makes the robot localization and mapping the key capability for a successful endoscopic capsule robot operation. Localization methods such as (Fluckiger and Nelson 2007; Rubin et al. 2006; Kim et al. 2008; Son et al. 2016) have the common drawback that they require extra sensors and hardware to be integrated to

✉ Mehmet Turan
mturan@student.ethz.ch

Yasin Almalioglu
yasin.almalioglu@boun.edu.tr

Helder Araujo
helder@isr.uc.pt

Ender Konukoglu
ender.konukoglu@vision.ee.ethz.ch

Metin Sitti
sitti@is.mpg.de

[1] Max Planck Institute for Intelligent Systems, Stuttgart, Germany

[2] Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

[3] Robotics Laboratory, University of Coimbra, Coimbra, Portugal

the robotic capsule system. Such extra sensors have their own drawbacks and limitations if it comes to their application in small-scale medical devices, e.g. space limitations, cost aspects, design incompatibilities, biocompatibility issues, and most importantly the interference of the sensors with the activation system of the capsule robot.

As a solution of these issues, vision-based localization and mapping methods (vSLAM) have attracted the attention for small-scale medical devices. With their low cost and small size, cameras are frequently used in localization applications where weight and power consumption are limiting factors, such as in the case of small-scale robots. However, many challenges posed by the GI tract and low quality cameras of the endoscopic capsule robots cause further difficulties in front of a vSLAM technique to be applied in a medical operation. Self-repetitiveness of the GI tract texture, non-rigid organ deformations, heavy reflections caused by the organ fluids, and lack of distinctive feature points on the GI tract tissue are further challenges in front of a reliable robotic operation. Moreover, the low frame rate and limited resolution of the current capsule camera systems also restrict the applicability of computer vision methods inside the GI tract. Especially feature tracking based visual localization methods have poor performance in the abdomen region compared to outdoor or indoor large scale environments where unique features can be found easier.

Figure 1 gives an overview of a modern vSLAM approach with its key components. A modern vSLAM method is expected to be equipped with reliable pose estimation and map reconstruction modules that is not affected by non-rigid deformations, sudden frame-to-frame movements, blur, noise, illumination changes, occlusions and large depth variations. Moreover, dynamic structure of the GI tract organs with heavy peristaltic motions require more than a static map; reconstructed parts of the map must be updated continuously as the organ structure changes during endoscopic operation. Besides, a failure recovery procedure relocalizing the robot after unexpected drifts is a further demand on a modern vSLAM system. The intra-operative 3D reconstruction of the explored inner organ simultaneous to tracking capsule robot position in real-time provides key information for the next generation actively controllable endoscopic robots which will be equipped with functionalities such as disease detection, local drug delivery and biopsy. Feature- based SLAM methods have been applied on endoscopic type of videos in the past decades (Mountney and Yang 2009; Casado et al. 2014; Stoyanov et al. 2010; Mountney and Yang 2010; Mountney et al. 2006; Qian et al. 2013; Mahmoud et al. 2016). However, besides sparse unrealistic map reconstruction, all of these methods suffer from heavy drifts and inaccurate pose estimations once low texture areas are entered. With that motivation, we developed a direct medical vSLAM method which shows high accuracy in terms of map reconstruction and pose estimation inside GI tract.

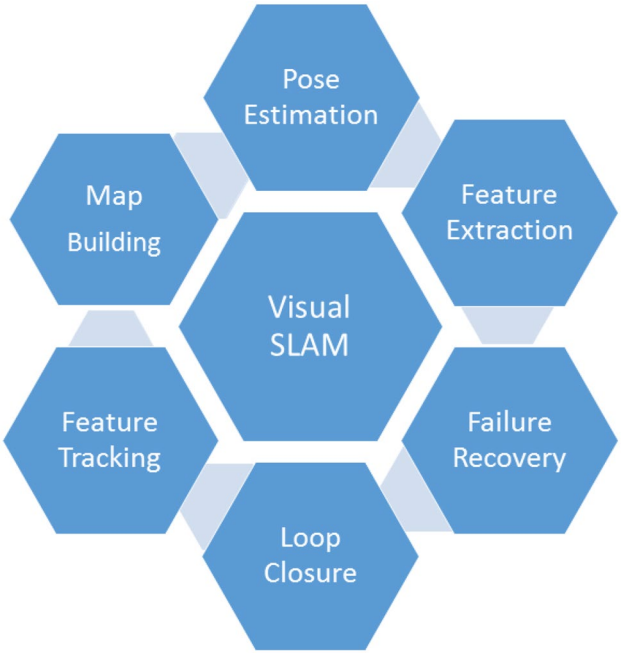


Fig. 1 Components of a modern vSLAM

2 Method

In that section, we first summarize the contributions of our paper and give details of the proposed method.

2.1 Contributions of the method

Inspired from large-scale RGB Depth SLAM approaches (Whelan et al. 2015; Newcombe et al. 2011), the proposed method is to the best of our knowledge the first fully dense, direct medical SLAM approach using GPU accelerated non-rigid frame-to-model fusion, joint volumetric-photometric pose estimation and dense model-to-model loop closure techniques. Figure 2 depicts the system architecture diagram and below the key steps of the proposed framework are summarized:

- Create depth image from RGB image based on shading;
- Divide visited organ parts into active and inactive areas. Only active areas are used for pose tracking and map fusion. Areas that do not appear in the scene for a certain period of time are assigned as inactive and not used in the estimation.
- For every new frame, search for its intersection with the active model and fuse them;

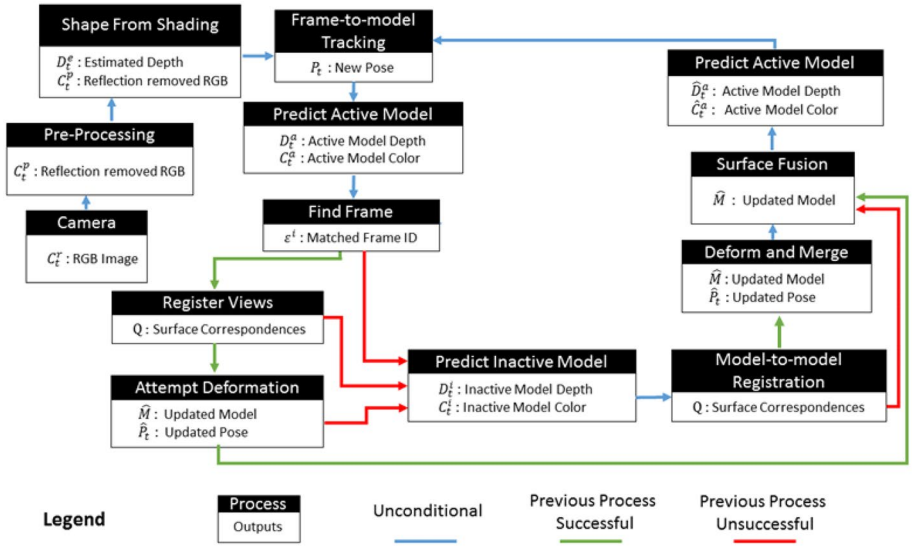


Fig. 2 Overview of the proposed medical SLAM method

- In case there exists an intersection of the active model with inactive model within the current frame, fuse the intersecting parts using loop closure and reactivate corresponding inactive parts.

The contributions of the approach described in this paper include:

- A vSLAM approach able to deal with specularities typically occurring in images of inner organs tissues;
- A direct vSLAM method able to handle non-rigid structures, including performing their non-sparse 3D reconstruction;
- A direct vSLAM approach jointly minimizing photometric-geometric constraints, including depth;

2.2 Preprocessing and depth image creation

The framework starts with a preprocessing module that suppresses specularities caused by inner organ fluids. Reflection detection is done by combining the gradient map of the input image with the peak values detected by an adaptive threshold. Once specularities detected, suppression is performed by inpainting. Next, GPU accelerated version of Tsai-Shah shading method is applied to create depth images. This method uses linear approximations to extract depth image from RGB input iteratively estimating slant, tilt and albedo values. For further details, the reader is referred to the original paper (Ping-Sing and Shah 1994). Figure 3 demonstrates examples of input RGB images, images after reflection suppression and depth images acquired by Tsai-Shah shading method.

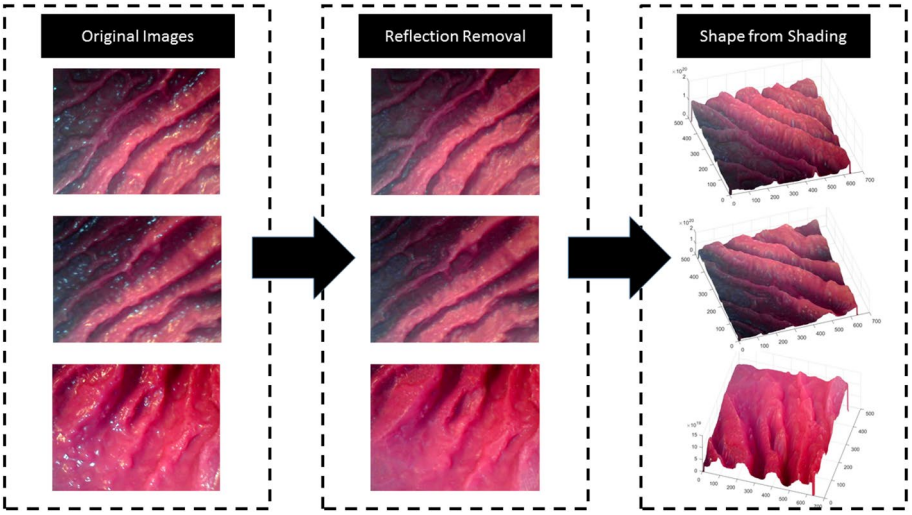


Fig. 3 Reflection suppression and shading-based depth image creation

2.3 Joint photometric and geometric pose estimation from a splattered surfel prediction

The input for pose estimation is the RGB image $\mathcal{C}$ and the depth image $\mathcal{D}$. We combine photometric and geometric pose estimation techniques. The camera pose of the endoscopic capsule robot is described by a transformation matrix $\mathbf{P}_t$:

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

Given the depth image $\mathcal{D}$, the 3D back-projection of a point $\mathbf{u}$ is defined as $\mathbf{p}(\mathbf{u}, \mathcal{D}) = \mathbf{K}^{-1}\boldsymbol{u}\mathrm{d}(\mathbf{u})$, 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 image $\mathcal{D}_t^l$ and the active depth model $\hat{\mathcal{D}}_{t-1}^a$:

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

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

$$E_{rgb} = \sum_{\mathbf{u}\in\boldsymbol{\Omega}} \left(I(\mathbf{u}, \mathcal{C}_t^l) - I(\pi(\mathbf{K}\exp(\hat{\xi})\mathbf{T}\mathbf{p}(\mathbf{u}, \mathcal{D}_t^l)), \hat{\mathcal{C}}_{t-1}^a)\right)^2 \tag{3}$$

The energy minimization function for joint photometric-geometric pose estimation is defined by:

$$E_{\text{track}} = E_{\text{icp}} + w_{\text{rgb}}E_{\text{rgb}}, \tag{4}$$

which is minimized using Gauss–Newton non-linear least-squares optimization.

2.4 Scene representation, deformation graph and loop closure

Due to strict real-time concerns of the approach, we use surfel-based scene reconstruction. 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 $\mathcal{G}^n$ has a timestamp $\mathcal{G}_{t_0}^n$, a position $\mathcal{G}_g^n \in \mathbb{R}^3$ and a set of neighboring nodes $\mathcal{N}(\mathcal{G}^n)$. The directed edges of the graph are neighbors of each node. A graph is connected up to a neighbor count k such that $\forall n, |\mathcal{N}(\mathcal{G}^n)| = k$. Each node also stores an affine transformation in the form of a 3×3 matrix $\mathcal{G}_{\mathbf{R}}^n$ and a 3×1 vector $\mathcal{G}_{\mathbf{t}}^n$. When deforming a surface, the $\mathcal{G}_{\mathbf{R}}^n$ and $\mathcal{G}_{\mathbf{t}}^n$ parameters of each node are optimized according to surface constraints. In order to apply a deformation graph to the surface, each surfel $\mathcal{M}^s$ identifies a set of influencing nodes in the graph $\mathcal{I}(\mathcal{M}^s, \mathcal{G})$. The deformed position of a surfel is given by:

$$\hat{\mathcal{M}}_{\mathbf{p}}^s = \phi(\mathcal{M}^s) = \sum_{n\in\mathcal{I}(\mathcal{M}^s,\mathcal{G})} w^n(\mathcal{M}^s)[\mathcal{G}_{\mathbf{R}}^n(\mathcal{M}_{\mathbf{p}}^s - \mathcal{G}_{\mathbf{g}}^n) + \mathcal{G}_{\mathbf{g}}^n + \mathcal{G}_{\mathbf{t}}^n] \tag{5}$$

while the deformed normal of a surfel is given by:

$$\hat{\mathcal{M}}_{\mathbf{p}}^s = \sum_{n\in\mathcal{I}(\mathcal{M}^s,\mathcal{G})} w^n(\mathcal{M}^s)\mathcal{G}_{\mathbf{R}}^{n-1^T}\mathcal{M}_{\mathbf{n}}^s, \tag{6}$$

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

$$w^n(\mathcal{M}^s) = (1 - ||\mathcal{M}_{\mathbf{p}}^s - \mathcal{G}_{\mathbf{g}}^n||_2/d_{\max})^2. \tag{7}$$

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.

3 Experiments and results

We evaluate the performance of our system both quantitatively and qualitatively in terms of trajectory estimation, surface reconstruction and computational performance.

3.1 Dataset and equipment

Figure 4 shows our experimental setup as a visual reference. We created our own endoscopic capsule robot dataset with ground truth. To make sure that our dataset is general and does not lead to overfitting, three different endoscopic cameras were used to capture the endoscopic videos. We mounted endoscopic cameras on our magnetically activated soft capsule endoscope (MASCE) systems as seen in Fig. 6. The videos were recorded from an oiled non-rigid, surgical stomach model Koken LM103—EDG (EsophagoGastroDuodenoscopy) Simulator. Some sample frames are shown in Fig. 5. To obtain 6-DoF localization ground truth, an OptiTrack motion tracking system consisting of eight infrared cameras and a tracking software was utilized. A total of 15 minutes of stomach videos was recorded containing over 10,000 frames. Finally, we scanned the open surgical stomach model using a 3D Artec Space Spider image scanner.

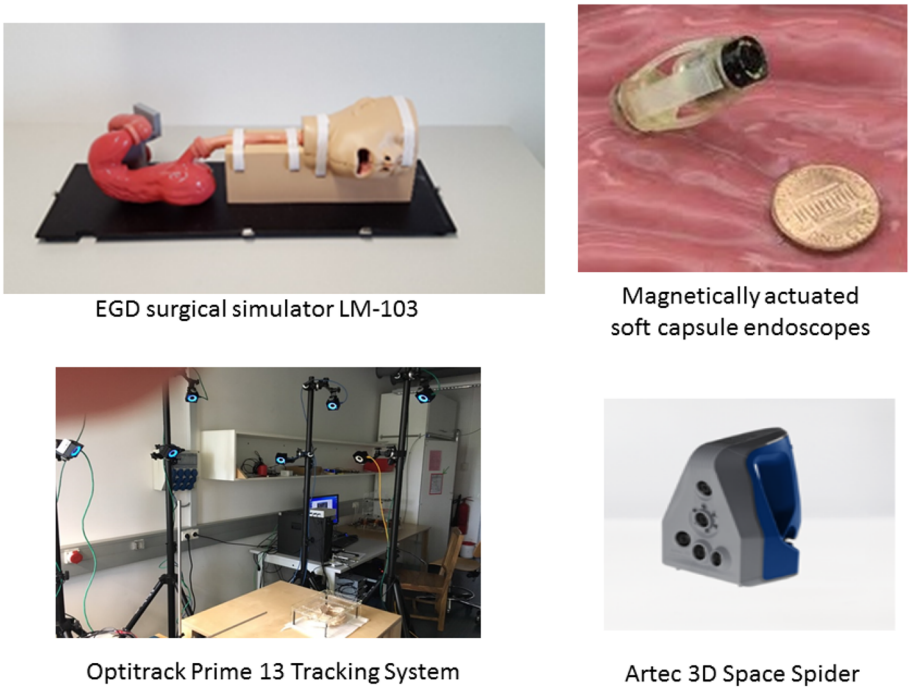


Fig. 4 Experimental setup

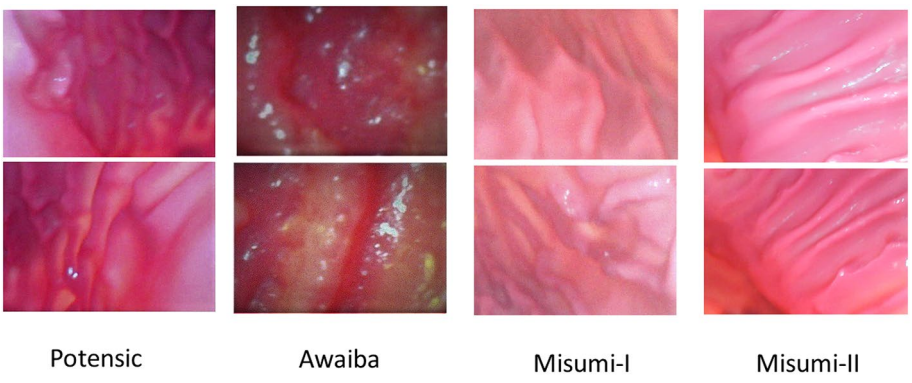


Fig. 5 Sample images from our dataset

This scan served as the ground truth for the quantitative evaluations of the 3D map reconstruction module.

3.2 Trajectory estimation

Table 1 demonstrates the results of the trajectory estimation for 7 different trajectories. The characteristics of the trajectories are as follows:

- Trajectory 1 is an uncomplicated path with slow incremental translations and rotations.
- Trajectory 2 follows a comprehensive scan of the stomach with many local loop closures.
- Trajectory 3 contains an extensive scan of the stomach with more complicated local loop closures.
- Trajectory 4 consists of more challenging motions including faster rotational and translational movements.
- Trajectory 5 consists of very loopy and complex motions.
- Trajectory 6 is the same as trajectory 5 but included added synthetic noise to allow checking the robustness of the system against noise.
- Before capturing trajectory 7, we added more paraffin oil into the simulator tissue to have stronger reflections. Similarly to trajectory 6, trajectory 7 consists of very loopy and complex motions including very fast rotations, translations and drifting.

Qualitative tracking results of the proposed direct medical SLAM compared to ORB SLAM and to ground truth are shown in Fig. 7. It is clearly observable that direct medical SLAM stays close to the ground truth except for minor deviations in loopy sections, whereas ORB SLAM has major

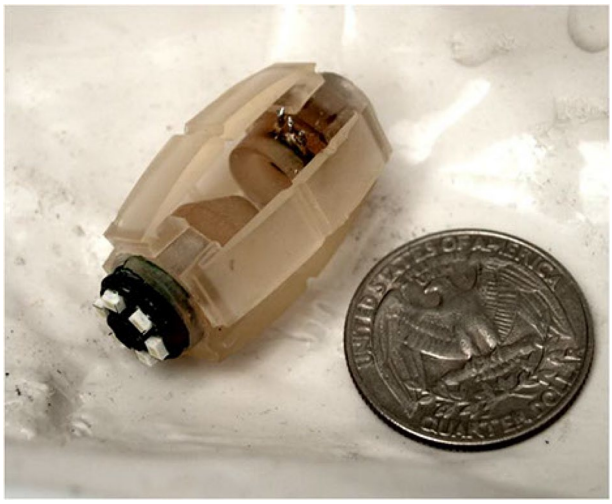


Fig. 6 Photo of the endoscopic capsule robot prototype used in the experiments

Table 1 Trajectory lengths and RMSE results in meters for different endoscopic cameras

Trajectory ID	POTENSIC	MISUMI	AWAIBA	LENGTH
1	0.015	0.019	0.020	0.414
2	0.018	0.020	0.023	0.513
3	0.017	0.021	0.025	0.432
4	0.032	0.037	0.042	0.478
5	0.035	0.039	0.045	0.462
6	0.038	0.043	0.048	0.481
7	0.041	0.044	0.049	0.468

deviations in many sections of the trajectories. For the quantitative analysis, we measured the root-mean-square of the Euclidean distances between the estimated camera poses and the ground truth. As seen in Table 1, the system performs very robustly and tracking accurately in all of the trajectories, not being affected by sudden movements, blur, noise or strong spectral reflections. Figure 9a, b represent rotational and translational RMSE results for different pose estimation strategies including frame-to-model alignment, photometric alignment, frame-to-frame alignment and ORB SLAM as a state-of-the art method. Results indicate that frame-to-model alignment clearly outperforms frame-to-frame alignment, photometric alignment and ORB SLAM. Besides, joint volumetric-photometric alignment outperforms photometric alignment indicating the significance of depth information for pose estimation. Figure 10a, b represent rotational and translational RMSE as a function of ICP weight in joint photometric-volumetric alignment (see Eq. 4). Both RMSEs decrease with higher ICP weights, reaching a minimum at $\omega = 87\%$ and $\omega = 85\%$, respectively.

3.3 Surface estimation

We scanned the non-rigid EGD (Esophagogastroduodenoscopy) simulator to obtain the ground truth 3D data. Reconstructed 3D surface and ground truth 3D data were aligned using iterative closest point algorithm (ICP). RMSE for the reconstructed surface was calculated using the absolute trajectory (ATE) RMSE measuring the root-mean-square of the Euclidean distances between estimated depth values and the corresponding ground truth values. RMSE results in Table 2 show that even in very challenging trajectories with 4–7 sudden movements, strong noise and reflections, our system is capable of providing a reliable and accurate 3D surface reconstruction. A sample 3D reconstruction procedure is shown in Fig. 8 for visual reference.

3.4 Computational performance

To analyze the computational performance of the system, we observed the average frame processing time across trajectories 1–4. The test platform was a desktop PC with an Intel Xeon E5-1660v3- CPU at 3.00, 8 cores, 32 GB of RAM and an NVIDIA Quadro K1200 GPU with 4 GB of memory. The execution time of the system depended on the number of surfels in the map, with an overall average of 48 ms per frame scaling to a peak average of 53 ms implying a worst case processing frequency of 18 Hz.

3.5 Comparison with ORB SLAM

We compared the proposed method with ORB SLAM using our endoscopic capsule dataset. We chose ORB SLAM due to its state-of-the-art performance in various tasks, publicly available code and its recent use in endoscopic applications. We make the following observations after a detailed theoretical and practical evaluation of the differences between the proposed medical SLAM and ORB SLAM:

- ORB SLAM is based on feature matching while direct medical SLAM uses joint photometric- geometric pose estimation. In our evaluation, we observed that for endoscopic images, direct pose estimation is advantageous as compared to feature-based methods because specularity, noise and presence of fewer robustly identifiable features reduce the matching accuracy across frames.

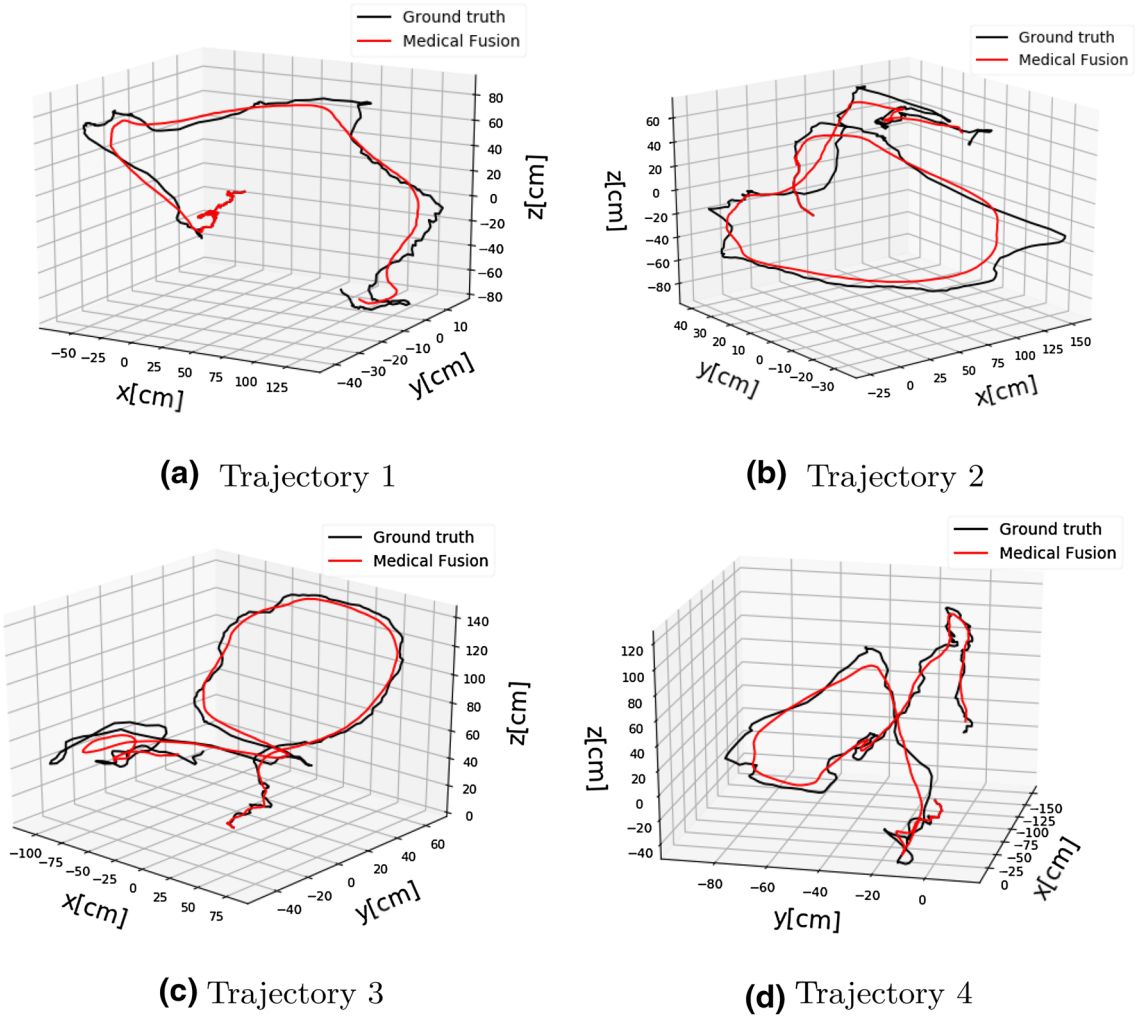


Fig. 7 Sample trajectories estimated by the proposed method, ORB SLAM and ground truth

- Direct medical SLAM needs a good initialization for pose estimation to avoid local minima while ORB SLAM does not require initialization.
- Direct medical SLAM employs a frame-to-model alignment strategy, which is robust to unexpected severe drifts inside GI tract. ORB SLAM on the other hand, performs frame-to-frame alignment and may have difficulties recovering from such drifts.
- Direct medical SLAM is computationally heavy while ORB SLAM can run on standard CPU in real-time. However, modern GPUs can be used to accelerate direct medical SLAM to near real-time as well.
- Direct medical SLAM tolerates larger motions between successive frames, while ORB SLAM expects smaller motions. However, we observed that both methods fail for very large inter-frame motion that leads to small overlap between successive frames.
- ORB SLAM's reconstruction is in the form of a sparse point cloud of the scanned inner organ, whereas direct medical SLAM creates a dense and high quality 3D map of the organ.
- Qualitative and quantitative comparisons depicted in Figs. 7, 9a, b indicate large deviations of ORB SLAM from ground truth, whereas our method is able to stay close to the ground truth even in loopy parts of the trajectories.

4 Conclusion

In this paper, we presented a direct and dense visual SLAM method for endoscopic capsule robots. Our system makes use of surfel-based dense data fusion in combination with frame-to-model tracking and non-rigid deformation. Experimental results suggest the effectiveness of the proposed system, both quantitatively and qualitatively, in occasionally looping endoscopic capsule robot trajectories and comprehensive inner organ scanning tasks. In future,

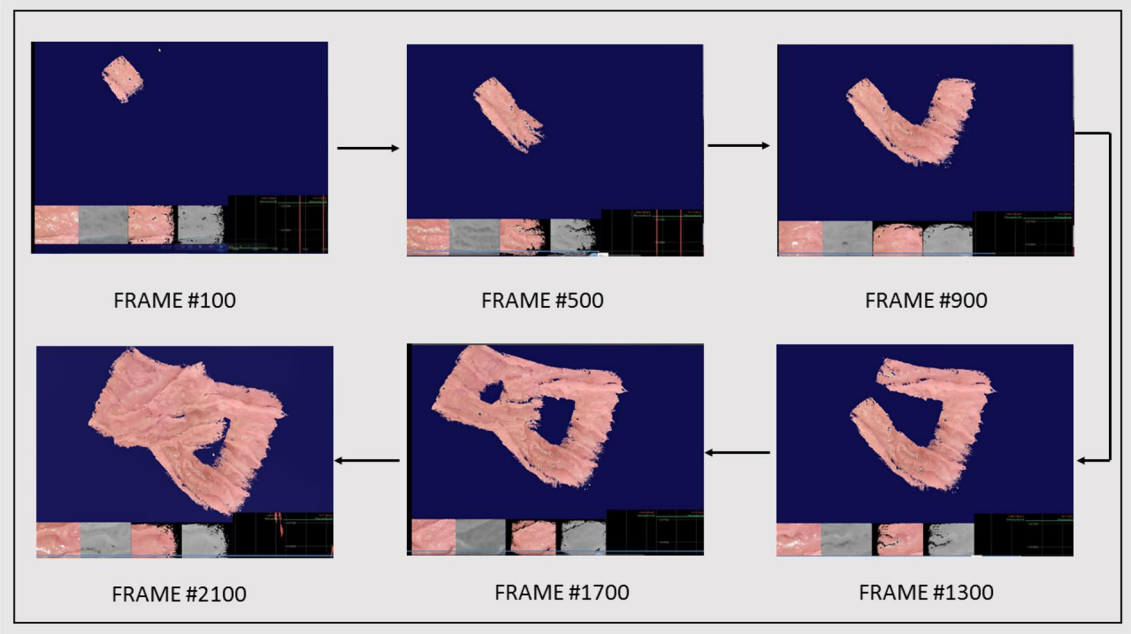
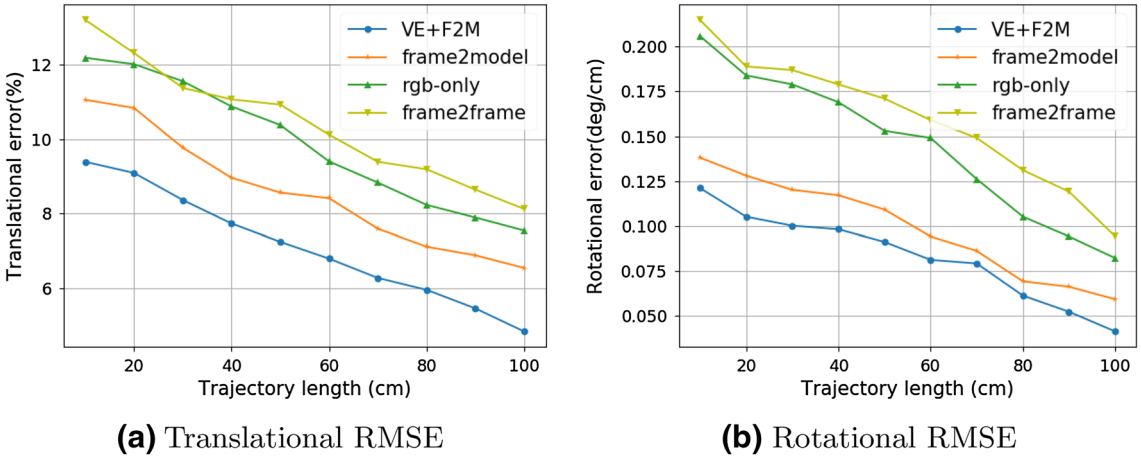
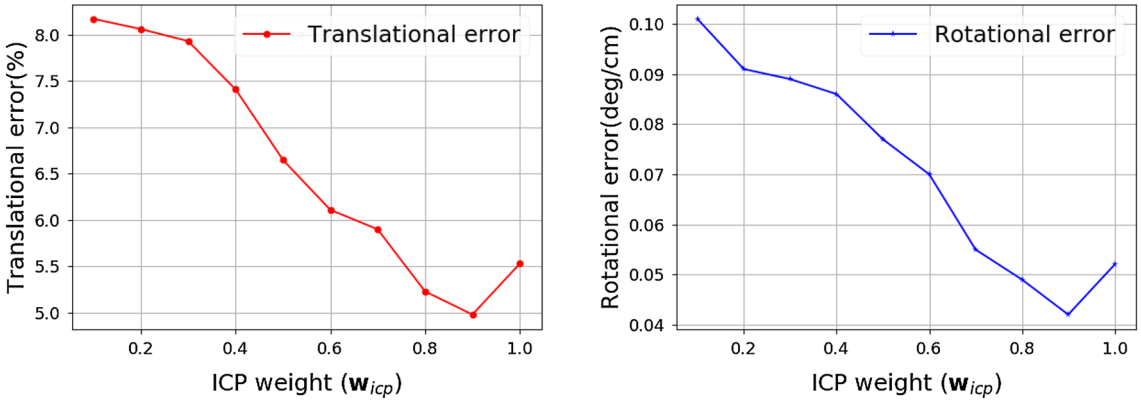


Fig. 8 Frame-by-frame 3D reconstruction of the soft stomach simulator surface by the proposed medical SLAM method



(a) Translational RMSE **(b)** Rotational RMSE

Fig. 9 RMSE results for frame-to-model alignment (frame2model), photometric alignment (rgb-only) and frame-to-frame alignment (frame-2frame)



(a) Translational RMSE **(b)** Rotational RMSE

Fig. 10 RMSE results vs ICP weight

Table 2 Trajectory length and RMSE in meters for 3D surface reconstruction for different endoscopic cameras

Trajectory ID	POTENSIC	MISUMI	AWAIBA	Length
1	0.023	0.026	0.028	0.414
2	0.025	0.029	0.032	0.513
3	0.026	0.030	0.034	0.432
4	0.029	0.033	0.035	0.478
5	0.032	0.035	0.038	0.462
6	0.034	0.037	0.041	0.481
7	0.035	0.042	0.044	0.468

we aim to extend our work into stereo capsule endoscopy applications to achieve even more accurate localization and mapping.

Acknowledgements Open Access Funding provided by Max Planck Society.

References

Carpi, F., Kastelein, N., Talcott, M., Pappone, C.: Magnetically controllable gastrointestinal steering of video capsules. IEEE Trans. Biomed. Eng. **58**(2), 231–234 (2011)

Casado, S., Gil, I., Montiel, J., Grasa, O.G., Bernal, E.: Visual slam for handheld monocular endoscope. Med. Imaging IEEE Trans. **33**(1), 135–146 (2014)

Fluckiger, M., Nelson, B.J.: Ultrasound emitter localization in heterogeneous media. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE **2007**, 2867–2870 (2007)

Keller, H., Juloski, A., Kawano, H., Bechtold, M., Kimura, A., Takizawa, H., Kuth, R.: Method for navigation and control of a magnetically guided capsule endoscope in the human stomach. In: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), IEEE, pp. 859–865 (2012)

Kim, K., Johnson, L.A., Jia, C., Joyce, J.C., Rangwalla, S., Higgins, P.D., Rubin, J.M.: Noninvasive ultrasound elasticity imaging (uei) of crohn's disease: animal model. Ultrasound Med. Biol. **34**(6), 902–912 (2008)

Liao, Z., Gao, R., Xu, C., Li, Z.-S.: Indications and detection, completion, and retention rates of small-bowel capsule endoscopy: a systematic review. Gastrointest. Endosc. **71**(2), 280–286 (2010)

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

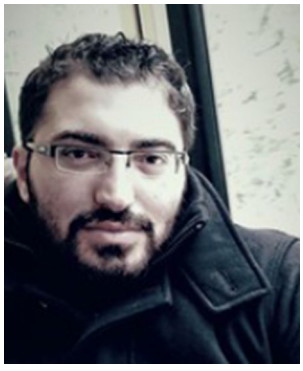
Mahoney, A.W., Wright, S.E., Abbott, J.J.: Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, pp. 5366–5371 (2013)

Mountney, P., Stoyanov, D., Davison, A., Yang, G.-Z.: Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 347–354 (2006)

Mountney, P., Yang, G.Z.: Dynamic view expansion for minimally invasive surgery using simultaneous localization, mapping, Visual slam for handheld monocular endoscope, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2009)

Mountney, P., Yang, G.-Z.: Motion compensated slam for image guided surgery. Med. Image Comput. Comput. Assist. Intervent. MICCAI (2010)

Nakamura, T., Terano, A.: Capsule endoscopy: past, present, and future. J. Gastroenterol. **43**(2), 93–99 (2008)
Newcombe, R.A, Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: real-time dense surface mapping and tracking. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, IEEE, pp. 127–136 (2011)
Pan, G., Wang, L.: Swallowable wireless capsule endoscopy: progress and technical challenges. Gastroenterol Res Pract (2012)
Ping-Sing, T., Shah, M.: Shape from shading using linear approximation. Image Vis. Comput. **12**(8), 487–498 (1994)
Qian, X., Sanchez, J., Sun, Y., Lin, B., Johnson, A.: Motion compensated slam for image guided surgery. Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery (2013)
Rubin, J.M., Xie, H., Kim, K., Weitzel, W.F., Emelianov, S.Y., Aglyamov, S.R., Wakefield, T.W., Urquhart, A.G., ODonnell, M.: Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans. J. Ultrasound Med. **25**(9), 1179–1186 (2006)
Sitti, M., Ceylan, H., Hu, W., Giltinan, J., Turan, M., Yim, S., Diller, E.: Biomedical applications of untethered mobile milli/microrobots. Proc. IEEE **103**(2), 205–224 (2015)
Son, D., Yim, S., Sitti, M.: A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors. IEEE/ASME Trans. Mechatron. **21**(2), 708–716 (2016)
Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.-Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, New York, pp. 275–282 (2010)
Than, T.D., Alici, G., Zhou, H., Li, W.: A review of localization systems for robotic endoscopic capsules. IEEE Trans. Biomed. Eng. **59**(9), 2387–2399 (2012)
Whelan, T., Leutenegger, S., Salas-Moreno, R.F., Glocker, B., Davison, A.J.: Elasticfusion: Dense slam without a pose graph. In: Robotics: Science and Systems, Vol. 11 (2015)
Yim, S., Goyal, K., Sitti, M.: Magnetically actuated soft capsule with the multimodal drug release function. IEEE/ASME Trans. Mechatron. **18**(4), 1413–1418 (2013)
x
Yim, S., Gultepe, E., Gracias, D.H., Sitti, M.: Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers. IEEE Trans. Biomed. Eng. **61**(2), 513–521 (2014)



Mehmet Turan received his Diploma Degree from the Information technology and Electronics engineering department of RWTH Aachen, Germany in 2012. He was a research scientist at UCLA (University of California Los Angeles) between 2013–2014 and a research scientist at the Max Planck Institute for Intelligent Systems between 2014-present. He is currently enrolled as a PhD Student at the ETH Zurich, Switzerland. He is also affiliated with Max Planck-ETH Center for Learning Systems, the first joint research center of ETH Zurich and the Max Planck Society. His research interests include SLAM (simultaneous localization and mapping) techniques for millimeter-scale medical robots and deep learning techniques for medical robot localization and mapping. He received DAAD fellowship between years 2005–2011 and Max Planck Fellowship between 2014-present. He has also received MPI-ETH Center fellowship between 2016-present.

Yasin Almalioglu received the BSc degree with honours in computer engineering from Bogazici University, Istanbul, Turkey in 2015. He was a research intern at CERN Geneva, Switzerland and Astroparticle and Neutrino Physics Group at ETH Zurich, Switzerland in 2013 and 2014, respectively. He is currently pursuing the MSc degree in computer engineering at Bogazici University, Istanbul, Turkey. His research interests include machine learning, Bayesian statistics, Monte Carlo methods, probabilistic graphical models, artificial neural networks and mobile robot localization. He received the Engin Arik Fellowship in 2013.

Helder Araujo is a Professor at the Department of Electrical and Computer Engineering of the University of Coimbra. His research interests include Computer Vision applied to Robotics, robot navigation and visual servoing. In the last few years he has been working on non-central camera models, including aspects related to pose estimation, and their applications. He has also developed work in Active Vision, and on control of Active Vision systems. Recently he has started work on the development of vision systems applied to medical endoscopy.

Ender Konukoglu, PhD, finished his PhD at INRIA Sophia Antipolis in 2009. From 2009 till 2012 he was a post-doctoral researcher at Microsoft Research Cambridge. From 2012 till 2016 he was a junior faculty at the Athinoula A. Martinos Center affiliated to Massachusetts General Hospital and Harvard Medical School. Since 2016 he is an Assistant Professor of Biomedical Image Computing at ETH Zurich. He is interested in developing computational tools and mathematical methods for analysing medical images with the aim to build decision support systems. He develops algorithms that can automatically extract quantitative image-based measurements, statistical methods that can perform population comparisons and biophysical models that can describe physiology and pathology.

Dr. Metin Sitti received the BSc and MSc degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1992 and 1994, respectively, and the PhD degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. He was a research scientist at UC Berkeley during 1999–2002. He has been a professor in the Department of Mechanical Engineering and Robotics Institute at Carnegie Mellon University, Pittsburgh, USA since 2002. He is currently a director at the Max Planck Institute for Intelligent Systems in Stuttgart. His research interests include small-scale physical intelligence, mobile microrobotics, bio-inspired materials and miniature robots, soft robotics, and micro-/nanomanipulation. He is an IEEE Fellow. He received the SPIE Nanoengineering Pioneer Award in 2011 and NSF CAREER Award in 2005. He received many best paper, video and poster awards in major robotics and adhesion conferences. He is the editor-in-chief of the Journal of Micro-Bio Robotics.

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 demonstrated 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** 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 surrouding tissues. 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]. Morover, 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

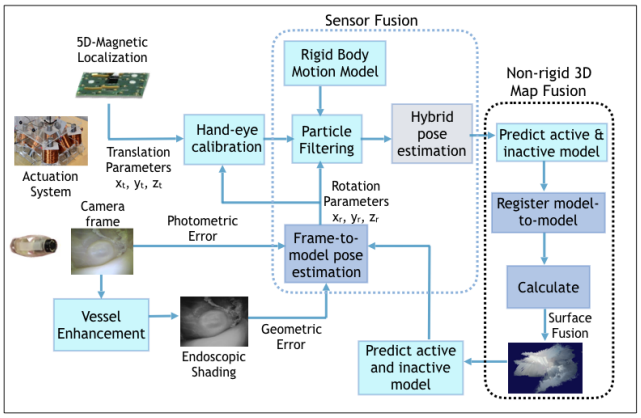


Fig. 1: System overview including 5-DoF magnetic localization, 6-DoF visual joint photometric-geometric frame-to-model pose optimization, inter-sensor calibration, particle filtering based sensor fusion, non-rigid deformations based frame-to-model map fusion.

[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

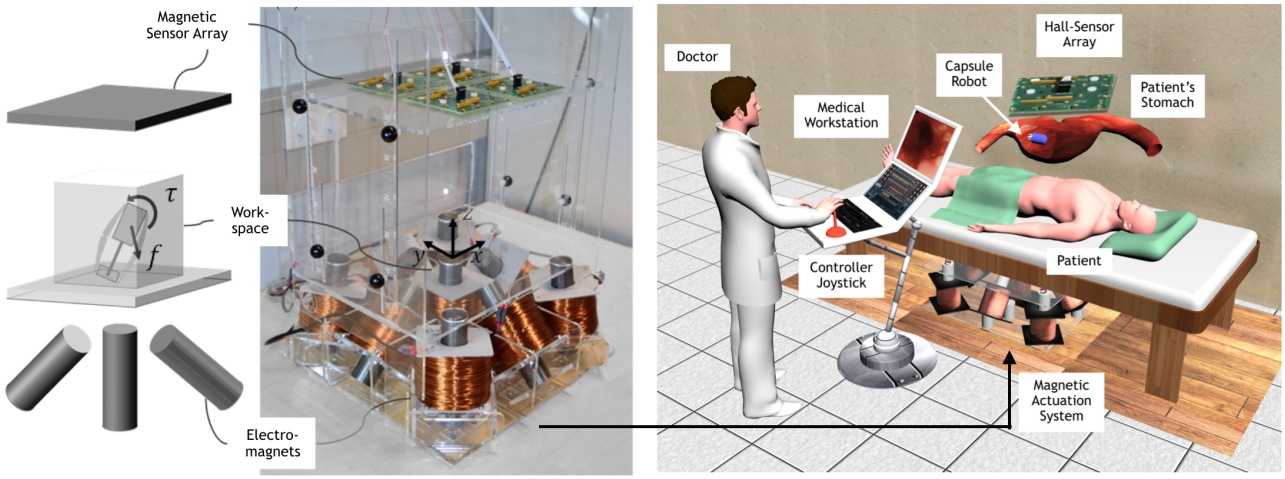


Fig. 2: Demonstration of the active endoscopic capsule robot operation using MASCE (Magnetically actuated soft capsule endoscope) designed for disease detection, drug delivery and biopsy-like operations in the upper GI-tract. MASCE is composed of a RGB camera, a permanent magnet, an empty space for drug chamber and a biopsy tool. Electromagnetic coils based actuation unit below the patient table excerts forces and torques to execute the desired motion. Medician operates the screening, drug delivery and biosy processes 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 and to execute desired therapeutic actions such as drug release and biopsy. Actuation system of the MASCE: The magnet exerts magnetic force and torque on the capsule in response to a controlled external magnetic field. The magnetic torque and forces are used to actuate the capsule robot and to release drug. Magnetic fields from the electromagnets generate the magnetic force and torque on the magnet inside MASCE so that the robot moves inside the workspace. Sixty-four three-axis magnetic sensors are placed on the top, and nine electromagnets are placed in the bottom.

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 paper, we propose that a non-rigidly deformable RGB Depth fusion method, which combines magnetic localization and visual pose estimation using particle filtering, can provide real-time, accurate localization and mapping for endoscopic capsule robots. We demonstrate the system in four different ex-vivo porcine stomachs by measuring its performance in terms of both surface mapping and capsule localization accuracy.

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:

- Estimate 3D position of the endoscopic capsule robot pose using magnetic localization system;
- Estimate 3D rotation of the endoscopic capsule robot pose using visual joint photometric-geometric frame-to-model pose optimization;
- Perform offline inter-sensor calibration between magnetic hall sensor array and capsule camera system;
- 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 estimation and deformation constraints defined by the graph equations;
- In case there exists an intersection of the active model with the inactive model within the current frame, fuse intersecting regions and deform the entire model non-rigidly.

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

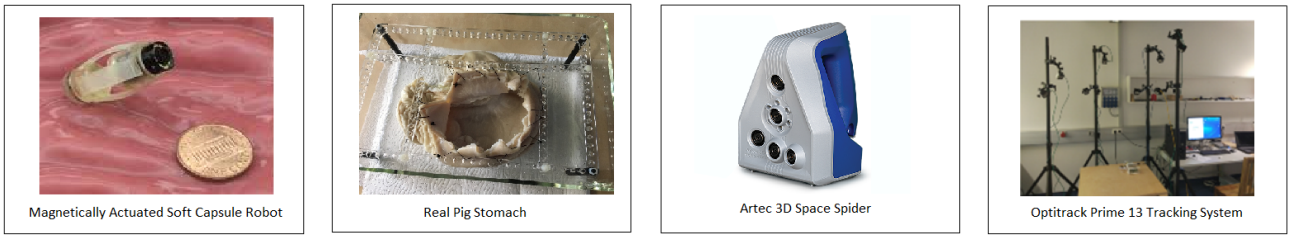


Fig. 3: Illustration of the experimental setup. MASCE is a magnetically actuated robotic capsule endoscope prototype which has a ringmagnet on the body. An electromagnetic coil array consisting of nine coils is used for the actuation of the MASCE. An opened and oiled porcine stomach simulator is used 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.

which is acquired by a Hall-effect sensor array. As a further step, second-order directional differentiation is applied to reduce the localization error. 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).

B. Visual Localization

We propose the use of a direct surfel map fusion method for actively controllable endoscopic capsule robots. 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. It performs joint volumetric and photometric alignment, frame-to-model predictive tracking, and dense model-to-model loop closure with non-rigid space deformation. Prior to using endoscopic video with such a method, the images must first be prepared.

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 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: 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:

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

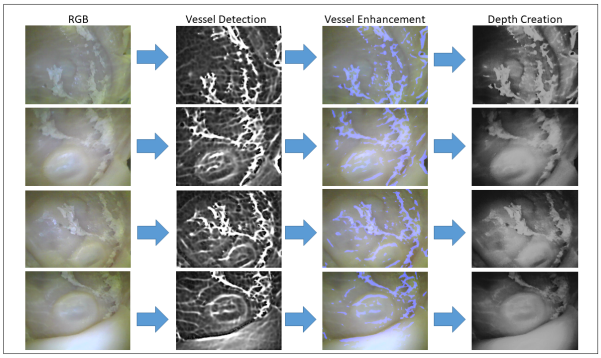


Fig. 4: For a given RGB frame, we extract the Hessian matrix and derive its eigenvalues and principal directions to detect the vessel. We convolve the input frame and final output to create multiple scale representations to identify the different vessels. After enhancement of vessel detected frame, we 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.

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 $\mathbf{T} \rightarrow \exp(\hat{\xi})\mathbf{T}$. 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 $\mathcal{G}^n$ has a timestamp $\mathcal{G}^n_{t_0}$, a position $\mathcal{G}^n_g \in \mathbb{R}^3$ and a set of neighboring nodes $\mathcal{N}(\mathcal{G}^n)$. The directed edges of the graph are neighbors of each node. A graph is connected up to a neighbor count k such that $\forall n, |\mathcal{N}(\mathcal{G}^n)| = k$. Each node also stores an affine transformation in the form of a 3×3 matrix $\mathcal{G}^n_{\mathbf{R}}$ and a 3×1 vector $\mathcal{G}^n_{\mathbf{t}}$. When deforming a surface, the $\mathcal{G}^n_{\mathbf{R}}$ and $\mathcal{G}^n_{\mathbf{t}}$ parameters of each node are optimized according to surface constraints. In order to apply a deformation graph to the surface, each surfel $\mathcal{M}^s$ identifies a set of influencing nodes in the graph $\mathcal{I}(\mathcal{M}^s, \mathcal{G})$. The deformed position of a surfel is given by:

$$\hat{\mathcal{M}}^s_{\mathbf{p}} = \phi(\mathcal{M}^s) = \sum_{n \in \mathcal{I}(\mathcal{M}^s, \mathcal{G})} w^n(\mathcal{M}^s)[\mathcal{G}^n_{\mathbf{R}}(\mathcal{M}^s_{\mathbf{p}} - \mathcal{G}^n_{\mathbf{g}}) + \mathcal{G}^n_{\mathbf{g}} + \mathcal{G}^n_{\mathbf{t}}] \tag{6}$$

while the deformed normal of a surfel is given by:

$$\hat{\mathcal{M}}^s_{\mathbf{p}} = \sum_{n \in \mathcal{I}(\mathcal{M}^s, \mathcal{G})} w^n(\mathcal{M}^s) {\mathcal{G}^n_{\mathbf{R}}}^{-1^T} \mathcal{M}^s_{\mathbf{n}}, \tag{7}$$

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

$$w^n(\mathcal{M}^s) = (1 - ||\mathcal{M}^s_{\mathbf{p}} - \mathcal{G}^n_{\mathbf{g}}||_2 / d_{\max})^2. \tag{8}$$

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.

C. Particle Filtering based Magneto-Visual Sensor Fusion

We developed a particle filtering based sensor fusion method for endoscopic capsule robots which provides robustness against sensor failure through the introduction of latent variables characterizing the sensor's reliability as either normal or failing, which are estimated along with the system state. The method is inspired by and modifed from [31]. As motion model, we use a rigid motion model (3D rotation and 3D translation) assuming 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, which include the scale factor, and the 3D rotation using visual information provided by the monocular endoscopic capsule camera.

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:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{v}_t) \tag{9}$$

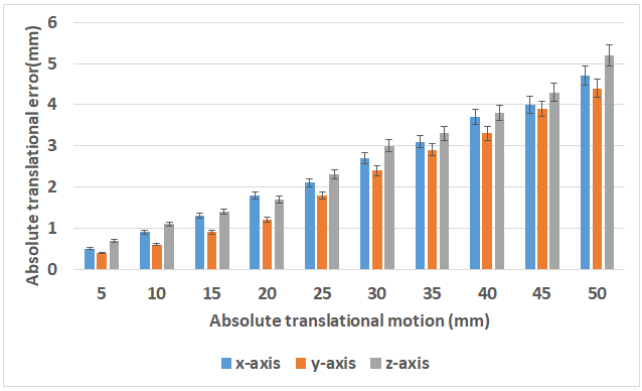
where f is a non-linear state transition function and $\mathbf{v_t}$ is white noise. t is the index of a time sequence, $t \in \{1, 2, 3, ...\}$. Observations of the pose are produced by n sensors $\mathbf{z}_{k,t} (k = 1, ..., n)$ in general, where the probability distribution $p(\mathbf{z}_{k,t}|\mathbf{x}_t)$ 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:

$$\mathbf{z}_{k,t} = h_{k,s_{k,t},t}(\mathbf{x}_t) + \mathbf{w}_{k,s_{k,t},t} \tag{10}$$

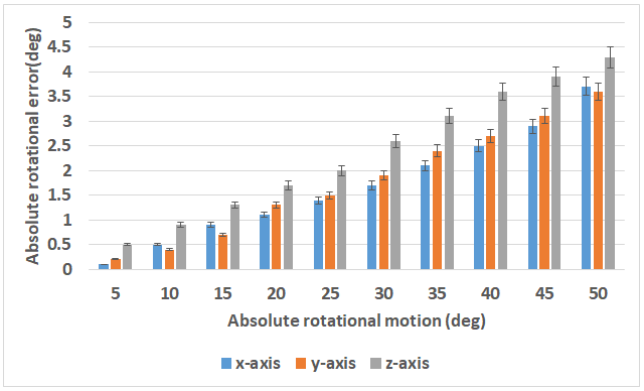
where $h_{k,s_{k,t},t}(\mathbf{x}_t)$ is the non-linear observation function and $\mathbf{w}_{k,s_{k,t},t}$ is the observation noise. The latent variable of the switch parameter $s_{k,t}$ is defined to be 0 if the sensor is in a failure state, which means that observation $\mathbf{z}_{k,t}$ is statistically independent of $\mathbf{x}_t$, and 1 if the sensor k is in its nominal state of work. 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:

$$Pr(s_{k,t} = j) = \alpha_{k,j,t}, \quad 0 \leq j \leq d_k \tag{11}$$

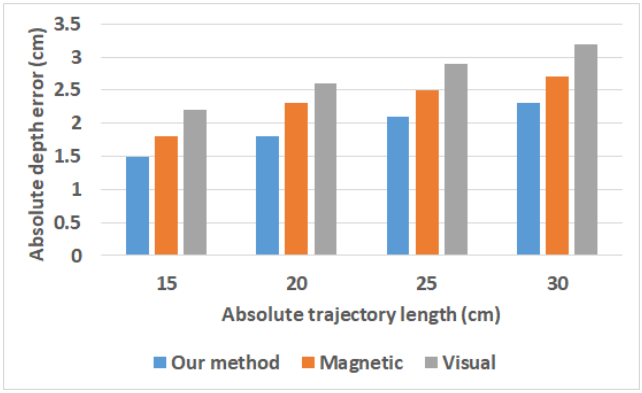
where $\alpha_{k,j,t} \geq 0$ and $\sum_{j=0}^{d_k} \alpha_{k,j,t} = 1$ with a Markov evolution property. The objective posterior density function $p(\mathbf{x}_{0:t}, \mathbf{s}_{1:t}, \alpha_{0:t}|\mathbf{z}_{1:t})$ and the marginal posterior probability $p(\mathbf{x}_t|\mathbf{z}_{1:t})$, in general, cannot be determined in a closed form due to its complex shape. However, sequential Monte Carlo



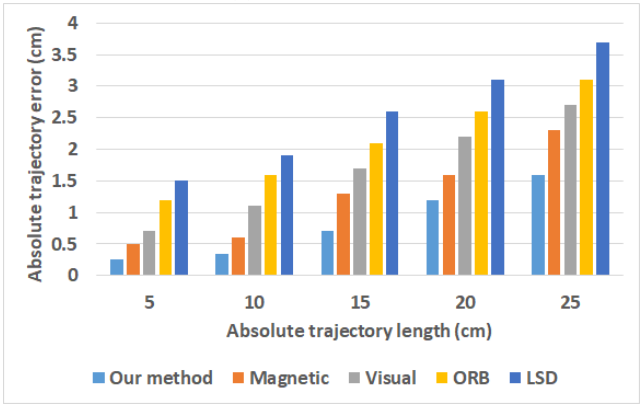
(a) Translational error



(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.

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.

Sensor Failure Detection and Handling: The proposed multi-sensor fusion approach is able to detect the sensor failure periods and to handle the failures, accordingly. As seen in Fig. 6, the posterior probabilities of the switch parameters $\mathbf{s}_{k,t}$ and the minimum mean square error (MMSE) estimates of $\alpha_{k,t}$ indicate an accurate detection of sensor failure states. Visual localization failed between seconds 14-36 due to very fast frame-to-frame motions and magnetic sensor failed between seconds 57-76 due to increased distance of the ringmagnet to the sensor array. Once a sensor failure is detected, the approach stops to use this sensor information until the failure state ends and uses prior information and rigid body motion model to predict the misssing information. Thanks to this switching option ability, MMSE is kept low during sensor failure as seen in Figure 6. In our sensor failure model, we do not make a Markovian assumption for the switch variable $\mathbf{s}_{k,t}$ but we do for its prior $\alpha_{k,t}$, resulting in a priori dependent on the past trajectory sections, which is more likely for the incremental endoscopic capsule robot motions. The model thus introduces a memory over the past sensor states rather than simply considering the last state. The length of the memory is tuned by the hyper-parameters $\sigma^{\alpha}_{k,t}$, leading to a long memory for large values and vice-versa. This is of particular interest when considering sensor failures. Our system detects automatically failure states. Hence, the confidence in the vision sensor decreases when visual localization fails recently due to occlusions, fast-frame-to frame changes etc. On the other hand, the confidence in magnetic sensor decreases if the magnetic localization fails due to noise interferences from environment and/or if the ringmagnet has a big distance to the magnetic sensor array.

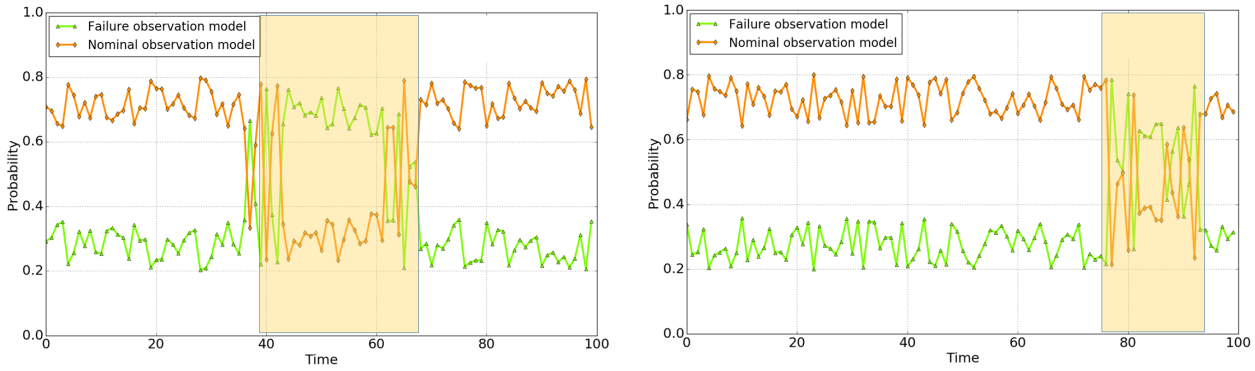


Fig. 6: The minimum mean square error (MMSE) of $\alpha_{k,t}$ for endoscopic RGB camera (left) and for magnetic localization system (right). The switch parameter, $\mathbf{s}_{k,t}$, and the confidence parameter $\alpha_{k,t}$ reflect the failure times accurately: Visual localization fails between $39-68$ seconds and magnetic localization fails between $78-92$ seconds. Both failures are detected confidentially.

D. Relative pose of magnetic and visual localization systems

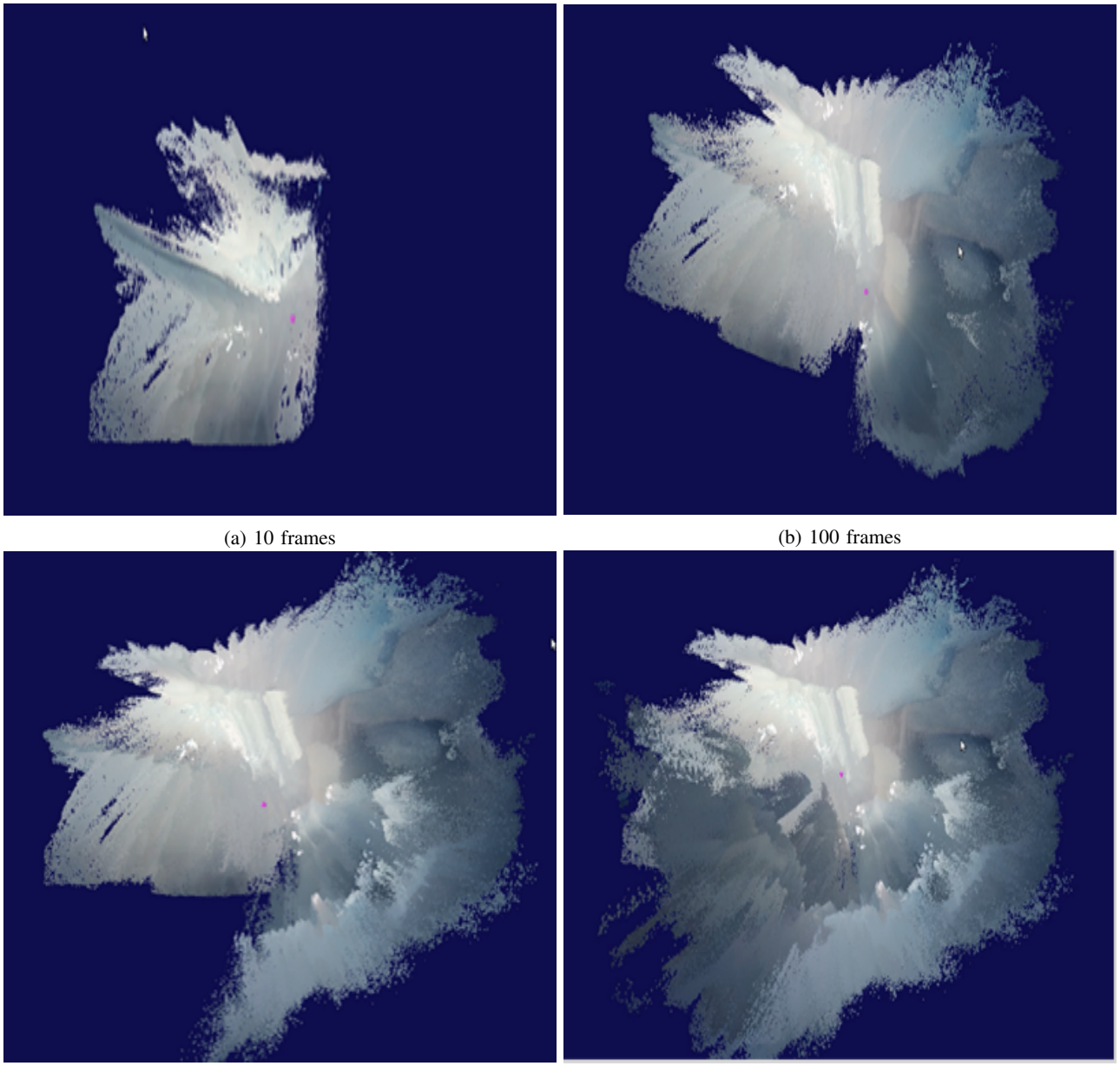
To relate the magnetic actuation and localization system (which is seen in Fig. 2) with the proposed vision system, the relative pose has to be estimated. The relative pose can be estimated using rigid motion from the capsule and the constraint of the rigid transformation between the magnetic sensor coordinate system and the camera coordinate system (as in eye-in-hand calibration). The vision system measures the pose of the camera, and the magnetic localization system measures the 5D pose of the magnet on the MASCE. The transformation between the coordinate frames attached to the ringmagnet and to the camera origin must be known, because the particle filter assumes that the two systems make measurements on the same system state, which in this case is a single rigid body pose associated with the capsule. In this case the magnetic system provides a 5-DoF pose while the vision system yields a 6-DoF pose. 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 estimated. The use of several motions allowed the estimation of the uncertainty in the parameters.

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

For the duration of the pose and map reconstruction evaluations, we have only utilized sequences where the Bayesian filtering algorithm confirmed that camera and magnetic sensor remained in the nominal sensor state. 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 reconstruction 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. 7, for visual reference.



(a) 10 frames (b) 100 frames

(c) 300 frames (d) 500 frames

Fig. 7: 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 .

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.

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

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

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 surfel-based dense reconstruction in combination with particle filter based fusion of magnetic and visual localization information and sensor failure detection. 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. In the future, *in vivo* testing is required to validate the accuracy and robustness of the approach in the challenging conditions of the GI tract. We also intend to extend our work into stereo capsule endoscopy applications to achieve even more accurate localization and mapping. In addition, an improved estimation of the relative pose between the coordinate systems of the sensors may result in improved accuracy.

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.

Neurocomputing 275 (2018) 1861–1870

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots

Mehmet Turan[a,d,*], Yasin Almalioglu[b], Helder Araujo[c], Ender Konukoglu[d], Metin Sitti[a]

[a] Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany
[b] Computer Engineering Department, Bogazici University, Istanbul, Turkey
[c] Institute for Systems and Robotics, University of Coimbra, Coimbra, Portugal
[d] Department of Information Technology and Electrical Engineering, Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 16 February 2017
Revised 23 August 2017
Accepted 6 October 2017
Available online 7 November 2017

Communicated by Wei Wu

Keywords:
Endoscopic capsule robot
Visual odometry
Sequential deep learning
RCNN
CNN
LSTM
Localization

ABSTRACT

Ingestible wireless capsule endoscopy is an emerging minimally invasive diagnostic technology for inspection of the GI tract and diagnosis of a wide range of diseases and pathologies. Medical device companies and many research groups have recently made substantial progresses in converting passive capsule endoscopes to active capsule robots, enabling more accurate, precise, and intuitive detection of the location and size of the diseased areas. Since a reliable real time pose estimation functionality is crucial for actively controlled endoscopic capsule robots, in this study, we propose a monocular visual odometry (VO) method for endoscopic capsule robot operations. Our method lies on the application of the deep recurrent convolutional neural networks (RCNNs) for the visual odometry task, where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used for the feature extraction and inference of dynamics across the frames, respectively. Detailed analyses and evaluations made on a real pig stomach dataset proves that our system achieves high translational and rotational accuracies for different types of endoscopic capsule robot trajectories.

1. Introduction

Following the advances in material science in last decades, untethered pill-size, swallowable capsule endoscopes with an onboard camera and wireless image transmission device have been developed and used in hospitals for screening the gastrointestinal tract and diagnosing diseases such as the inflammatory bowel disease, the ulcerative colitis and the colorectal cancer. Unlike standard endoscopy, endoscopic capsule robots are non-invasive, painless and more appropriate to be employed for long duration screening purposes. Moreover, they can access difficult body parts that were not possible to reach before with standard endoscopy (e.g., small intestines). Such advantages make pill-size capsule endoscopes a significant alternative screening method over standard endoscopy [1–5]. However, current capsule endoscopes used in hospitals are passive devices controlled by peristaltic motions of the inner organs. The control over capsule's position, orientation, and functions would give the doctor a more precise reachability of targeted body parts and more intuitive and correct diagnosis opportunity [6–10]. Therefore, several groups have recently proposed active, remotely controllable robotic capsule endoscope prototypes equipped with additional functionalities such as local drug delivery, biopsy and other medical functions [2,11–19]. However, an active motion control needs feedback from a precise and reliable real time pose estimation functionality. In last decade, several localization methods [4,20–23] were proposed to calculate the 3D position and orientation of the endoscopic capsule robot such as fluoroscopy [4], ultrasonic imaging [20–23], positron emission tomography (PET) [4,23], magnetic resonance imaging (MRI) [4], radio transmitter based techniques and magnetic field based techniques [16]. The common drawback of these localization methods is that they require extra sensors and hardware design. Such extra sensors have their own deficiencies and limitations if it comes to their application in small scale medical devices such as space limitations, cost aspects, design incompatibilities, biocompatibility issue and the interference of sensors with activation system of the device.

* Corresponding author at: Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany.
E-mail addresses: mturan@student.ethz.ch (M. Turan), yasin.almalioglu@boun.edu.tr (Y. Almalioglu), helder@isr.uc.pt (H. Araujo), ender.konukoglu@vision.ee.ethz.ch (E. Konukoglu), sitti@is.mpg.de (M. Sitti).

https://doi.org/10.1016/j.neucom.2017.10.014

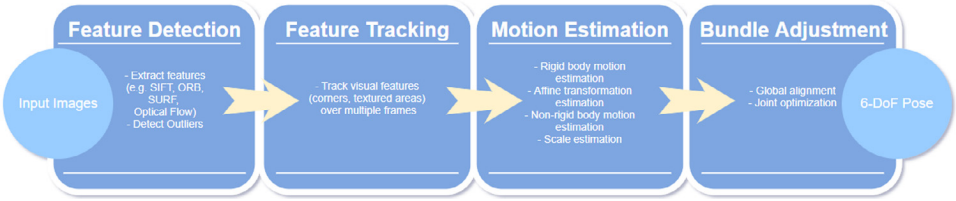


Fig. 1. Traditional visual odometry pipeline.

As a solution of these issues, a trend of visual odometry methods have attracted the attention for the localization of such small scale medical devices. A classic visual odometry pipeline typically consisting of camera calibration, feature detection, feature matching, outliers rejection (e.g. RANSAC), motion estimation, scale estimation and global optimization (bundle adjustment) is depicted in Fig. 1. Although some state-of-the-art algorithms based on this traditional pipeline have been applied for the visual odometry task of the hand-held endoscopes in the past decades, their main deficiency is tracking failures in low textured areas. In last years, deep learning (DL) techniques have been dominating many computer vision related tasks with some promising result, e.g. object detection, object recognition, classification problems etc. Contrary to these high-level computer vision tasks, VO is mainly working on motion dynamics and relations across sequence of images, which can be defined as a sequential learning problem. With that motivation, we propose a novel monocular VO algorithm based on deep recurrent convolutional neural networks (RCNNs). Since it is designed in an end-to-end fashion, it does not need any module from the classic VO pipeline to be integrated. The main contributions of our paper are as follows:

- To the best of our knowledge, this is the first monocular VO approach through deep learning techniques developed for the endoscopic capsule robot and hand-held standard endoscope localization.
- Neither prior knowledge nor parameter tuning is needed to recover the absolute trajectory scale contrary to monocular traditional VO approach.
- A novel RCNN architecture is introduced which can successfully model sequential dependence and complex motion dynamics across endoscopic video frames.
- A real pig stomach dataset and a synthetic human simulator dataset with 6-DoF ground truth pose labels and 3D scan are recorded, which we are considering to publish for the sake of other researchers in that area.

The proposed method solves several issues faced by typical visual odometry pipelines, e.g. the need to establish a frame-to-frame feature correspondence, vignetting, motion blur, specularity or low signal-to-noise ratio (SNR). We think that DL based endoscopic VO approach is more suitable for such challenge areas since the operation environment (GI tract) has similar organ tissue patterns among different patients which can be learned by a sophisticated machine learning approach easily. Even the dynamics of common artefacts such as vignetting, motion blur and specularity across frame sequences could be learned and used for a better pose estimation.

As the outline of this paper, Section 2 introduces the proposed RCNN based localization method in detail. Section 3 presents our dataset and the experimental setup. Section 4 shows our experimental results, we achieved for 6-DoF localization of the endoscopic capsule robot. Section 5 gives future directions.

2. System overview and analysis

Our architecture makes use of inception modules for feature extraction and RNN for sequential modelling of motion dynamics to regress the robot's orientation and position in real time (5.3 ms per frame). It takes two consecutive endoscopic RGB Depth frames each with timestamp and regresses the 6-DoF pose of the robot without need of any extra sensor. For the depth image creation from RGB input images, we used shape from shading (SfS) technique of Tsai and Shah, which is based on the following assumptions [24]:

- The object surface is Lambertian;
- The light comes from a single point light source;
- The surface has no self-shaded areas.

For more details of the Tsai–Shah SfS method, the reader is referred to the original paper of the authors. In past couple of years, some powerful CNN architectures, such as GoogleNet [25], VGG16 [26], ResNet50 [27] have been developed and evaluated for various high level computer vision tasks, e.g. object detection, object recognition and classification [25,28–30]. One major drawback of CNN architectures is the fact that they only analyse just-in-moment information, whereas VO is rather dependent on the correlative information across frames. Unlike traditional feed-forward artificial neural networks, RCNN can use its internal memory to process arbitrarily long sequences by its directed cycles between the hidden units. Therefore, we think that RCNN architectures are more suitable than CNN architectures for VO tasks. The proposed deep EndoVO (endoscopic visual odometry) approach works as follows:

Algorithm 1 Deep EndoVO.

1: Take two consecutive input RGB images.
2: Create the depth images from RGB images using Tsai–Shah SfS method.
3: Subtract mean RGB Depth value of the training set from the RGB Depth images.
4: Stack the preprocessed RGB Depth frame pair to form a tensor.
5: Serve the tensor into the stack of inception modules to create the feature vector.
6: Feed the feature representation into the RNN layers.
7: Estimate the 6-DoF relative pose.

The proposed DL network consists of three inception layers and two LSTM layers concatenated sequentially. The inception layers, imitating visual cortex of human beings, are basically extracting multi-level features; i.e, features of different sizes such as small details, middle-size or larger features (see Fig. 3b). The final inception layer passes the feature representation into the RNN modules (see Fig. 3a). RNNs are very suitable for modelling the dependencies across image sequences and for creating a temporal motion model since it has a memory of hidden states over time and has directed cycles among hidden units, enabling the current hidden state to be a function of arbitrary sequences of inputs (see Fig. 3a).

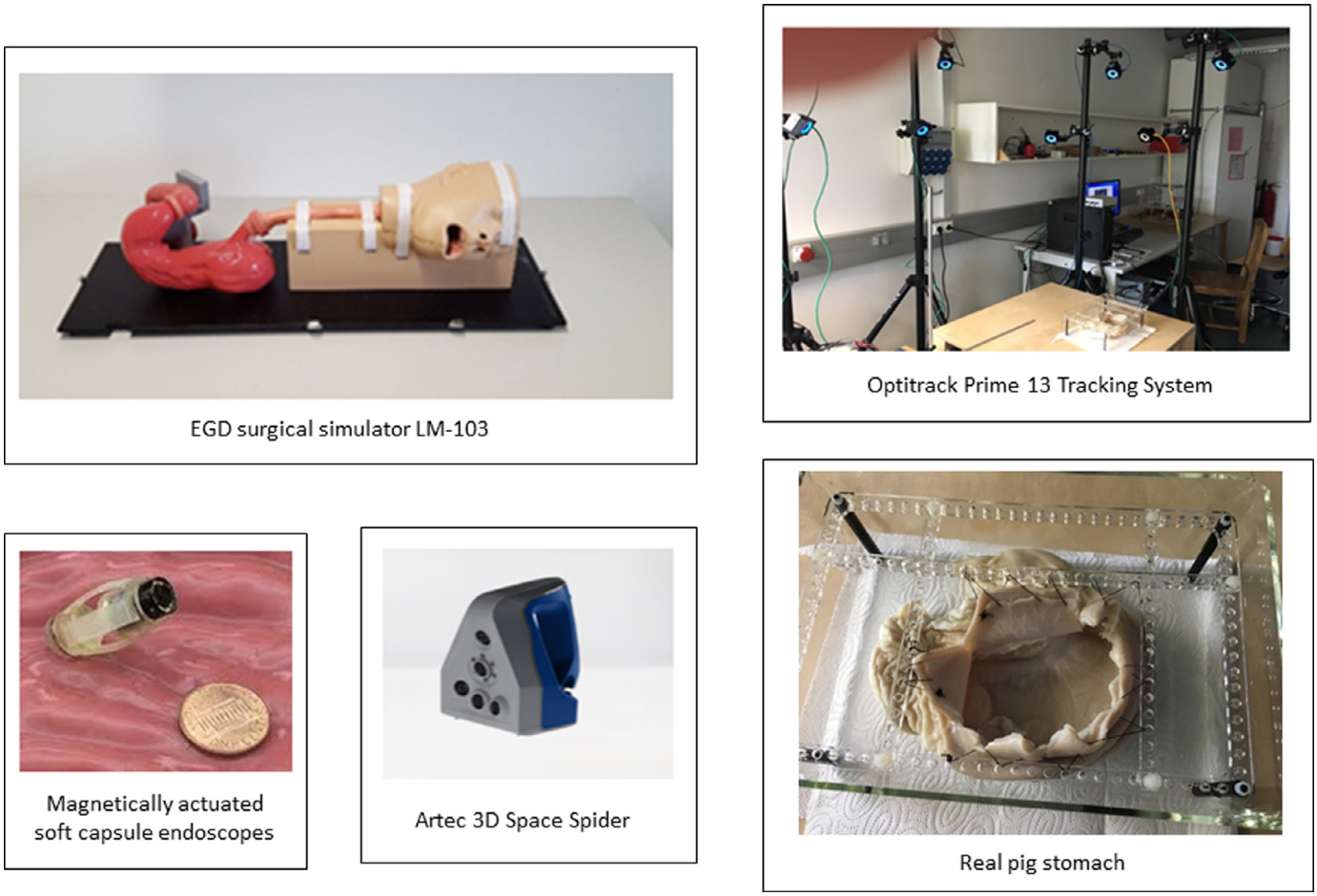
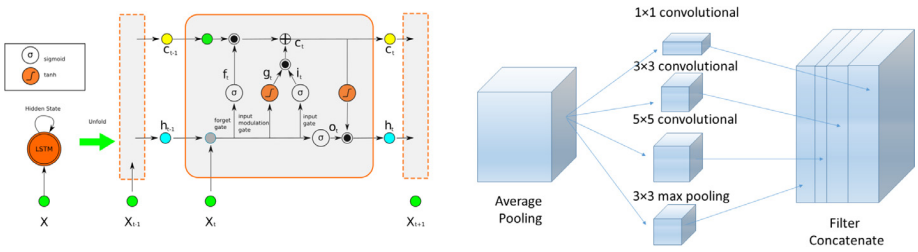


Fig. 2. Experimental overview.



(a) Information flow through the hidden units of the LSTM [31].

(b) Inception layer[25]

Fig. 3. The structure of the LSTM and inception layers of the proposed model is shown.

Thus, using RNN, the pose estimation of the current frame benefits from information encapsulated in previous frames [32,33]. Given a set of inception features x_k at time k, RNN updates at time step k, W denote corresponding weight matrices of the hidden units, b the bias vector, and H an element-wise hyperbolic tangent based activation function. Long short-term memory (LSTM) is more suitable than RNN to exploit longer trajectories since it avoids the vanishing gradient problem of RNN resulting in a higher capacity of learning long-term relations among the sequences by introducing memory gates such as input, forget and output gates and hidden units of several blocks. The input gate controls the amount of new information flowing into the current state, the forget gate adjusts the amount of existing information that remains in the memory and the output gate decides which part of the information triggers the activations. The folded LSTM and its unfolded version over time are shown in Fig. 3a along with the internal structure of a LSTM memory cell. It can be seen that unfolded LSTMs correspond to timestamps. Given the input vector x_k at time k, the output vector h_{k-1} and the cell state vector c_{k-1} of the previous LSTM unit, the LSTM updates at time step k according to the following equations, where σ is sigmoid non-linearity, tanh is hyperbolic tangent non-linearity, W terms denote corresponding weight matrices, b terms denote bias vectors, i_k, f_k, g_k, c_k and o_k are input gate, forget gate, input modulation gate, the cell state and output gate at time k, respectively [31]:

$$f_k = \sigma(W_f \cdot [x_k, h_{k-1}] + b_f) \qquad i_k = \sigma(W_i \cdot [x_k, h_{k-1}] + b_i)$$
$$g_k = \tanh(W_g \cdot [x_k, h_{k-1}] + b_g) \qquad c_k = f_k \odot c_{k-1} + i_k \odot g_k$$
$$o_k = \sigma(W_o|cdot[x_k, h_{k-1}] + b_o) \qquad h_k = o_k \odot \tanh(c_k)$$

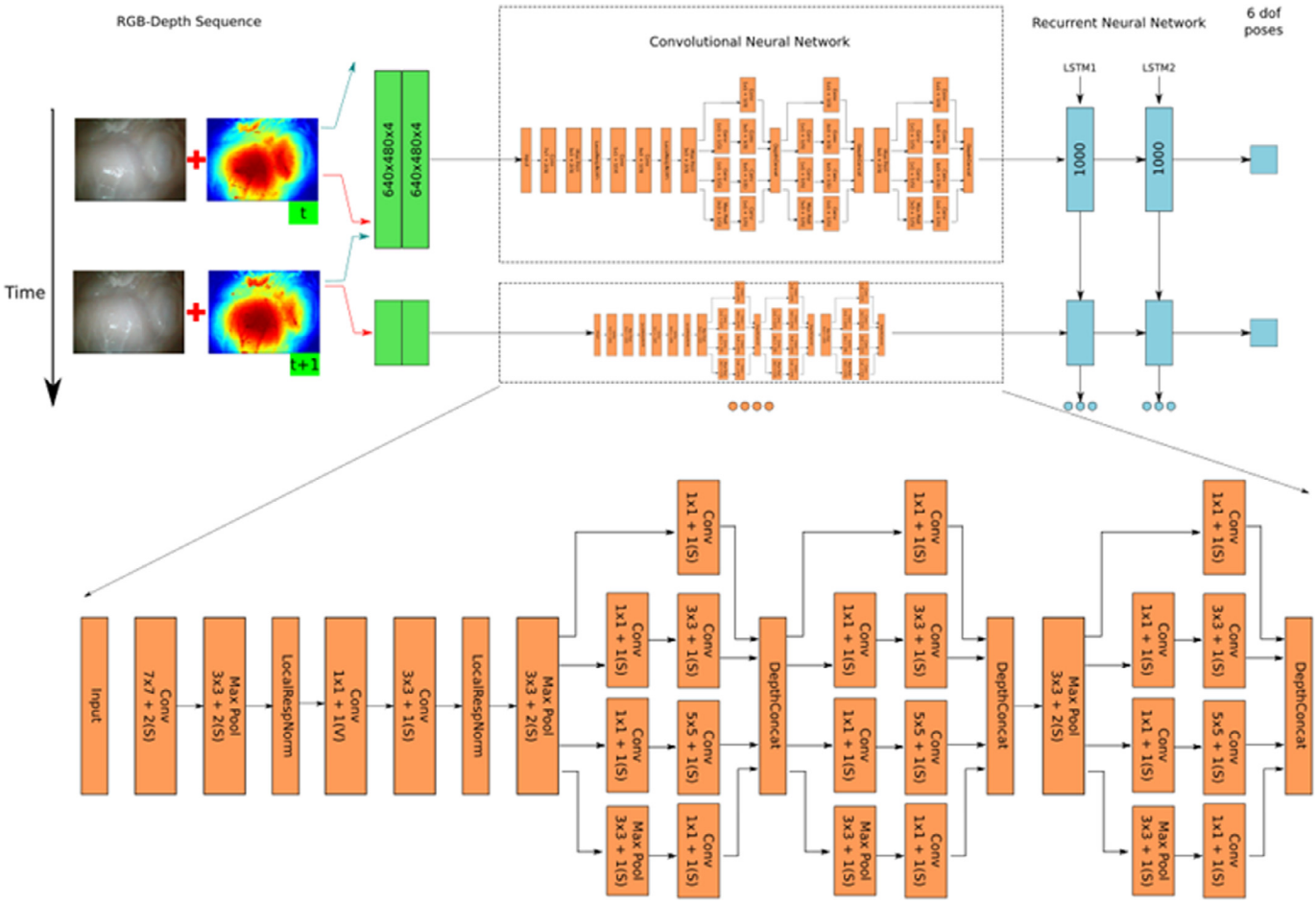


Fig. 4. Architecture of the proposed RCNN based monocular VO system.

Although the LSTM is prone to vanishing gradient problem of RNN and is capable to detect the long-term dependencies, its learning capacity can be increased further by stacking multiple LSTM layers vertically. Thus, our deep RNN consists of two LSTM layers with the output sequence of the first one forming the input sequence of the second one each containing 1000 hidden units, as illustrated in Fig. 4. The proposed system, which learns translational and rotational motions simultaneously to regress the 6-DoF pose, is trained on Euclidean loss using Adam optimization method with the following objective loss function:

$$\text{loss}(I) = \|\hat{x} - x\|_2 + \beta\|\hat{q} - q\|_2 \tag{1}$$

where x is the translation vector and q is the rotation vector. The pseudo-code to calculate the loss value is given in Algorithm 2. In our loss function, a balance β must be kept between the orientation and translation loss values which are highly coupled each other as they are learned from the same model weights. Experimental results show that the optimal β is given by the ratio between the loss values of predicted positions and orientations at the end of training session [30].

Algorithm 2 Pseudo code to calculate the loss over the network.

```
procedure CalculateLoss
    loss ← 0
    for layer in layers do
        for top, loss_weight in layer.tops, layer.loss_weights do
            loss ← loss + loss_weight × sum(top)
```

The back-propagation algorithm is used to calculate the gradients of RCNN weights, which are passed to the Adam optimization method to compute adaptive learning rates for each parameter employing the first-order gradient-based optimization of the stochastic objective function. In addition to saving exponentially decaying average of past squared gradients, v_t, Adam optimization keeps exponentially decaying average of past gradients, m_t that is similar to momentum. The update equations are given as

$$(m_t)_i = \beta_1(m_{t-1})_i + (1-\beta_1)(\nabla L(W_t))_i \tag{2}$$

$$(v_t)_i = \beta_2(v_{t-1})_i + (1-\beta_2)(\nabla L(W_t))_i^2 \tag{3}$$

$$(W_{t+1})_i = (W_t)_i - \alpha\frac{\sqrt{1-(\beta_2)_i^t}}{1-(\beta_1)_i^t}\frac{(m_t)_i}{\sqrt{(v_t)_i + \varepsilon}} \tag{4}$$

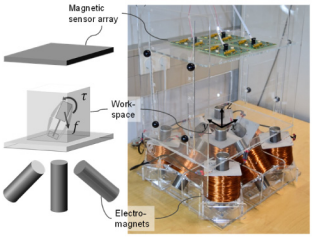
We used default values proposed by [34] for the parameters β_1, β_1 and ε: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$.

3. Dataset

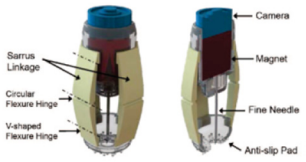
This section demonstrates the experimental setup of the proposed study, introduces our magnetically actuated soft capsule endoscopes (MASCE) and explains how the training and testing datasets were recorded.

3.1. Magnetically actuated soft capsule endoscopes (MASCE)

Our capsule prototype is a magnetically actuated soft capsule endoscope (MASCE) designed for disease detection, drug delivery

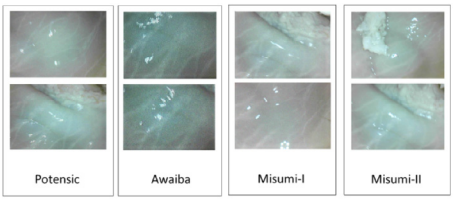


(a) Actuation system of the MASCE [19]

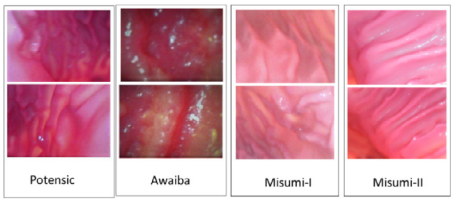


(b) Exterior (left) and section view (right) of MASCE [19]

Fig. 5. MASCE design features and actuation unit.



(a) Sample frames recorded on a real pig stomach



(b) Sample frames recorded on EGD simulator

Fig. 6. Sample frames from the datasets used in the experiments.

Table 1
Endoscopic camera specifications used for the experiments.

(a) Awaiba Naneye Endoscopic Camera		(b) Misumi-V3506-2ES Camera	
Resolution	250×250 pixel	Resolution	400×400 pixel
Footprint	$2.2 \times 1.0 \times 1.7$ mm	Diameter	8.2 mm
Pixel size	3×3 μm^2	Pixel size	5.55×5.55 μm^2
Frame rate	44 fps	Frame rate	30 fps
(c) Misumi-V3506-2ES Camera		(d) Potensic Mini Camera	
Resolution	640×480 pixel	Resolution	1280×720 pixel
Diameter	8.6 mm	Diameter	8.8 mm
Pixel size	6.0×6.0 μm^2	Pixel size	10.0×10.0 μm^2
Frame rate	30 fps	Frame rate	30 fps

and biopsy operations in the upper gastrointestinal tract. The prototype is composed of a RGB camera, a permanent magnet, a fine-needle and a drug chamber (see Fig. 5 for visual reference). The magnet exerts magnetic force and torque to the robot in response to a controlled external magnetic field [19]. The magnetic torque and forces are used to actuate the capsule robot and to release drug and deliver the needle through the hole in the bottom of the capsule. Magnetic fields from the electromagnets generate the magnetic force and torque on the magnet inside MASCE so that the robot moves inside the workspace. Sixty-four three-axis magnetic sensors are placed on the top, and nine electromagnets are placed in the bottom [19].

3.2. Training dataset

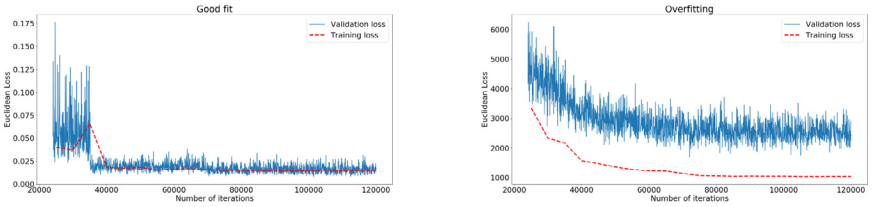
We created two groups of training datasets. The first training dataset was recorded on five different real pig stomachs (see Fig. 2), whereby the second dataset which was only used for training purposes, was captured using a non-rigid open GI tract model EGD (esophagus gastro duodenoscopy) surgical simulator LM-103 (see Fig. 2). To ensure that our algorithm is not tuned to a specific camera model, four different commercial endoscopic cameras were employed, specifications of which are shown in Table 1, accordingly. For each pig stomach-camera combination, 2000 frames were acquired which makes for four cameras and five pig stomachs 40, 000 frames, in total. Sample real pig stomach frames are shown in Fig. 6a for visual reference. As a second training dataset, for each of four cameras, we captured 10, 000 frames on an EGD human stomach simulator making 40, 000 frames, in total. Sample synthetic training frames are shown in Fig. 6b for visual reference. During video recording, Optitrack motion tracking system consisting of eight Prime-13 cameras and a tracking software was utilized to obtain 6-DoF localization ground truth data in a sub-millimeter precision (see Fig. 2) which was used as a gold standard for the evaluations of the pose estimation accuracy.

3.3. Testing dataset

We created a testing dataset recorded using five different real pig stomachs, which were not used for the training section. For each pig stomach-camera combination, 2000 frames are acquired making 40, 000 frames, in total. We did not capture any synthetic dataset for the testing session since it is less realistic due to obvious patterns of such artificial simulators. For all of the video records, again Optitrack motion tracking system was utilized to obtain 6-DoF localization ground truth.

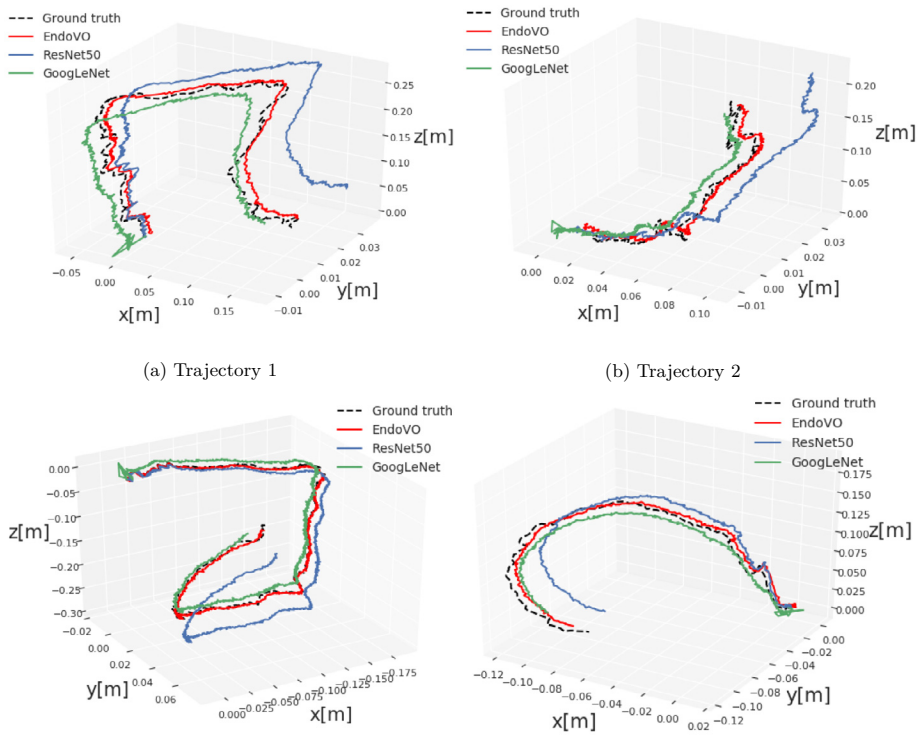
4. Evaluations and results

Architecture was trained using Caffe library and NVIDIA Tesla K40 GPU. Using back-propagation-through-time method, the weights of hidden units were trained for up to 200 epochs with an initial learning rate of 0.001. Overfitting meaning that the noise or random fluctuations in the training data are picked up and learned as concepts by the model, whereas these concepts do not apply to a new data and negatively affect the ability of the model to make generalizations, was prevented using dropout and early stopping techniques (see Fig. 10). Dropout regularization technique introduced by [35] is an extremely effective and simple method to avoid overfitting. It samples a part of the whole network and updates its parameters based on the input data. Early stopping is another widely used technique to prevent overfitting of a complex neural network architecture which was optimized by a gradient-based method. The approach is executed by splitting the dataset into a training and a validation set to evaluate the generalization capability of the model.



(a) Change in the loss values for a good fitting (b) Change in the loss values for overfitting

Fig. 7. The decrease in the training and validation loss values. In overfitting case, the training loss gets smaller than the validation loss. However, the loss values are balanced for a good fit.

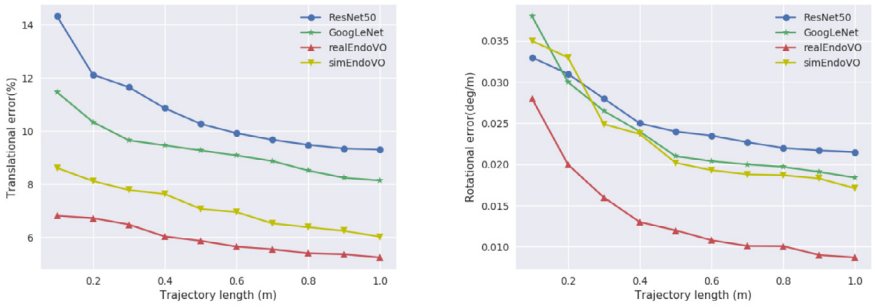


(a) Trajectory 1 (b) Trajectory 2

(c) Trajectory 3 (d) Trajectory 4

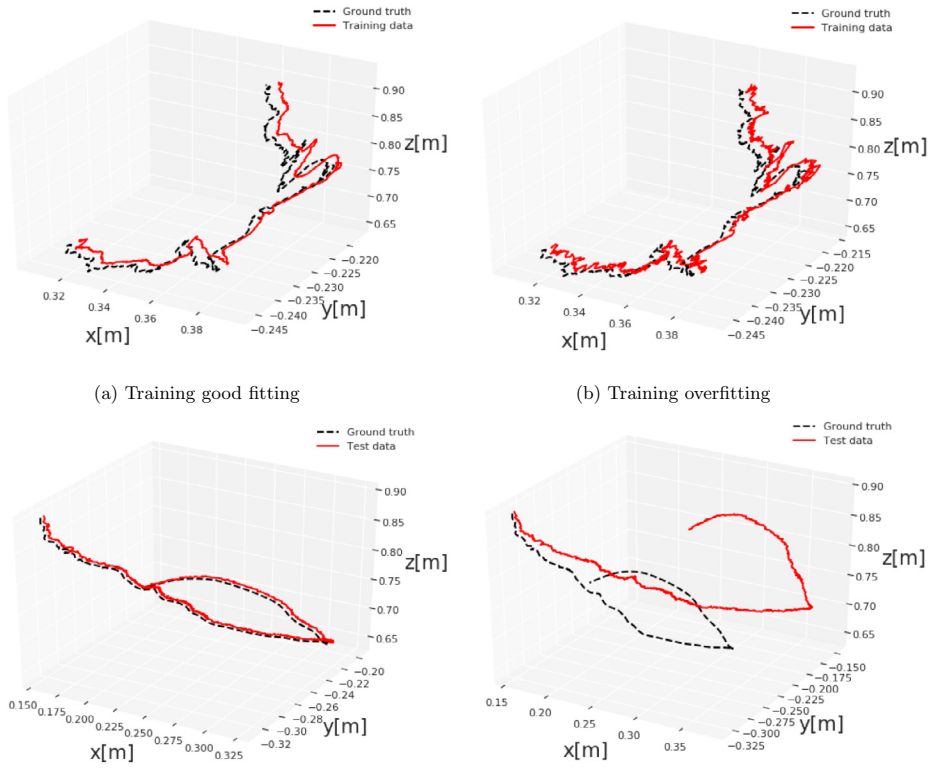
Fig. 8. Sample ground truth trajectories and estimated trajectories predicted by the DL based VO models. As seen, deep EndoVO is the closest to the ground truth trajectories. The scale is calculated and maintained correctly by the models.

For the testing sessions, only real pig stomach recordings were used to ensure real world conditions. Additionally, we strictly avoided to use any frame from the training session for the testing session. Two separate experiments were conducted, whereas training session of the first experiment was performed using only the synthetic training dataset (see Fig. 7b) which we call simEndoVO and training session of the second experiment was performed using frames from both synthetic and real pig stomach dataset (see Fig. 7b and a) which we call realEndoVO. The performance of the simEndoVO and realEndoVO approaches were analysed using averaged root mean square errors (RMSEs) for translational and rotational motions. For various trajectories with different complexity levels of motions, including uncomplicated paths with slow incremental translations and rotations, comprehensive scans with many local loop closures and complex paths with sharp rotational and translational movements, we performed testings on both simEndoVO and realEndoVO comparing them with GoogLeNet and ResNet50 architectures which were modified to regress 6-DoF pose values by removing softmax layer and integrating a fully-connected (FC) layer and an affine regressor layer. The average translational and rotational RMSEs for simEndoVO, realEndoVO, GoogLeNet and ResNet50 networks against different path lengths are shown in Fig. 9, respectively. The results depicted indicate, that realEndoVO clearly outperforms GoogLeNet and ResNet50, whereas simEndoVO slightly outperforms them. We presume that the effective use of LSTM in EndoVO architecture enabled learning motion dynamics across frame sequences, which is not feasible by architectures working with the principle of just-in-moment information processing; i.e. GoogleNet and ResNet50. The results in Fig. 9 also indicate that the training procedure including both simulator and real dataset was more informative than training only with simulator dataset. On the other hand, the accuracies achieved by the modified GoogLeNet are slightly better than accuracies achieved by the modified ResNet50, proving the superiority of inception layers over residual networks for feature extraction related tasks. Derived from RMSEs calculated, the rotational motion parameters seem to be more prone to overfitting compared to translational motion parameters (see Fig. 10 for visual reference). The reason for that observation could be the fact that inner organ scanning procedures generally contain more translational motions than rotational mo-



(a) Trajectory length vs translation error (b) Trajectory length vs totation error

Fig. 9. Deep EndoVO outperforms both of the other models in terms of translational and rotational position estimation.



(a) Training good fitting (b) Training overfitting

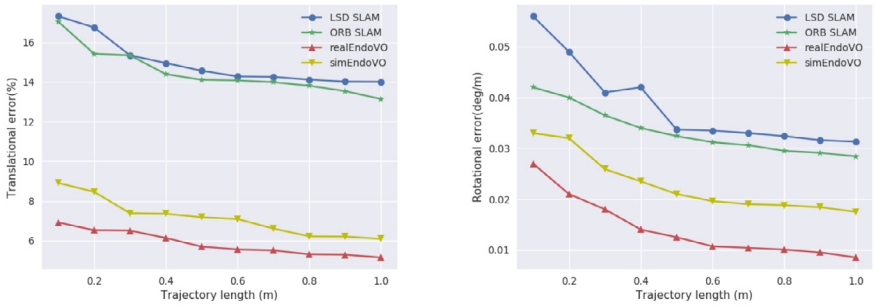
(c) Test good fitting (d) Test overfitting

Fig. 10. The affect of good fitting and overfitting. The first and the second rows show over-fitted and well-fitted models, respectively. As seen in subfigures, the model learns the details and noise in the training data to an undesired extent that it negatively impacts the performance of the model on the test data.

tions resulting in a better learning for translations. As the length of the trajectory increases, both the translational and rotational error of all the proposed models significantly decrease (see Fig. 9). Some sample ground truth and estimated trajectories for realEndoVO, GoogLeNet and ResNet50 are shown in Fig. 8 for visual reference. As seen in these sample trajectories, realEndoVO is able to stay close to the ground truth pose values for even sharp crispy motions, contrary to realEndoVO; GoogLeNet and ResNet50 path estimations which deviate drastically from the ground truth path values. Even for very fast and challenge paths such as Fig. 8a and c, the deviations of realEndoVO from the ground truth still remain in an acceptable range for medical operations. In addition to that, it is clearly seen that all of the three evaluated neural network architectures are able to estimate the scale very accurately without using any prior information or post alignment techniques contrary to traditional VO. Solving the scale ambiguity for monocular camera based VO makes our proposed DL based method more beneficial than traditional VO approach. As opposed to the traditional VO pipeline (see Fig. 1), the DL-based VO do not require any explicit feature extraction, matching, outlier detection or multi-scale bundle adjustment-like parameter tuning requiring operations, which can be seen as further benefits of the proposed approach.

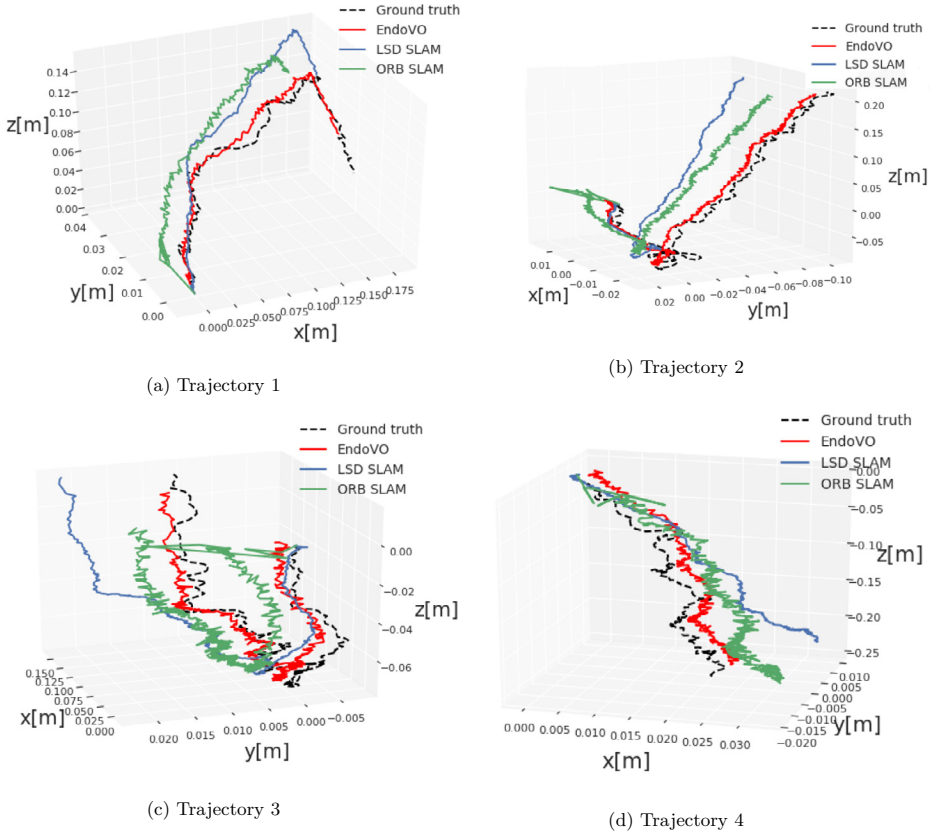
4.1. Comparisons of deep EndoVO with state-of-the-art SLAM methods

In this subsection, we compare the performance of the proposed deep EndoVO with two of the widely used state-of-the-art SLAM methods; i.e. large-scale direct monocular SLAM (LSD SLAM) [36] and the oriented fast and rotated brief SLAM (ORB SLAM)



(a) Trajectory length vs translation error (b) Trajectory length vs rotation error

Fig. 11. Deep EndoVO outperforms the state-of-the-art SLAM methods ORB SLAM and LSD SLAM in both the translation and orientation estimation.



(a) Trajectory 1 (b) Trajectory 2

(c) Trajectory 3 (d) Trajectory 4

Fig. 12. The ground truth and the trajectory plots acquired via deep EndoVO, LSD SLAM and ORB SLAM. Deep EndoVO is the closest to the ground truth trajectories compared to the state-of-the-art SLAM methods.

[37]. LSD SLAM is a direct image alignment-based method which optimizes the geometry using all of the image intensities. In addition to higher accuracy and robustness particularly in environments with little key points, this provides substantially more information about the geometry of the environment, which can be very valuable for medical robot applications, as well. ORB SLAM on the other hand, relies on feature point extraction and tracking to estimate camera pose and 3D map the environment. Even though it gives very promising results for feature-rich areas, its main deficiency appears once the robot enters poorly featured areas. Tracking failures are commonly observable for poorly featured GI tract tissues making ORB SLAM less proper for our case. We believe that our deep EndoVO architectures makes an optimal use of both direct and feature point information to estimate the pose. The average translational and rotational RMSEs for simEndoVO, realEndoVO, LSD SLAM and ORB SLAM, shown in Fig. 11 indicate that both simEndoVO and realEndoVO clearly outperform LSD SLAM and ORB SLAM in terms of pose accuracy. Sample trajectory estimations shown in Fig. 12 visualize clearly that the tracking capability of the proposed deep EndoVO is much more robust and reliable compared to LSD SLAM and ORB SLAM. In many parts of the trajectories, ORB SLAM and LSD SLAM deviate from the ground truth trajectory drastically, whereas deep EndoVO is still able to stay close to the ground truth values even for most challenge trajectory sections (see Fig. 12b and c).

5. Conclusion

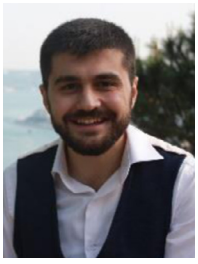
In this study, we presented, to the best of our knowledge, the first deep VO method for endoscopic capsule robot and standard hand-held endoscope operations. The proposed system is able to achieve simultaneous representation learning and sequential mod-

elling of motion dynamics across frames by concatenating the inception modules with RNN layers. Many issues faced by traditional VO techniques such as feature correspondence establishment in low textured areas, high reflections, motion blur and low image quality are handled by the proposed deep EndoVO successfully. Since it is trained in an end-to-end manner, there is no need to carefully fine-tune the parameters of the system. As a future step, we consider to combine deep EndoVO with some functionalities from the traditional VO pipelines such as RANSAC for outlier detection and bundle fusion for globally consistent pose estimation etc to avoid drifts. Moreover, we consider to develop a stereo version of the proposed deep EndoVO approach.

References

[1] Z. Liao, R. Gao, C. Xu, Z.-S. Li, Indications and detection, completion, and retention rates of small-bowel capsule endoscopy: a systematic review, Gastrointestinal Endosc. 71 (2) (2010) 280–286.
[2] T. Nakamura, A. Terano, Capsule endoscopy: past, present, and future, J. Gastroenterol. 43 (2) (2008) 93–99.
[3] G. Pan, L. Wang, Swallowable wireless capsule endoscopy: progress and technical challenges, Gastroenterol. Res. Pract. 2012 (2012), Article ID 841691, 9 pages.
[4] T.D. Than, G. Alici, H. Zhou, W. Li, A review of localization systems for robotic endoscopic capsules, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2387–2399.
[5] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, E. Diller, Biomedical applications of untethered mobile milli/microrobots, Proc. IEEE 103 (2) (2015) 205–224.
[6] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, M. Sitti, A non-rigid map fusion-based RGB-depth slam method for endoscopic capsule robots, arXiv preprint, arXiv:1705.05444 (2017).
[7] M. Turan, Y. Almalioglu, E. Konukoglu, M. Sitti, A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots, arXiv preprint, arXiv:1705.05435 (2017).
[8] M. Turan, A. Abdullah, R. Jamiruddin, H. Araujo, E. Konukoglu, M. Sitti, Six degree-of-freedom localization of endoscopic capsule robots using recurrent neural networks embedded into a convolutional neural network, arXiv preprint, arXiv:1705.06196 (2017).
[9] M. Turan, Y.Y. Pilavci, R. Jamiruddin, H. Araujo, E. Konukoglu, M. Sitti, A fully dense and globally consistent 3d map reconstruction approach for GI tract to enhance therapeutic relevance of the endoscopic capsule robot, arXiv preprint arXiv:1705.06524 (2017).
[10] M. Turan, Y.Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, M. Sitti, Sparse-then-dense alignment based 3d map reconstruction method for endoscopic capsule robots, arXiv preprint, arXiv:1708.09740 (2017).
[11] M.K. Goenka, S. Majumder, U. Goenka, Capsule endoscopy: present status and future expectation, World J. Gastroenterol. 20 (29) (2014) 10024–10037.
[12] F. Munoz, G. Alici, W. Li, A review of drug delivery systems for capsule endoscopy, Adv. Drug Delivery Rev. 71 (2014) 77–85.
[13] F. Carpi, N. Kastelein, M. Talcott, C. Pappone, Magnetically controllable gastrointestinal steering of video capsules, IEEE Trans. Biomed. Eng. 58 (2) (2011) 231–234.
[14] H. Keller, A. Juloski, H. Kawano, M. Bechtold, A. Kimura, H. Takizawa, R. Kuth, Method for navigation and control of a magnetically guided capsule endoscope in the human stomach, in: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), IEEE, 2012, pp. 859–865.
[15] A.W. Mahoney, S.E. Wright, J.J. Abbott, Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet, in: 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 5366–5371.
[16] S. Yim, E. Gultepe, D.H. Gracias, M. Sitti, Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers, IEEE Trans. Biomed. Eng. 61 (2) (2014) 513–521.
[17] A.J. Petruska, J.J. Abbott, An omnidirectional electromagnet for remote manipulation, in: 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 822–827.
[18] D. Son, S. Yim, M. Sitti, A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors, IEEE/ASME Trans. Mechatron. 21 (2) (2016) 708–716.
[19] D. Son, M.D. Dogan, M. Sitti, Magnetically actuated soft capsule endoscope for fine-needle aspiration biopsy, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 1132–1139.
[20] M. Fluckiger, B.J. Nelson, Ultrasound emitter localization in heterogeneous media, in: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007, pp. 2867–2870.
[21] J.M. Rubin, H. Xie, K. Kim, W.F. Weitzel, S.Y. Emelianov, S.R. Aglyamov, T.W. Wakefield, A.G. Urquhart, M. ODonnell, Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans, J. Ultrasound Med. 25 (9) (2006) 1179–1186.
[22] K. Kim, L.A. Johnson, C. Jia, J.C. Joyce, S. Rangwalla, P.D. Higgins, J.M. Rubin, Noninvasive ultrasound elasticity imaging (UEI) of crohn's disease: animal model, Ultrasound Med. Biol. 34 (6) (2008) 902–912.
[23] S. Yim, M. Sitti, 3-d localization method for a magnetically actuated soft capsule endoscope and its applications, IEEE Trans. Robot. 29 (5) (2013) 1139–1151.
[24] T. Ping-Sing, M. Shah, Shape from shading using linear approximation, Image Vision Comput. 12 (8) (1998) 487–498.
[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2014). 1409.1556.
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[28] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, M. Milford, On the performance of convnet features for place recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 4297–4304.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.
[30] A. Kendall, M. Grimes, R. Cipolla, Posenet: a convolutional network for real-time 6-dof camera relocalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2938–2946.
[31] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to Forget: Continual Prediction with LSTM, 1999.
[32] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, D. Cremers, Image-based localization with spatial LSTMs, arXiv preprint, arXiv:1611.07890 (2016).
[33] S. Wang, R. Clark, H. Wen, N. Trigoni, Deepvo: towards end-to-end visual odometry with deep recurrent convolutional neural networks, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2043–2050.
[34] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980(2014).
[35] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[36] J. Engel, T. Schöps, D. Cremers, Lsd-slam: large-scale direct monocular slam, in: European Conference on Computer Vision, Springer, 2014, pp. 834–849.
[37] R. Mur-Artal, J. Montiel, J.D. Tardós, Orb-slam: a versatile and accurate monocular slam system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.

Mehmet Turan received his diploma degree from the Information technology and Electronics Engineering Department of RWTH Aachen, Germany in 2012. He was a research scientist at UCLA (University of California Los Angeles) between 2013–2014 and a research scientist at the Max Planck Institute for Intelligent Systems between 2014–present. He is currently enrolled as a Ph.D. student at the ETH Zurich, Switzerland. He is also affiliated with Max Planck-ETH Center for Learning Systems, the first joint research center of ETH Zurich and the Max Planck Society. His research interests include SLAM (simultaneous localization and mapping) techniques for milli-scale medical robots and deep learning techniques for medical robot localization and mapping. He received DAAD fellowship between years 2005–2011 and Max Planck Fellowship between 2014–present. He has also received MPI-ETH Center fellowship between 2016–present.



Yasin Almalioglu received the B.Sc. degree with honours in computer engineering from Bogazici University, Istanbul, Turkey in 2015. He was a research intern at CERN Geneva, Switzerland and Astroparticle and Neutrino Physics Group at ETH Zurich, Switzerland in 2013 and 2014, respectively. He is currently pursuing the M.Sc. degree in computer engineering at Bogazici University, Istanbul, Turkey. His research interests include machine learning, Bayesian statistics, Monte Carlo methods, probabilistic graphical models, artificial neural networks and mobile robot localization. He received the Engin Arik Fellowship in 2013.

Helder Araujo is a professor at the Department of Electrical and Computer Engineering of the University of Coimbra. His research interests include computer vision applied to robotics, robot navigation and visual servoing. In the last few years he has been working on non-central camera models, including aspects related to pose estimation, and their applications. He has also developed work in active vision, and on control of active vision systems. Recently he has started work on the development of vision systems applied to medical endoscopy.

Ender Konukoglu, Ph.D., finished his Ph.D. at INRIA Sophia Antipolis in 2009. From 2009 till 2012 he was a post-doctoral researcher at Microsoft Research Cambridge. From 2012 till 2016 he was a junior faculty at the Athinoula A. Martinos Center affiliated to Massachusetts General Hospital and Harvard Medical School. Since 2016 he is an assistant professor of Biomedical Image Computing at ETH Zurich. His is interested in developing computational tools and mathematical methods for analysing medical images with the aim to build decision support systems. He develops algorithms that can automatically extract quantitative image-based measurements, statistical methods that can perform population comparisons and biophysical models that can describe physiology and pathology.

Dr. Metin Sitti received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1992 and 1994, respectively, and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. He was a research scientist at UC Berkeley during 1999–2002. He has been a professor in the Department of Mechanical Engineering and Robotics Institute at Carnegie Mellon University, Pittsburgh, USA since 2002. He is currently a director at the Max Planck Institute for Intelligent Systems in Stuttgart. His research interests include small-scale physical intelligence, mobile microrobotics, bio-inspired materials and miniature robots, soft robotics, and micro-/nanomanipulation. He is an IEEE fellow. He received the SPIE Nanoengineering Pioneer Award in 2011 and NSF CAREER Award in 2005. He received many best paper, video and poster awards in major robotics and adhesion conferences. He is the editor-in-chief of the Journal of Micro-Bio Robotics.

Unsupervised Odometry and Depth Learning for Endoscopic Capsule Robots

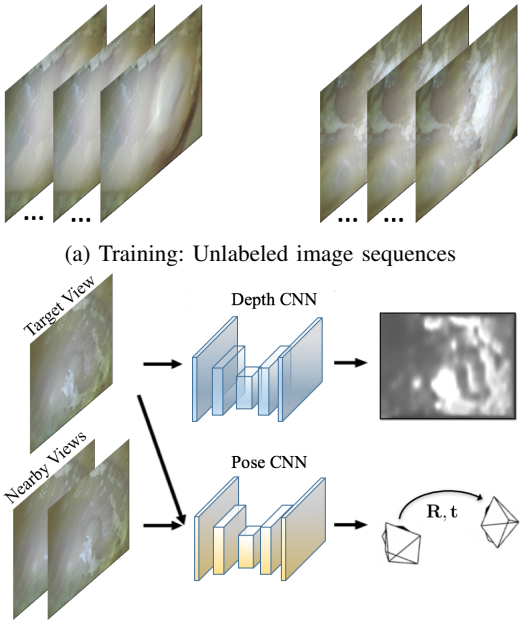
Mehmet Turan[1], Evin Pinar Ornek[2], Nail Ibrahimli [2], Can Giracoglu[2], Yasin Almalioglu[3], Mehmet Fatih Yanik [4], and Metin Sitti[5]

***Abstract*— In the last decade, many medical companies and research groups have tried to convert passive capsule endoscopes as an emerging and minimally invasive diagnostic technology into actively steerable endoscopic capsule robots which will provide more intuitive disease detection, targeted drug delivery and biopsy-like operations in the gastrointestinal(GI) tract. In this study, we introduce a fully unsupervised, real-time odometry and depth learner for monocular endoscopic capsule robots. We establish the supervision by warping view sequences and assigning the re-projection minimization to the loss function, which we adopt in multi-view pose estimation and single-view depth estimation network. Detailed quantitative and qualitative analyses of the proposed framework performed on non-rigidly deformable ex-vivo porcine stomach datasets proves the effectiveness of the method in terms of motion estimation and depth recovery.**

I. Introduction

Advancements in various fields of science and technology in the last decade has opened new pathways for non-invasive examination of patient's body and detailed investigation about diseases. Hospitals are using innovative ways to provide accurate data from inside of the human body. As an emerging example, various diseases such as colorectal cancer and inflamatory bowel disease are diagnosed by the usage of swallowable capsule endoscopes, which are non-invasive, painless, suitable to be used for long duration screening purposes which can access difficult body parts (e.g.,small intestines) better than standard endoscopy. Such benefits make swallowable, non-tethered capsule endoscopes an exciting alternative over standard endoscopy [1], [2].

Current capsule endoscope technology employed in GI tract monitoring and disease detection consists of passive devices which are locomated by random peristaltic motions. The doctor would have an easier access to fine-scale body parts and could make more intuitive and correct diagnosis in case of a precise and reliable control over the position of the capsule. Many research groups attempted to build remotely controllable active endoscopic capsule robot systems with additional functionalities such as local drug delivery, biopsy and other medical functions [2]–[19], which are, on the other hand, heavily dependent on a real-time and precise pose estimation capability.



(a) Training: Unlabeled image sequences

(b) Testing: Pose and depth prediction

Fig. 1: Unsupervised training approach consists of two separate neural networks, one for depth prediction and another one for multi-view pose estimation. It requires unlabeled image sequences from different temporal points to establish a supervision basis. Models produce pose estimation between two views from different perspectives parameterized as 6-DoF motion, and depth prediction as a disparity map for a given view.

In this work, we propose a novel real-time localization and depth estimation approach for endoscopic capsule robots which mimic the remarkable ego-motion estimation and scene reconstruction capabilities of human beings by training an unsupervised deep neural network. The proposed network consists of two simultaneously trained sub networks, the first one assigned for depth estimation via encoder-decoder strategy, the second assigned to regress the camera pose in 6-DoF. The model observes sequences of monocular images and aims to interpret them to estimate executed camera motion in 6-DoF and the depth map of the observed scene as shown in Fig. 1. Our framework estimates the camera motion and depth information in an end-to-end and unsupervised

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

[2]Evin Pinar Ornek, Nail Ibrahimli, Can Giracoglu is with the Informatics Faculty, Technical University of Muenich, Germany evin.oernek, nail.ibrahimli, can.giracoglu@tum.de

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

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

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

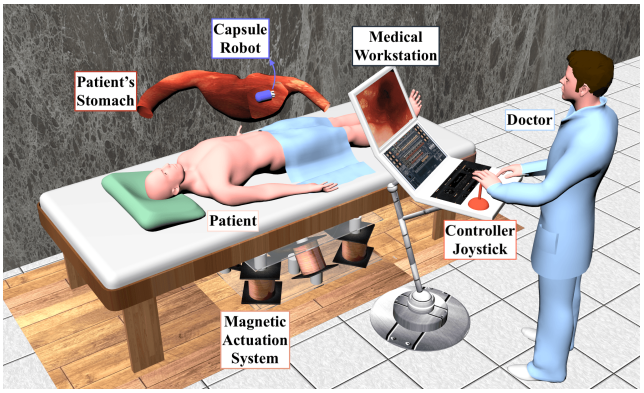


Fig. 2: Demonstration of the active endoscopic capsule robot operation using MASCE (Magnetically actuated soft capsule endoscope) designed for disease detection, drug delivery and biopsy-like operations in the upper GI-tract. MASCE is composed of a RGB camera, a permanent magnet, an empty space for drug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

fashion directly from input pixels. Training is performed using only unlabeled monocular frames in a similar way to prior works such as [20]–[22].

We formulate the entire pose estimation and map reconstruction pipeline for endoscopic capsule robots as a consistent and systematic learning concept which can improve its performance every day by collecting streamed data belonging to numerous patients undertaken to endoscopic capsule robot and standard endoscopy investigations in hospitals over the world. This way, we want to mimic and transfer a continuous learning functionality from medical doctors into medical robots domain, where experience and adaptation to unexpected novel situations can be much more critical to real-world scenarios.

To summarize, main contributions of our paper are as follows:

- To best of our knowledge, this is the first unsupervised odometry and depth estimation approach for both the endoscopic capsule robots and hand-held standard endoscopes.
- Since the network learns in a fully unsupervised manner, no ground truth pose and/or depth values are required to train the neural network.
- Neither prior knowledge nor parameter tuning is needed to recover the trajectory and depth, contrary to traditional visual odometry(VO) and deep learning(DL) based supervised odometry approaches.
- We simultaneously train a reliability mask which identifies pixels distorted by camera occlusions, non-rigid organ deformations and/or non-Lambertian surface. Such a mask is very crucial for vision based methods applied on endoscopic type of images since occlusions, non-rigid deformations and specularities violating Lambertian surface properties commonly occur in endoscopic types of images.

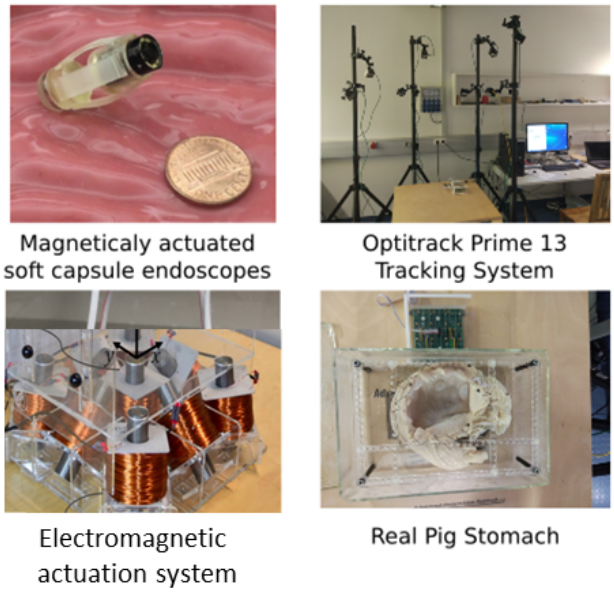


Fig. 3: Illustration of the experimental setup. MASCE is a magnetically actuated robotic capsule endoscope prototype which has a ringmagnet on the body. An electromagnetic coil array consisting of nine coils is used for the actuation of the MASCE. The ringmagnet exerts magnetic force and torque on the capsule in response to the external magnetic field provide by the electromagnetic coil array. Magnetic torque and forces are also used to release drug, as well. OptiTrack system consisting of eight infrared cameras is employed for the ground truth pose estimation. An opened and oiled porcine stomach simulator is used to represent human stomach.

Evaluations we made on non-rigidly deformable porcine stomach videos prove the success of our depth estimation and localization approach. As the outline of this paper, the previous work in endoscopic capsule odometry is discussed in Section II. Section III introduces the proposed method with its mathematical background in detail and the unsupervised DL architecture. Section IV shows our experimental quantitative and qualitative results achieved for 6-DoF localization and depth recovery. Finally, Section V mentions some bottlenecks and gives future directions for our project. Our code will be made available at https://github.com/mpi/deep-unsupervised-endovo.

II. Background

In the last decade, several localization methods [23]–[27] were proposed to calculate the 3D position and orientation of the endoscopic capsule robot such as fluoroscopy [23],

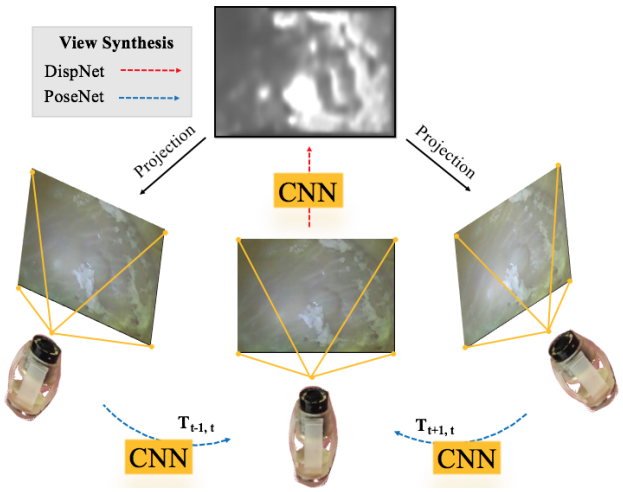


Fig. 4: Training input consists of sequential images from different perspectives, which are noted by $< I_{t-1}, I_t, I_{t+1} >$. After view synthesis creates the supervision baseline, PoseNet is trained to estimate relative motion change between $< I_{t-1}, I_t >$ and $< I_t, I_{t+1} >$, whereas DispNet learns to predict depth for the target image $< I_t >$.

ultrasonic imaging [24]–[27], positron emission tomography (PET) [23], [27], magnetic resonance imaging (MRI) [23], radio transmitter based techniques and magnetic field based techniques. The common drawback of these localization methods is that they require extra sensors and hardware design. Such extra sensors have their own drawbacks and limitations if it comes to their application in small scale medical devices such as space limitations, cost aspects, design incompatibilities, biocompatibility issues and the interference of the sensors with the activation system of the device.

As a solution of these issues, a trend of VO methods have attracted the attention for endoscopic capsule localization. A classic VO pipeline typically consists of many hand-engineered parts such as camera calibration, feature detection, feature matching, outliers rejection (e.g. RANSAC), motion estimation, scale estimation and global optimization (bundle adjustment). Although some state-of-the-art algorithms based on this traditional pipeline have been developed and proposed for endoscopic VO task in the past decades, their main deficiencies such as tracking failures in low textured areas, sensor occlusion issues, lack of handling non-rigid organ deformation still remain. In last couple of years, DL techniques have been dominating many computer vision related tasks with numerous promising result, e.g. object detection, object recognition, classification problems etc. Contrary to these high-level computer vision tasks, VO is mainly working on motion dynamics and relations across sequence of images, which can be defined as a sequential learning problem.

Our proposed method solves several issues faced by typical VO pipelines, e.g the need to establish a frame-to-frame feature correspondence, vignetting artefacts, motion blur, specularity or low signal-to-noise ratio (SNR). We think that DL based endoscopic VO approach is more suitable for such challenge areas since the operation environment(GI tract) has similar organ tissue patterns among different patients which can be learned by a sophisticated machine learning approach easily. Even the dynamics of common artefacts such as non-rigidness, sensor occlusions, vignetting, motion blur and specularity across frame sequences could be learned and used for a better pose estimation, whereas our unsupervised odometry learning method additionally solves the common problem of missing labels on medical datasets from inner body operations [4], [6].

III. METHOD

Different from supervised VO learning [2], [4], [6], where camera poses and/or depth ground truths are required to train the neural network, the core idea underlying our unsupervised pose and depth prediction method is to make use of the view synthesis constraint as the supervision metric, which forces the neural network to synthesize target image from multiple source images acquired from different camera poses. This synthesis is performed using estimated depth image, estimated target camera pose values in 6-DoF and nearby color values from source images. In addition, a reliability mask is trained to detect sensor occlusions, non-rigid deformations of the soft organ tissue and lack of textures inside the explored organ.

A. View synthesis as supervision metric

To provide a supervision to the neural network, view synthesis is accomplished by training with consecutive images. As input, we take a sequence of 3 consecutive frames, and choose the middle frame as a target frame. Sequences are denoted by $< I_{t-1}, I_t, I_{t+1} >$ where I_t is the target view and rest of images are source views $I_s =< I_{t-1}, I_{t+1} >$, which are used to render the target image (see Fig. 4). The objective function of the view synthesis is:

$$\mathscr{L}_{vs} = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)| \tag{1}$$

where p is pixel coordinate, and $\hat{I}_s$ is the source view I_s warped to the target view making use of the estimated depth image $\hat{D}_t$ and 4×4 camera transformation matrix $\hat{T}_{t \to s}$ [29]. Let p_t represent the homogeneous pixel coordinates in the target view, and K be the camera intrinsics matrix.

p_t is projected coordinate on the source view and p_s is acquired by:

$$p_s \sim K \hat{T}_{t \to s} \hat{D}_t(p_t) K^{-1} p_t \tag{2}$$

Note that the value of p_s is not discrete. To find the expected intensity value at that position, bilinear interpolation among four discrete neighbors of p_s is used [30]:

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i \in \{top, bottom\}, j \in \{left, right\}} w^{ij} I_s(p_s^{ij}) \tag{3}$$

Let w^{ij} be the proximity value between projected and neighboring pixels summing up to one and $\hat{I}_s$ be the estimated mean intensity for projected pixel p_s.

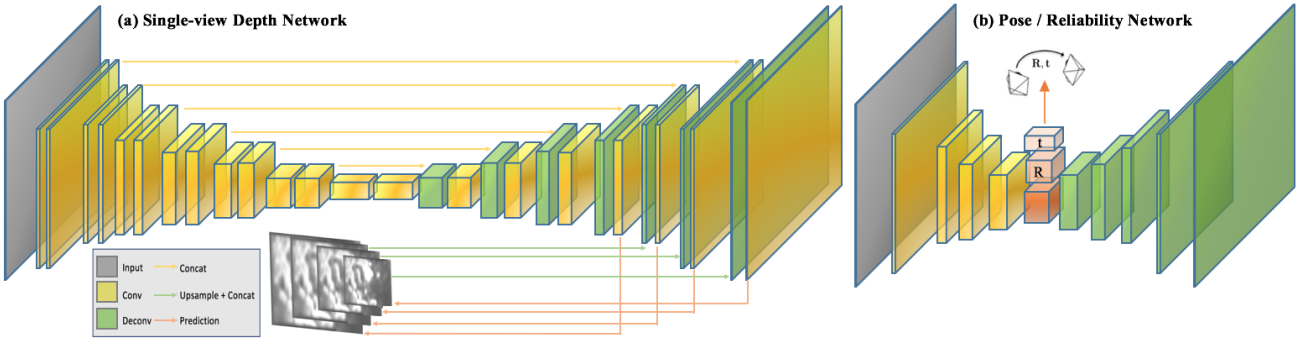


Fig. 5: The proposed neural network architecture for pose/reliability/depth map estimation. The width and height of illustrated blocks reflect the spatial dimensions of layers and output channels which are based on an encoder-decoder design. (a) Single-view depth prediction model is adopted by DispNet [28]. ReLu activations follow the middle convolution layers. Kernel size for first four layers are 7, 7, 5, 5 respectively, and rest of the layers have kernel size 3. (b) Pose/reliability estimation network is motivated by SFM-Learner [21] model and it has decoder-encoder design, as well. The encoder part has five feature extraction layers which are shared for both pose and reliability mask estimation. The pose results are gathered after the encoder network, which has $6*(N-1)$ output channels for 6-DoF motion parameters. The encoder part is followed by a decoder, which has 5 deconvolutional layers, consisting ReLU activations in between.

View synthesis approach assumes that camera sensor is not occluded, non-rigid deformations are avoided and explored organ surface obeys Lambertian surface rules enabling photometric error minimization between target and source views. These assumptions are frequently violated in endoscopic type of videos:

1) Sensor occlusions occur often due to peristaltic organ motions.
2) Inner organs have in general a non-rigid structure meaning deformations cannot be completely avoided.
3) Organ fluids cause specularities which violate the Lambertian surface rules.

To overcome these, we trained a soft reliability mask which labels each target-source pixel pair as reliable to be used for view-synthesis or believed to violate assumptions because of being affected by occlusions, non-rigid deformations and/or specularities. Incorporating the soft-reliability mask $\hat{E}_s$, the view synthesis equation is updated as:

$$\mathscr{L}_{vs} = \sum_{<I_{t-1},I_{t+1}>\in\mathscr{S}} \sum_{p} \hat{E}_s(p)|I_t(p) - \hat{I}_s(p)| \ . \qquad (4)$$

Minimizing this energy function without regularizer will force mask to be zero across the whole image domain. To overcome this problem and obtain a reasonable mask, a regularization term is to use which describes the prior knowledge about reliability mask. Hence, let $\mathscr{L}_{reg}(\hat{E}_s^l)$ be the regularization term that minimizes the cross-entropy loss and prevents trivial solutions. Finally, since gradients are derived from differences between four neighbors and corresponding pixel intensities of source and target frames, a smoothness loss $\mathscr{L}_{smooth}^l$ is needed. The multiscale pyramid and smoothness loss for gradients are extracted from larger spatial regions. This leads to the following energy function:

$$\mathscr{L}_{final} = \sum_l \mathscr{L}_{vs}^l + \lambda_s \mathscr{L}_{smooth}^l + \lambda_e \sum_s \mathscr{L}_{reg}(\hat{E}_s^l) \qquad (5)$$

Here, s indexes source images, l indexes images from different scales, λ_s is the regularization weight for depth smoothness, and λ_e is the weight for reliability mask.

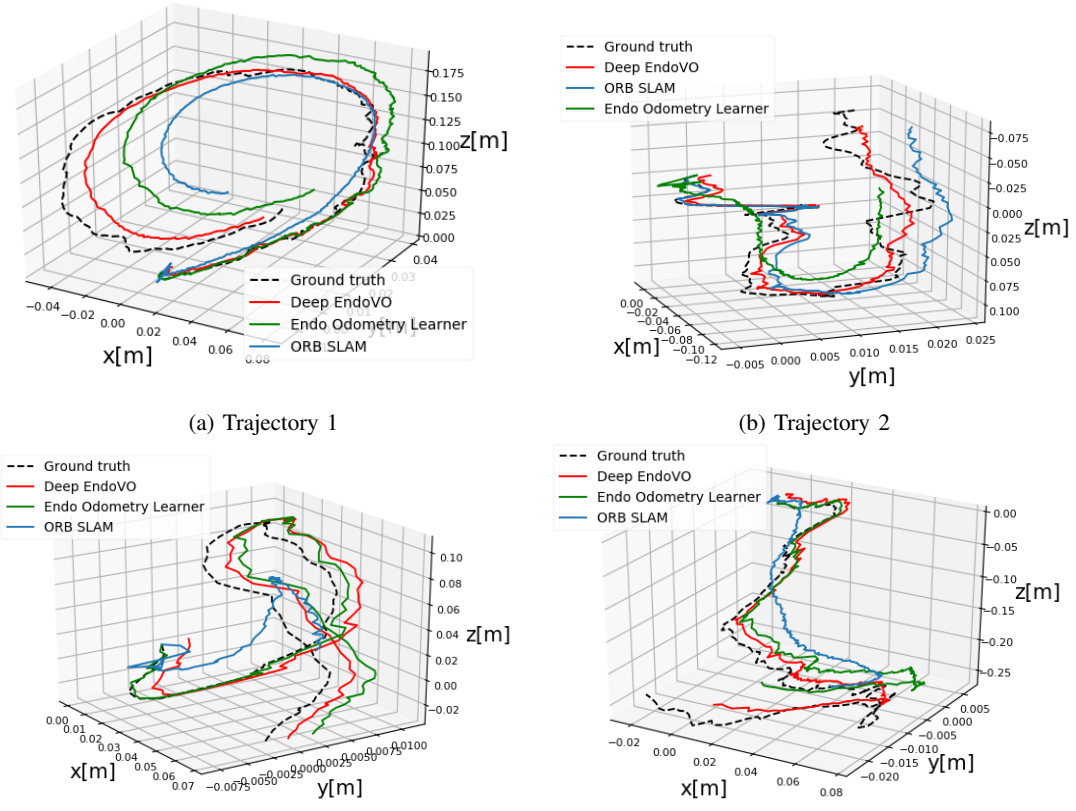
B. Network architecture

As mentioned earlier, our problem is estimating odometry in a textureless scene by using only sequenced RGB frames as input. Since classical methods fail to cope with this problem, we use DL methods where we get our motivation from recent works [31] and [21] which propose improvements by autoencoder based architectures. Our overall DL model as shown in Fig. 5 consists of two end-to-end frameworks.

The first architecture is employed to predict single-view depths by creating disparity map outputs. The encoder-decoder convolutional layers are followed by a prediction layer, whose outputs are constrained by $1/(\alpha*sigmoid(x)+\beta)$ with $\alpha = 10$ and $\beta = 0.1$ to ensure that predictions occur in a desirable interval.

The second network tries to estimate relative pose, parameterized by SE(3) motions between views, and the reliability mask. The encoder part for pose estimation and reliability mask are same, where they share weights in the first five feature extractor convolutional layers and divide into two tracks afterwards. Pose is estimated by encoder's $6*(N-1)$ channels, as translation and rotation parameters. The decoder part consists of five deconvolutional layers and generates multiscale mask predictions. There are four output channels for each prediction layer, and each two of them predict the reliability for input source-target pairs by softmax normalization.

Both networks are trained and optimized jointly. On the other hand, both networks can be tested and evaluated independently. Testing and training pipelines are illustrated in Fig. 1.



(a) Trajectory 1

(b) Trajectory 2

(c) Trajectory 3

(d) Trajectory 4

Fig. 6: Sample trajectories comparing the unsupervised learning method with ORB SLAM, EndoVO and OptiTrack ground truth in millimetric scale. Deep EndoVO shows the best odometry estimations, whereas ORB SLAM fails to track some fine-scale motions. Tracking performance of unsupervised odometry lies inbetween of ORB SLAM and Deep EndoVO; many fine-scale motions are successfully caught in detail, however there is still a certain amount of drift.

IV. EVALUATION AND RESULTS

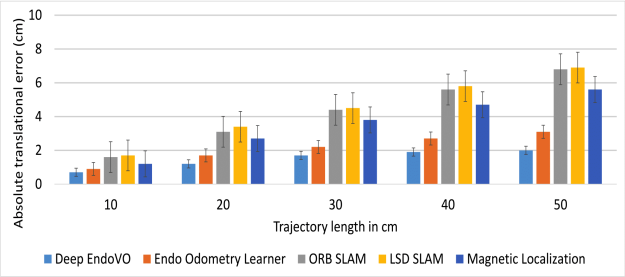
A. Dataset and Transfer learning

We used transfer learning to have an initialization for neural network weights since we lack huge amounts of labeled data. For pretraining, DL model proposed by Zhou et al. [21] is employed. The model is implemented with publicly available Tensorflow framework and pretrained with the KITTI dataset. Batch normalization is used for all of the layers except the outputs. Adam optimization is chosen to increase the convergence rate, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 0.1 and mini-batch size of 8. We used the model which was trained with $50K$ images and converged after 150K iterations. The model requires sequential images with size 128 x 416. On top of the model pretrained by a KITTI dataset, we fine-tuned the architecture with our domain data from endoscopic capsule robot by employing a GeForce GTX 1070 model GPU. Our dataset was collected in an experimental setup for an ex-vivo parcine stomach shown in Fig. 3 and it contains $12K$ frames with ground truth odometry obtained by OptiTrack visual tracking system. In this experiment, we fix the length of input image sequences into three frames. We used $10K$ frames for training, $1K$ for cross validation and $1K$ for evaluation and testing.

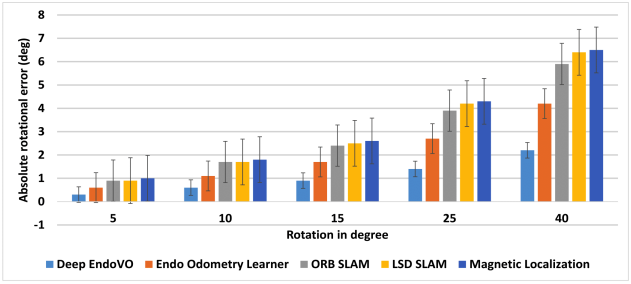
B. Pose estimation and Odometry benchmark

Our pose estimation network is tested with $1K$ frames. The network outputs the pose predictions as 6-DoF motion (Euclidean coordinates for translation and rotation) between sequences. Ground truth data was established with the OptiTrack mechanism. Some examples from odometry outputs can be seen in Fig. 6. Here, we illustrate only short sequences qualitatively. It can be seen that the main trajectory results successfully differentiate the major displacements with a minor amount of drift.

We compare our ego-motion estimation method with monocular ORB-SLAM [32], Deep EndoVO [2], LSD SLAM [33] using Absolute Trajectory Error (ATE) [32] for the alignment with the ground truth. As shown in Fig. 6 and error bars in Fig. 7a, 7b, our method outperforms ORB SLAM and LSD SLAM which are state-of-the art widely used SLAM methods. Because of the geometric and photometric properties of scenes, these methods fail to find and match proper keypoints. Magnetic localization also outperforms ORB-SLAM and LSD-SLAM, because magnetic localization does not depend on textural geometry of the



(a) Translational error results



(b) Rotational error results

Fig. 7: Translational (a) and rotational (b) error results for ORB SLAM, LSD SLAM, Deep EndoVO, magnetic localization and our proposed supervised method. It is clear that in both rotational and translational motions, our unsupervised odometry outperforms ORB SLAM, LSD SLAM and magnetic localization, whereas Deep EndoVO shows best performance. For example, for trajectory length of 10 cm, Deep EndoVO and our method results in a translational error less than 1 cm, and others are slightly above 1 cm. In terms of rotational motion, a 5 degree change has an effect of less than 1 degree in Deep EndoVO and our method, however rest of the methods are closer to 1 degree. Translational results indicate that the proposed method shows robustness for increasing trajectory lengths and remains close to the ground truth trajectory. The trajectory length increase from 10 cm to 50 cm results a change of more than 4 cm in both ORB SLAM and LSD SLAM methods, whereas our error increases around 1 cm.

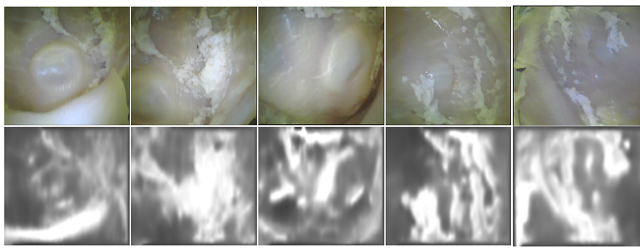


Fig. 8: Sample disparity map estimations from ex-vivo porcine stomach dataset. Even though depth estimations lack fine-scale details in low textured areas, major depth differences were successfully caught.

(a) Without fine-tuning (KITTI) (b) Original image (c) After transfer learning

Fig. 9: Disparity map outputs before and after fine-tuning on top of KITTI. (a) shows the estimation without fine-tuning. Since there is no object in front of the camera in KITTI images, the resulting disparity maps have a dark region in the center. Moreover, the disparity map has a poor quality. After transfer learning and training with porcine stomach dataset in addition to KITTI images, the quality of the disparity map drastically increases and the dark hole in the center of the image dissapears (c).

scene. Even though the proposed method is unsupervised, its translational and rotational accuracies are comparable with Deep EndoVO approach which is a supervised odometry learning method.

C. Depth Estimation

The neural network model creates depth estimation as a disparity map for a given view. Some estimation results can be seen in Fig. 8. It is clear that major depth differences are captured by the network. However, since stomach surface is non-Lambertian and the light source is attached to camera, it becomes more challenging to reproduce a robust algorithm. In the disparity map output of the network, it is observable that there are minor errors at some low textured regions or on high gradient parts such as sharp edges. However, our improvement on overall depth estimation with fine-tuning can be seen in Fig. 9.

V. CONCLUSIONS

In this paper we applied unsupervised DL method for estimating VO and depth for endoscopic capsule robot videos. Even though our method performs comparably well to supervised EndoVO method and outperforms existing state of the arts SLAM algorithms ORB and LSD SLAM, some playroom for the improvements of the method still remains:

- Accuracy of the results can be improved by increasing sequence size of inputs. As well, additional training data generated by augmentation techniques could improve the performance of the method for cases where non-rigid deformations, occlusions and heavy specularities exist.
- Since our capsule robot also uses rolling shutter camera, instead using KITTI dataset captured by global shutter camera, we could also incorporate Cityscapes dataset captured by rolling shutter camera.
- The quality of estimated depth maps can be improved by combining the depth output of our method with shading based depth estimation. In that way, a more realistic and therapeutically relevant 3D reconstruction of the

explored inner organ could be achieved.

- The dependency of the proposed method on the camera intrinsics matrix makes it rather impractical to be used for random videos streaming from hospitals with unknown calibration matrix.
- It would be interesting to extend our network to perform further tasks such as tissue segmentation and disease detection.

REFERENCES

[1] 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.

[2] 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.

[3] M. K. Goenka, S. Majumder, and U. Goenka, "Capsule endoscopy: Present status and future expectation," *World J Gastroenterol*, vol. 20, no. 29, pp. 10 024–10 037, 2014.

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] T. Nakamura and A. Terano, "Capsule endoscopy: past, present, and future," *Journal of gastroenterology*, vol. 43, no. 2, pp. 93–99, 2008.

[10] F. Munoz, G. Alici, and W. Li, "A review of drug delivery systems for capsule endoscopy," *Advanced drug delivery reviews*, vol. 71, pp. 77–85, 2014.

[11] F. Carpi, N. Kastelein, M. Talcott, and C. Pappone, "Magnetically controllable gastrointestinal steering of video capsules," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 231–234, 2011.

[12] H. Keller, A. Juloski, H. Kawano, M. Bechtold, A. Kimura, H. Takizawa, and R. Kuth, "Method for navigation and control of a magnetically guided capsule endoscope in the human stomach," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*. IEEE, 2012, pp. 859–865.

[13] A. W. Mahoney, S. E. Wright, and J. J. Abbott, "Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 5366–5371.

[14] S. Yim, E. Gultepe, D. H. Gracias, and M. Sitti, "Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 2, pp. 513–521, 2014.

[15] A. J. Petruska and J. J. Abbott, "An omnidirectional electromagnet for remote manipulation," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 822–827.

[16] 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

[17] 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

[18] 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

[19] 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

[20] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 781–788.

[21] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, vol. 2, no. 6, 2017, p. 7.

[22] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.

[23] 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.

[24] M. Fluckiger and B. J. Nelson, "Ultrasound emitter localization in heterogeneous media," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 2867–2870.

[25] J. M. Rubin, H. Xie, K. Kim, W. F. Weitzel, S. Y. Emelianov, S. R. Aglyamov, T. W. Wakefield, A. G. Urquhart, and M. ODonnell, "Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans," *Journal of Ultrasound in Medicine*, vol. 25, no. 9, pp. 1179–1186, 2006.

[26] K. Kim, L. A. Johnson, C. Jia, J. C. Joyce, S. Rangwalla, P. D. Higgins, and J. M. Rubin, "Noninvasive ultrasound elasticity imaging (uei) of crohn's disease: animal model," *Ultrasound in medicine & biology*, vol. 34, no. 6, pp. 902–912, 2008.

[27] 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.

[28] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," *CoRR*, vol. abs/1512.02134, 2015. [Online]. Available: http://arxiv.org/abs/1512.02134

[29] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291. International Society for Optics and Photonics, 2004, pp. 93–105.

[30] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," *CoRR*, vol. abs/1605.03557, 2016.

[31] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *CoRR*, vol. abs/1704.07804, 2017.

[32] 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.

[33] 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.

Endo-VMFuseNet: A Deep Visual-Magnetic Sensor Fusion Approach for Endoscopic Capsule Robots

Mehmet Turan[1], Yasin Almalioglu[2], Hunter B. Gilbert[3], Alp Eren Sari[4], Ufuk Soylu[5], and Metin Sitti[6]

***Abstract*— In the last decade, researchers and medical device companies have made major advances towards transforming passive capsule endoscopes into active medical robots. One of the major challenges is to endow capsule robots with accurate perception of the environment inside the human body, which will provide necessary information and enable improved medical procedures. We extend the success of deep learning approaches from various research fields to the problem of sensor fusion for endoscopic capsule robots in the case of asynchronous and asymmetric sensor data without any need of calibration between sensors. The results performed on real pig stomach datasets show that our method achieves high precision for both translational and rotational movements and contains various advantages over traditional sensor fusion techniques.**

I. INTRODUCTION

A fundamental requirement for medical mobile robots is the ability to accurately localize the robot during the medical operation. External and internal sensor systems, which are used to determine position and orientation coordinates of the robot, compete for on-board space and may interference with the actuation system of the capsule robot, leading to inaccuracies in terms of pose estimation [1]–[4]. Moreover, different sensors used in medical milliscale robot localization have their own particular strengths and weaknesses, which makes sensor data fusion an attractive solution. Monocular visual-magnetic odometry approaches, for example, have received considerable attention in the medical robotic sensor fusion literature [2], [5]–[9]. Fig. 1 shows a traditional sensor fusion pipeline for a camera and an additional external sensor, which in general use Kalman filter variations and particle filters for the fusion task. However, these methods suffer from inaccurate pose estimations, and they also require strict calibration and synchronization between sensors. Moreover, it is hard to find a probability density function which exactly describes the signal-to-noise ratio (SNR) of the sensors, yet the precision of the pose estimation heavily depends on the accuracy of the predicted noise model. In the last years, deep learning (DL) techniques have shown great promise in many computer vision related tasks, e.g., object detection, object recognition, classification problems, etc [4], [10]. Inspired by the recent success of deep-learning models for processing raw, high-dimensional data, we propose in this paper a sequence-to-sequence deep sensor fusion approach for endoscopic capsule robot localization which has several important advantages:

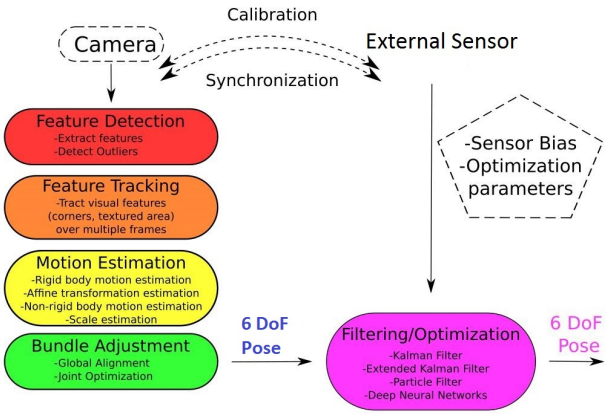


Fig. 1: Classical sensor fusion pipeline

- No spatial and temporal calibration is required between the sensors;
- The method is agnostic to sensor type and dimensionality;
- The neural network training procedure automatically learns the eye-in-hand calibration for each sensor.

We demonstrate that our proposed neural network-based fusion method can successfully fuse 6-degree-of-freedom (DoF) and 5-DoF sensor data and clearly outperforms Extended Kalman Filter (EKF)-based sensor fusion, which additionally requires spatial and temporal calibration between sensors.

This paper is organized as follows: Section II gives a survey of sensor fusion techniques for endoscopic capsule robot localization. Section III explains our method in detail. Section IV introduces the experimental setup and dataset used for the experiments. Section V shows the qualitative and quantitative results for our method compared with the endoscopic visual odometry approach and magnetic localization. And finally, we conclude with future directions.

[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 Engineering Department, Bogazici University, Turkey yasin.almalioglu@boun.edu.tr

[3]Hunter Gilbert is with the Department of Mechanical Engineering, Louisiana State University, Baton Rouge, LA 70803, USA hbgilbert@lsu.edu

[4]Ufuk Soylu is with the Electrical and Electronics Department, Middle East Technical University, Turkey ufuk.soylu@metu.edu.tr

[5]Alp Eren Sari is with the Electrical and Electronics Department, Middle East Technical University, Turkey sari.eren@metu.edu.tr

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

II. Related Work

Localization techniques for endoscopic capsule robots can be categorized into three main groups: electromagnetic wave-based techniques; magnetic field strength-based techniques and hybrid techniques [11].

Many different electromagnetic wave-based techniques have been developed, including received signal strength (RSS), time of flight and difference of arrival (ToF and TDoA), angle of arrival (AoA), and RF identification (RFID)-based methods [1], [12]–[15]. The advantage of electromagnetic wave-based techniques is that these techniques are not affected by the quasi-static magnetic field that is used for actuation. In contrast, magnetic field strength-based localization techniques must compensate for the actuation field. On the other hand, the disadvantage of electromagnetic wave-based techniques is that high-frequency electromagnetic waves are attenuated more by the human body than quasi-static magnetic fields.

In magnetic localization systems, the magnetic field generator and magnetic sensor system are the essential components. The magnetic field generator can be designed in different ways: a permanent magnet, an embedded secondary coil, or a tri-axial magnetoresistive sensor. Magnetic sensors located outside the human body detect the magnetic flux density in order to estimate the location of the capsule (e.g., [6], [16], [17]). The first advantage of magnetic field strength-based localization techniques is that they can be coupled with magnetic locomotion systems using magnetic levitation, magnetic steering, and remote magnetic manipulation. The second advantage is that low frequency magnetic fields are not attenuated by the human body. On the other hand, the disadvantage is possible interference from the environment, which requires additional hardware for handling the localization problem.

The third group of localization techniques, the hybrid techniques, utilize the integration of different sources such as magnetic sensors, RF sensors, and RGB sensors. The integration of data from different sources can produce higher quality, more reliable data. Therefore, hybrid localization techniques are promising for building accurate and robust systems. These techniques include the fusion of RF electromagnetic signal, video, and magnetic sensor data with a Kalman filter. The first group of hybrid techniques fuses RF and video signal [18], [19]. In the second group, RF signal and magnetic data are fused for the localization of the capsule robot [18], [20], [21]. In the third group of hybrid techniques, video and magnetic data are fused for the localization of the capsule robot [22].

As alternatives, there are methods which utilize computed tomography (CT), X-rays, MRI or γ rays [23], and ultrasound sensing [24]. However, each of these techniques has some drawbacks: radiation hazards should be avoided if possible, MRI devices are expensive and introduce additional restrictions on the capsule design, and ultrasound imagers capture only planar images that might not intersect the capsule robot.

III. Deep Sensor Fusion for Uncalibrated, Unsynchronized, and Asymmetric Data

We propose an end-to-end deep sensor fusion technique consisting of multi-rate Long Short-Term Memories (LSTMs) for frequency adjustment and a core LSTM unit. Our deep fusion architecture is inspired and modified from [25]. The main advantage of our fusion technique is that it eliminates the need for the separate calibration and synchronization steps of traditional sensor fusion pipelines. Our sensor fusion pipeline is shown in Fig. 2. An endoscopic visual odometry (EVO) approach is applied for 6-DoF visual localization [7], whereas a 2D array (8x8) of mono-axial Hall-effect sensors is used for 5-DoF magnetic localization. Multi-rate LSTMs process 50 Hz data coming from magnetic sensors, converting it to 30 Hz data, the same data rate of the monocular camera, whereas the core LSTM unit fuses 6-DoF visual odometry-based pose information and 5-DoF magnetic-sensor-based localization information. To summarize, our main contributions are as follows:

- We present, to the best of our knowledge, the first deep learning-based sensor fusion method for endoscopic capsule robot localization.
- Our method automatically handles the calibration and synchronization between sensors through pipeline structure and neural network training.

The input to the network is the 6-DoF localization data acquired by the EVO approach and the 5-DoF magnetic localization data from a 2D Hall effect sensor array [26]. The output of the network is a 6-DoF vector, describing relative rigid body motion of the endoscopic capsule robot from frame-to-frame.

A. Endoscopic Visual Odometry

In this subsection, we will introduce briefly our endoscopic visual odometry approach. For every input RGB image, its depth image is created using the perspective shape-from-shading algorithm by [2], [27]. For the pose estimation from RGB and depth images, an energy-minimization-based technique is developed containing both optical flow (OF) based sparse feature correspondence and dense pose alignment based on volumetric and photometric energy minimization [28]–[30].

The coarse global alignment based on optical flow serves as the initialization of the dense alignment, which uses GPU hardware acceleration to incorporate all of the available information provided by the input image in an interactive-rate system. Such a dense alignment approach is very helpful for pose estimation in low textured areas, where sparse methods are prone to fail. Taking into account the complete sequence of the previous frame history, the sparse alignment module attempts to establish OF correspondences between an input frame and the previous frames to provide a coarse initial global pose optimization, whereas the dense optimization

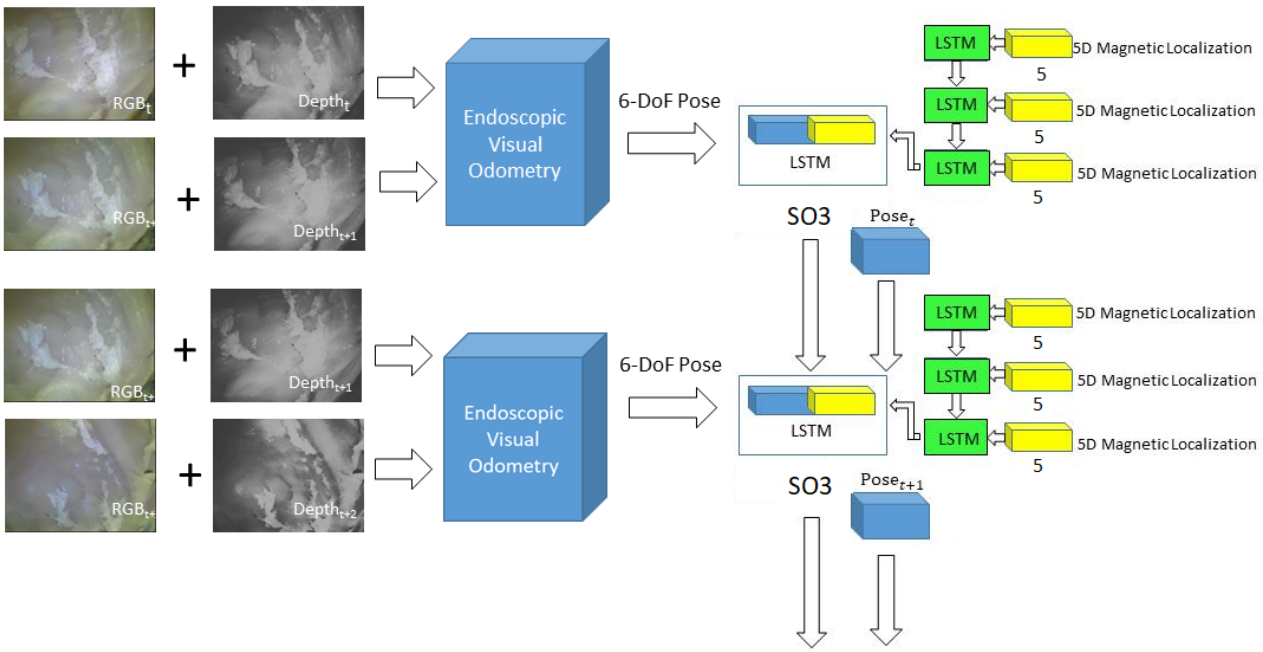


Fig. 2: Deep learning architecture of Endo-VMFuseNet.

stage serves for a fine-scale refinement. Additionally, OF-based global optimization allows for continuous loop-closure and re-localization. Such a re-localization capability is essential to recover from tracking failures in case of unexpected drifts inside the GI-tract. Inspired from the pose estimation strategies proposed by [28]–[30], the energy minimization equation of our coarse-then-fine approach is as follows:

$$\mathbf{X} = (R_o, t_o, ..., R_{|S|}, t_{|S|})^T \tag{1}$$

$$E_{\text{align}}(\mathbf{X}) = \omega_{\text{sparse}} E_{\text{sparse}}(\mathbf{X}) + \omega_{\text{dense}} E_{\text{dense}}(\mathbf{X}) \tag{2}$$

for $|S|$ frames, where ω_{sparse} and ω_{dense} are weights assigned to sparse and dense matching terms, and $E_{\text{sparse}}(\mathbf{X})$ and $E_{\text{dense}}(\mathbf{X})$ are the sparse and dense matching terms, respectively, such that:

$$E_{\text{sparse}}(\mathbf{X}) = \sum_{(\text{i}=1)}^{|S|} \sum_{(\text{j}=1)}^{|S|} \sum_{(\text{k},1) \in \text{C(i,j)}} ||\tau_i P_{\text{i,k}} - \tau_j P_{\text{j,k}}||^2 \tag{3}$$

Here, $P_{\text{i, k}}$ is the k^{th} detected feature point in the i^{th} frame. $C(i, j)$ is the set of all pairwise correspondences between the i^{th} and the j^{th} frame. The Euclidean distance over all the detected feature matches is minimized once the best rigid transformation, τ_i, is found. Dense pose estimation is described as follows [28]–[30]:

$$E_{\text{dense}}(\tau) = \omega_{\text{photo}} E_{\text{photo}}(\tau) + \omega_{\text{geo}} E_{\text{geo}}(\tau) \tag{4}$$

whereas,

$$E_{\text{photo}}(\mathbf{X}) = \sum_{(\text{i,j}) \in \text{E}} \sum_{\text{k}=0}^{|\text{I}_\text{i}|} ||I_i(\omega(d_{\text{i,k}})) - I_j(\omega(\tau_j^{-1} \tau_i d_{\text{i,k}}))||_2^2 \tag{5}$$

and,

$$E_{\text{geo}}(\mathbf{X}) = \sum_{(\text{i,j}) \in \mathbf{E}} \sum_{\text{k}=0}^{|\text{D}_\text{i}|} [\mathbf{n}_{\text{i,k}}^T (\mathbf{d}_{\text{i,k}} - \tau_i^{-1} \tau_j \omega^{-1} (D_j(\omega(\tau_j^{-1} \tau_i d_{\text{i,k}}))))]^2 \tag{6}$$

with τ_i being rigid camera transformation, $P_{\text{i,k}}$ the k^{th} detected feature point in i^{th} frame, $\mathbf{n}_{\text{i,k}}$ is the normal of the k^{th} pixel in the i^{th} input frame, $\mathbf{d}_{\text{i,k}}$ is the 3D position associated with the k^{th} pixel of the i^{th} depth frame, $C(i, j)$ being the set of pairwise correspondences between the i^{th} and j^{th} frame. The set of rigid camera transforms is denoted as τ, the function ω is the perspective projection, D is the depth of the input frame, and I is the gradient of the luminance of frame's color.

B. Magnetic Localization System

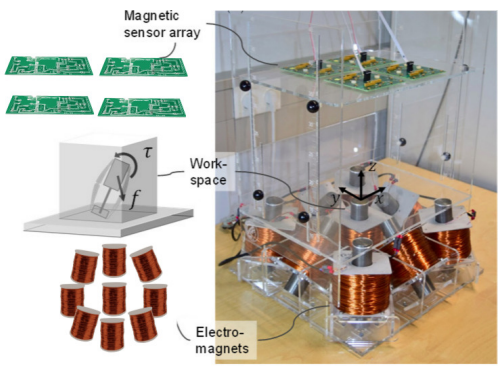


Fig. 3: Photo and schematic of our magnetic localization system using an array of hall-effect sensors externally.

Our magnetic localization technique [31] is able to measure 5-DoF absolute pose values for an untethered meso-scale magnetic robot. As shown in Fig. 3, the system consists of a magnetic sensor system for localization and electromagnets for actuation of the magnetic capsule robot. A Hall-effect sensor array measures magnetic field at several locations from the magnetic capsule robot, whereas a computer-controlled electromagnetic coil array provides actuator's magnetic field. The core idea of our localization technique is separation of capsule's magnetic field from actuator's known magnetic field, which is realized by subtracting actuator's magnetic field component from the measured magnetic data. Finally, noise effects are reduced by second-order directional differentiation. For curious readers, further details of the magnetic localization technique can be found in [31].

C. Deep learning based Sensor Fusion

Recurrent Neural Networks (RNNs) are suitable for modelling the dependencies across data sequences and for creating a temporal motion model thanks to its memory of hidden states over time. This allows the pose estimation for the current time to benefit from the prior information of past sensor data in a similar way to the way that statistical filters use prior distributions to estimate posterior ones. To address the vanishing and exploding gradients that are the most common challenges in designing and training RNNs, a particular form of RNN, which is called LSTM, was introduced by [32]. The information flow through LSTM is shown in Fig. 4. The LSTM model has a memory cell c_t that encodes the knowledge that is observed up to time step t. Gates control the behaviour of the cell. They are the layers that are multiplicatively applied, and can keep or discard a value from the gated layer. Three gates are used in the LSTM, which control whether to forget the current cell value in the forget gate, if it should read its input in the input gate and whether to output the new cell value in the output gate. $\odot$ is the element-wise multiplication with a gate value, $\sigma(\cdot)$ is the sigmoid non-linearity and f is the forget gate.

Our deep RNN model is constructed by concatenating the core-LSTM on top of two multi-rate LSTMs with inputs from the EVO and the magnetic localization system as illustrated in Fig. 2. Each LSTM layer has 200 hidden states. The system learns translational and rotational movements simultaneously. To regress the 6-DoF pose, we trained the RNN architecture on Euclidean loss using the Adam optimization method with the following objective loss function:

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \|\hat{\mathbf{q}} - \mathbf{q}\|_2 \qquad (7)$$

where $\mathbf{x}$ is the translation vector and $\mathbf{q}$ is the Euler vector for a rotation. A balance β must be kept between the orientation and translation loss values which are highly coupled as they are learned from the same model weights [33]. Experimental results showed that the optimal β was given by the ratio between expected error of position and orientation at the end of training session.

The back-propagation algorithm is used to determine the gradients of RNN weights. These gradients are passed into the Adam optimization method which is a stochastic gradient descent algorithm based on estimation of first and second-order moments. The moments of the gradient are calculated using exponential moving average in addition to exponentially decaying average of past gradients, which also corrects the bias.

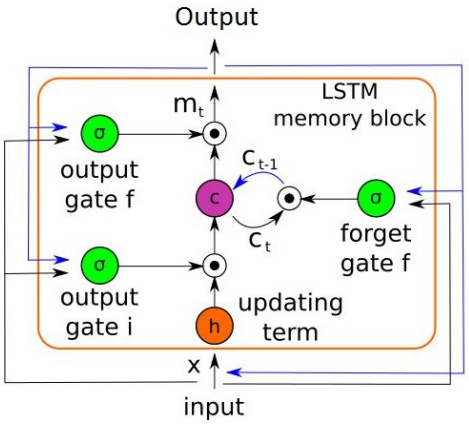


Fig. 4: Information flow through the hidden units of the LSTM.

IV. Experimental Setup

A. Magnetically Actuated Soft Capsule Endoscopes

Our capsule prototype is a magnetically actuated soft capsule endoscope (MASCE) which is designed to be used in the upper gastrointestinal tract for disease detection, drug delivery, and biopsy operations. The prototype is composed of an RGB camera, a permanent magnet, and a drug chamber (see Fig. 5 for visual reference). The magnet produces force and torque in response to a controlled external magnetic field, which are used to actuate the capsule robot and to release drugs in specifically targeted locations. A 2D

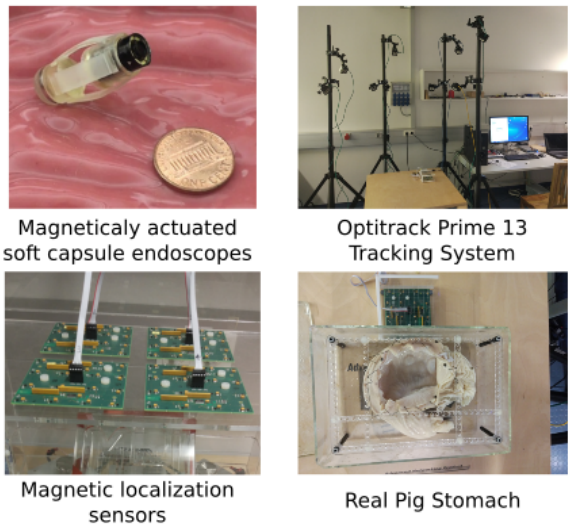


Fig. 5: An overview of the experimental setup

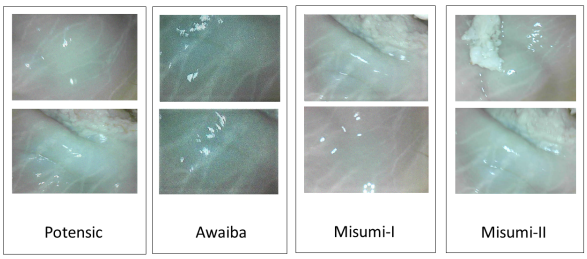


Fig. 6: Sample images from real pig stomach dataset

magnetic sensor array is placed on top of the workspace and electromagnets are placed at the bottom of the workspace, with the patient located in between. Magnetic fields from the electromagnets generate the magnetic force and torque on the ring magnet around MASCE so that the robot moves inside the workspace. The coordinate system in Fig. 3 shows the origin and orientation of the workspace.

B. Dataset

The dataset was recorded on five different real pig stomachs (Fig. 5). In order to ensure that our algorithm is not tuned to a specific camera model, four different commercial endoscopic cameras were employed. For each pig stomach and camera combination, 3000 frames were acquired, which makes 60000 frames for four cameras and five pig stomachs in total. The endoscopic capsule robot was actuated using magnetic actuation system. 40000 frames were used for training the RNNs, whereas the remaining 20000 frames were used for evaluation. Sample real pig stomach frames are shown in Fig. 6 for visual reference. During video recording, an Optitrack motion tracking system consisting of eight Prime-13 cameras and the manufacturer's tracking software was utilized to obtain 6-DoF localization ground-truth-data with sub-millimetre accuracy (see Fig. 5) which was used as a gold standard for the evaluations of the pose estimation accuracy. The tracking system consistently produces positional error less than 0.3 mm and rotational error less than 0.05°.

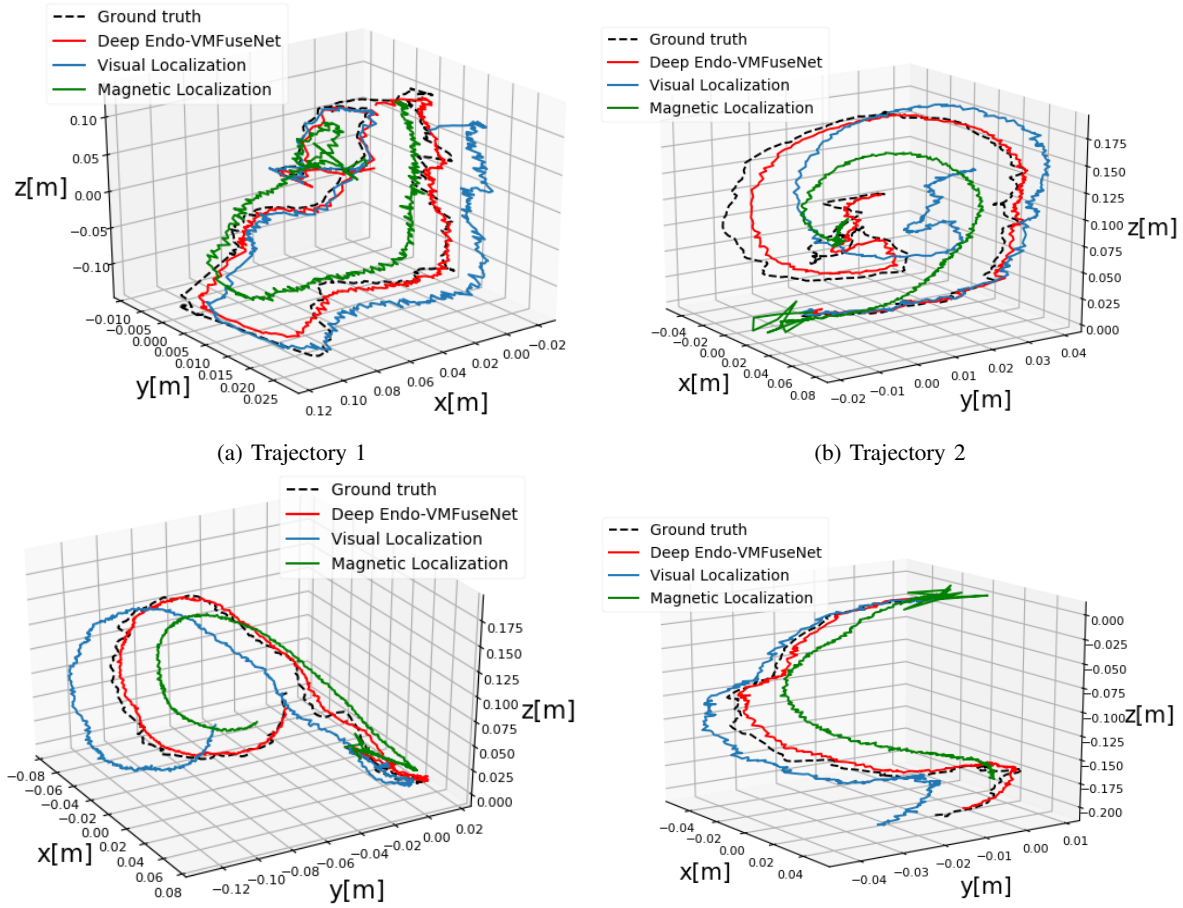
V. Results and Discussion

The RNN architecture was trained using the Caffe library on an NVIDIA Tesla K80 GPU. Using the back-propagation-through-time method, the weights of the hidden units were trained for up to 200 epochs with an initial learning rate of 0.001. Overfitting,which would make the resulting pose estimator inapplicable in other scenarios, was prevented using dropout and early stopping techniques. The dropout regularization technique, which samples a part of the whole network and updates its parameters based on the input data [34], is an extremely effective and simple method to avoid overfitting. Early stopping is another widely used technique to prevent overfitting of a complex neural network architecture optimized by a gradient-based method. We strictly avoided the use of any image frames from the training session for the testing session.

The performance of the deep Endo-VMFuseNet approach was analysed using averaged root mean square error (RMSE) estimation for translational and rotational motions. For trajectories of various complexity, such as uncomplicated paths with slow incremental translations and rotations, and comprehensive scans with many local loop closures and complex paths with fast rotational and translational movements, we performed tests on deep Endo-VMFuseNet comparing with EVO localization and magnetic localization. The average translational and rotational RMSEs for deep Endo-VMFuseNet, EVO localization, magnetic localization and EKF-based sensor fusion against different path lengths are shown in Fig. 8, respectively. A spatial and temporal calibration between the camera and magnetic sensor is performed before EKF-based sensor fusion. The results indicate that deep Endo-VMFuseNet clearly outperforms EKF-based sensor fusion for both translational and rotational localization, whereas both sensor fusion methods perform better than visual and magnetic localization. We presume that the effective use of the LSTM architecture in Endo-VMFuseNet architecture enabled learning from asynchronous and uncalibrated sensor array. The results also indicate that Endo-VMFuseNet is capable of handling asynchronous data (50 Hz magnetic data and 30 Hz visual data), by interpreting the localization information from the current magnetic data and previous visual and magnetic localization information saved by internal hidden memory of LSTM units. Moreover, we can conclude that Endo-VMFuseNet is also able to handle asymmetric sensor data, i.e the missing 6th degree of magnetic localization by making use of existing 6-DoF EVO and 5-DoF magnetic sensor information from current and previous frames. Some sample ground-truth and estimated trajectories for Endo-VMFuseNet, EVO localization and magnetic localization are shown in Fig. 7d for visual reference. As seen in sample trajectories, Endo-VMFuseNet is able to stay close to the ground-truth pose values for even complex, fast rotational and translational motions, where both EVO and magnetic localization by themselves clearly deviate from the ground-truth trajectory. Thus, we can conclude that Endo-VMFuseNet makes effective use of both sensor data streams.

VI. CONCLUSIONS

In this study, we presented, to the best of our knowledge, the first sensor fusion method based on deep learning for endoscopic capsule robots. The proposed fusion architecture is able to achieve simultaneous learning and sequential modelling of motion dynamics across sensor streams by concatenating the core LSTM with two multi-rate LSTMs. Many issues faced by traditional sensor fusion techniques such as external calibration of sensors, synchronization between sensors and issue of unsensed degrees of freedom in one or more sensors are successfully handled by deep Endo-VMFuseNet. Since it is trained in an end-to-end manner, there is no need to carefully hand-tune the parameters of the system.



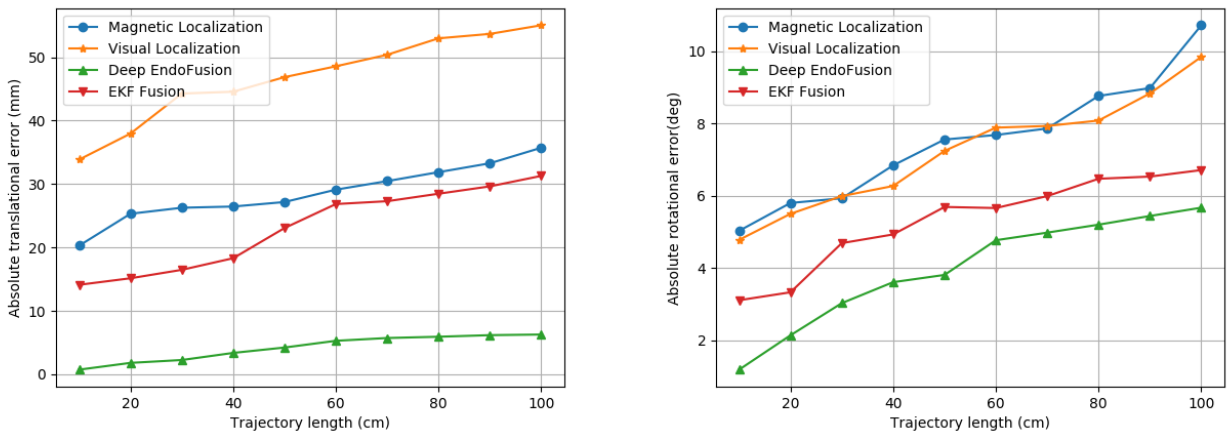
(a) Trajectory 1

(b) Trajectory 2

(c) Trajectory 3

(d) Trajectory 4

Fig. 7: Sample ground-truth trajectories and estimated trajectories predicted by the DL-based sensor fusion approach. As seen, deep Endo-VMFuseNet is the closest to the ground truth trajectories.



(a) Trajectory length vs translation error

(b) Trajectory length vs rotation error

Fig. 8: Deep Endo-VMFuseNet outperforms both of the other models in terms of translational and rotational position estimation.

VII. Acknowledgements

H.G. thanks the Alexander von Humboldt foundation for financial support. This work is funded by the Max Planck Society.

References

[1] Y. Ye, "Bounds on RF cooperative localization for video capsule endoscopy," Ph.D. dissertation, Worcester Polytechnic Institute, 2013.

[2] G. Ciuti, M. Visentini-Scarzanella, A. Dore, A. Menciassi, P. Dario, and G.-Z. Yang, "Intra-operative monocular 3D reconstruction for image-guided navigation in active locomotion capsule endoscopy," in *Biomedical Robotics And Biomechatronics (Biorob), 2012 4th Ieee Ras & Embs International Conference On*. IEEE, 2012, pp. 768–774.

[3] M. Turan, Y. Y. Pilavci, R. Jamiruddin, H. Araujo, 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," *arXiv preprint arXiv:1705.06524*, 2017.

[4] 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.

[5] M. Turan, Y. Almalioglu, H. Araujo, T. Cemgil, and M. Sitti, "Endosensorfusion: Particle filtering-based multi-sensory data fusion with switching state-space model for endoscopic capsule robots using recurrent neural network kinematics," *arXiv preprint arXiv:1709.03401*, 2017.

[6] K. M. Popek, A. W. Mahoney, and J. J. Abbott, "Localization method for a magnetic capsule endoscope propelled by a rotating magnetic dipole field," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 5348–5353.

[7] 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, 2017.

[8] 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*, pp. 1–15, 2017.

[9] 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.

[10] M. Turan, A. Abdullah, R. Jamiruddin, H. Araujo, E. Konukoglu, and M. Sitti, "Six degree-of-freedom localization of endoscopic capsule robots using recurrent neural networks embedded into a convolutional neural network," *arXiv preprint arXiv:1705.06196*, 2017.

[11] I. Umay, B. Fidan, and B. Barshan, "Localization and tracking of implantable biomedical sensors," *Sensors*, vol. 17, no. 3, p. 583, 2017.

[12] Y. Wang, R. Fu, Y. Ye, U. Khan, and K. Pahlavan, "Performance bounds for RF positioning of endoscopy camera capsules," in *Biomedical Wireless Technologies, Networks, and Sensing Systems (BioWireleSS), 2011 IEEE Topical Conference on*. IEEE, 2011, pp. 71–74.

[13] D. Fischer, R. Schreiber, D. Levi, and R. Eliakim, "Capsule endoscopy: the localization system," *Gastrointestinal Endoscopy Clinics*, vol. 14, no. 1, pp. 25–31, 2004.

[14] L. Wang, C. Hu, L. Tian, M. Li, and M. Q.-H. Meng, "A novel radio propagation radiation model for location of the capsule in gi tract," in *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*. IEEE, 2009, pp. 2332–2337.

[15] J. Hou, Y. Zhu, L. Zhang, Y. Fu, F. Zhao, L. Yang, and G. Rong, "Design and implementation of a high resolution localization system for in-vivo capsule endoscopy," in *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*. IEEE, 2009, pp. 209–214.

[16] C. Di Natali, M. Beccani, N. Simaan, and P. Valdastri, "Jacobian-based iterative method for magnetic localization in robotic capsule endoscopy," *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 327–338, 2016.

[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] Y. Geng and K. Pahlavan, "Design, implementation, and fundamental limits of image and RF based wireless capsule endoscopy hybrid localization," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 1951–1964, 2016.

[19] G. Bao, K. Pahlavan, and L. Mi, "Hybrid localization of microrobotic endoscopic capsule inside small intestine by data fusion of vision and RF sensors," *IEEE Sensors Journal*, vol. 15, no. 5, pp. 2669–2678, 2015.

[20] I. Umay and B. Fidan, "Adaptive magnetic sensing based wireless capsule localization," in *Medical Information and Communication Technology (ISMICT), 2016 10th International Symposium on*. IEEE, 2016, pp. 1–5.

[21] ——, "Adaptive wireless biomedical capsule tracking based on magnetic sensing," *International Journal of Wireless Information Networks*, vol. 24, no. 2, pp. 189–199, 2017.

[22] J. D. Gumprecht, T. C. Lueth, and M. B. Khamesee, "Navigation of a robotic capsule endoscope with a novel ultrasound tracking system," *Microsystem technologies*, vol. 19, no. 9-10, pp. 1415–1423, 2013.

[23] T. D. Than, G. Alici, S. Harvey, G. OKeefe, H. Zhou, W. Li, T. Cook, and S. Alam-Fotias, "An effective localization method for robotic endoscopic capsules using multiple positron emission markers," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1174–1186, 2014.

[24] K. Arshak and F. Adepoju, "Capsule tracking in the gi tract: A novel microcontroller based solution," in *Sensors Applications Symposium, 2006. Proceedings of the 2006 IEEE*. IEEE, 2006, pp. 186–191.

[25] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem." in *AAAI*, 2017, pp. 3995–4001.

[26] D. Son, M. D. Dogan, and M. Sitti, "Magnetically actuated soft capsule endoscope for fine-needle aspiration biopsy," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1132–1139.

[27] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Metric depth recovery from monocular images using shape-from-shading and specularities," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 25–28.

[28] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 3, p. 24, 2017.

[29] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph." Robotics: Science and Systems, 2015.

[30] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[31] 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, Oct. 2016.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[34] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

EndoSensorFusion: Particle Filtering-Based Multi-sensory Data Fusion with Switching State-Space Model for Endoscopic Capsule Robots

Mehmet Turan[1], Yasin Almalioglu[2], Hunter Gilbert[3], Helder Araujo[4], Taylan Cemgil[5], and Metin Sitti[6]

***Abstract*— A reliable, real time, multi-sensor fusion functionality is crucial for localization of actively controlled capsule endoscopy robots, which are an emerging, minimally invasive diagnostic and therapeutic technology for the gastrointestinal (GI) tract. In this study, we propose a novel multi-sensor fusion approach based on a particle filter that incorporates an online estimation of sensor reliability and a non-linear kinematic model learned by a recurrent neural network. Our method sequentially estimates the true robot pose from noisy pose observations delivered by multiple sensors. We experimentally test the method using 5 degree-of-freedom (5-DoF) absolute pose measurement by a magnetic localization system and a 6-DoF relative pose measurement by visual odometry. In addition, the proposed method is capable of detecting and handling sensor failures by ignoring corrupted data, providing the robustness expected of a medical device. Detailed analyses and evaluations are presented using *ex vivo* experiments on a porcine stomach model, proving that our system achieves high translational and rotational accuracies for different types of endoscopic capsule robot trajectories.**

I. INTRODUCTION

Milli-scale, untethered, mobile robots have the potential to make a major impact on healthcare. Swallowable capsule endoscopes with an on-board camera and wireless image transmission device have been commercialized and used in hospitals (FDA approved) since 2001. These devices have enabled access to regions of the GI tract that were impossible to access before, and have reduced discomfort and sedation-related loss of work [1]–[3]. However, with systems that are commercially available today, capsule endoscopy cannot provide precise (centimeter- to millimeter-accurate) localization of diseased areas, and active, wireless control remains a highly active area of research. Several groups have recently proposed remotely controllable robotic capsule endoscopes that are equipped with additional functionalities, such as localized drug delivery, biopsy and other medical functions [4]–[8]. Accurate and robust localization would not only provide better diagnostic information in passive devices, but would also improve the reliability and safety of active control strategies like remote magnetic actuation.

In the last decade, many different approaches have been developed for real-time endoscopic capsule robot localization, including received signal strength (RSS), time of flight and time difference of arrival (ToF and TDoA), angle of arrival (AoA), and radiofrequency identification (RFID)-based methods [4], [9]. Recently, it has also been shown that the permanent magnets which are added to capsule robots to facilitate remote magnetic actuation can be simultaneously used for precise localization [10]. This strategy has a clear advantage for miniaturization: the permanent magnet provides two essential functions rather than one.

Hybrid techniques based on the combination of different measurements can improve both the accuracy and the reliability of the location measurement system. Sensor fusion techniques have been applied to wireless capsule endoscopes, and several combinations of sensor types have been investigated. Most of the techniques that have been demonstrated for data fusion have been based on Kalman filtering. The first subgroup of hybrid techniques fuses radio frequency (RF) signals and video for localization of the capsule robot [11]. Geng et al. assert that using RF signals and video data can result in millimetric accuracy, whereas previous techniques were able to achieve only a few centimeters accuracy.

In the second group, RF signal and magnetic localization are fused to locate the capsule robot [12], [13]. In these studies, a localization method that has high accuracy for simultaneous position and orientation estimation has been investigated. In the third group of hybrid techniques, video-based tracking and magnetic localization are fused [14]. In [14], the authors introduced a technique that combines ultrasound imaging and magnetic field-based localization.

Although some of these state-of-the-art sensor fusion techniques have achieved remarkable accuracy for the tracking and localization task of a capsule robot, they are not able to detect and autonomously handle sensor faults, and additionally several techniques using RF localization require complex signal corrections to account for attenuation and propagation of RF signals inside human body tissues. In addition, most previous models use relatively simple dynamic models for the capsule, whereas performance would be greatly improved by a more accurate model of the system. Although most of the existing techniques have been based on Kalman filters, these filters typically work best for linear systems, whereas the dynamics of capsule robots are generally nonlinear. Lastly, previously demonstrated methods generate inaccurate estimations in cases where noise from the environment and

[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 Engineering Department, Bogazici University, Turkey yasin.almalioglu@boun.edu.tr

[3]Hunter Gilbert is with the Mechanical Engineering Department, Louisiana State University, Baton Rouge, LA 70803, USA hbgilbert@lsu.edu

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

[5]Taylan Cemgil is with the Computer Engineering Department, Bogazici University, Turkey taylan.cemgil@boun.edu.tr

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

the actuation system interferes with one or more components of the localization system.

In this paper, we propose a novel multi-sensor fusion algorithm for capsule robots, which is based on a switching state space model and particle filtering. The dynamic model of the endoscopic capsule robot is based on Recurrent Neural Networks (RNNs). The resulting method naturally handles both nonlinear motion models and sensor faults. The main contributions of our paper are as follows:

- To the best of our knowledge, this is the first multi-sensor data fusion approach that combines a switching observation model, a particle filter approach, and a recurrent neural network developed for the endoscopic capsule robot and hand-held endoscope localization.
- We propose a sensor failure detection system for endoscopic capsule robots based on probabilistic graphical models with efficient proposal distributions applied onto the particle filtering. The approach can be generalized to any number of sensors and any mobile robotic platforms.
- No manual formulation is required to determine a probability density function that describes the motion dynamics, contrary to traditional particle filter and Kalman filter based methods.

The paper is organized as follows. Section II introduces the sensor fusion algorithms and the RNN-based dynamic model. Section III describes the experiments used to verify the proposed methods for a wireless capsule endoscope in an ex vivo porcine model. Section IV includes the results and discussion of the experiments, and section VI concludes with future directions.

II. Sensor Fusion and Modeling Approach

The particle filter is a Bayesian filtering method that computes the posterior probability density functions (pdf) of sequentially obtained state vectors $\mathbf{x_t} \in \mathscr{X}$, which are suggested by (complete or partial) sensor measurements. For the capsule robot, the state $\mathbf{x_t}$ is composed of the 6-DoF pose, which is assumed to propagate in time according to a general model:

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \mathbf{v}_t) \\ \mathbf{y}_t &= g(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{aligned} \tag{1}$$

where f is a non-linear state transition function and $\mathbf{v_t}$ is white noise distributed. The function g encodes the transition of a hidden state $\mathbf{y}_t$. t is the index of a time sequence, i.e. $t \in \{1,2,3,...\}$. In general, the hidden state $\mathbf{y}_t$ might represent, at a minimum, the rigid body velocity. However, it could also include other dynamic factors like acceleration or jerk, and may also be used to represent environmental state.

6-DoF pose state estimation with a high precision is a complex problem, which often requires multi-sensor input or sequential observations. In our capsule, we have two sensor systems, one being a 5-DoF magnetic sensor array and the other one being an endoscopic monocular RGB camera (these subsystems are described later). Generally speaking, observations of the pose are produced by n sensors $\mathbf{z}_{k,t}(k = 1,...,n)$, where the probability distribution $p(\mathbf{z}_{k,t}|\mathbf{x}_t)$ is known for each sensor.

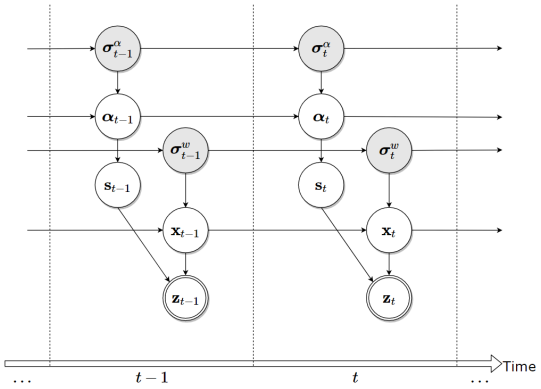


Fig. 1: The overall switching state-space model. The double circles denote observable variables and the gray circles denote hyper-parameters.

A. The Sequential Bayesian Model and Problem Statement

We estimate the 6-DoF pose states, which rely on latent (hidden) variables, by using the Bayesian filtering approach. The probabilistic graphical model that shows the relations between all of the variables is shown in Fig. 1. 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:

$$\mathbf{z}_{k,t} = h_{k,s_{k,t},t}(\mathbf{x}_t) + \mathbf{w}_{k,s_{k,t},t} \tag{2}$$

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 latent variable of the switch parameter $s_{k,t}$ is defined to be 0 if the sensor is in a failure state, which means that observation $\mathbf{z}_{k,t}$ is independent of $\mathbf{x}_t$, and 1 if the sensor k is in its nominal state of work. 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:

$$Pr(s_{k,t} = j) = \alpha_{k,j,t}, \quad 0 \leq j \leq d_k \tag{3}$$

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

B. Proposal Distributions

In this section, we formulate the optimal proposal distributions in terms of minimizing the variance of the weights

and effective approximations, in cases where sampling from the optimal distributions is not feasible. The particles are extended from time $t-1$ to time t according to the importance distribution denoted by $q(\cdot)$.

- $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(i)}, \sigma_t^{w(i)}, \hat{\mathbf{s}}_t^{(i)}, \mathbf{z}_t)$ is approximated by an unscented Kalman filter (UKF) step for particle i:

$$\hat{\mathbf{x}}_{t|t}^{(i)} = \hat{\mathbf{x}}_{t|t-1}^{(i)} + \sum_{k=1}^{n} \hat{s}_{k,t}^{(i)} K_{k,t}^{(i)} \hat{v}_{k,t}^{(i)}$$

where $\hat{\mathbf{x}}_{t|t-1}^{(i)} = f(\mathbf{x}_{t-1}^{(i)})$, n is the number of sensors, $\hat{v}_{k,t}^{(i)}$ is the residual, and $K_{k,t}^{(i)}$ is the Kalman gain sequentially obtained by UKF. Finally,

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(i)}, \sigma_t^{w(i)}, \hat{\mathbf{s}}_t^{(i)}, \mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_{t|t}^{(i)}, P_{t|t}^{(i)})$$

where the error covariance matrix, $P_{t|t}^{i}$ is obtained by the UKF step with the process noise of $\sigma_t^{w(i)}$.

- In switching state-space models, the switch parameters with self-adaptive prior are more efficient than a fixed prior approach [15], [16]. The optimal proposal distribution for switch variable that represents the state of a sensor is given by

$$Pr(\mathbf{s}_{k,t}|\mathbf{x}_{t-1}^{(i)}, \alpha_{k,t-1}^{(i)}, \mathbf{z}_{k,t}) = \frac{\alpha_{k,s_{k,t},t-1}^{(i)} p(\mathbf{z}_{k,t}|s_{k,t}, \mathbf{x}_{t-1}^{(i)})}{\sum_{j=0}^{d_k} \alpha_{k,s_{k,t},t-1}^{(i)} p(\mathbf{z}_{k,t}|j, \mathbf{x}_{t-1}^{(i)})} \tag{4}$$

which is approximated by applying UKF to pdfs $p(\mathbf{z}_{k,t}|j, \mathbf{x}_{t-1}^{(i)})$ for $j = 0, \ldots, d_k$

$$p(\mathbf{z}_{k,t}|j, \mathbf{x}_{t-1}^{(i)}) \simeq \mathcal{N}(h_{k,j,t}(\hat{x}_{t|t-1}^{(i)}), S_{k,j,t}^{(i)}) \tag{5}$$

where $\hat{x}_{t|t-1}^{(i)} = f(\mathbf{x}_{t-1}^{(i)})$ is the state prediction and $S_{k,j,t}^{(i)}$ is the approximated innovation covariance matrix approximated by UKF. Hence, the proposal distribution for the switch parameter $\mathbf{s}_{k,t}$ is given by

$$\begin{aligned} &q(\mathbf{s}_{k,t}|\mathbf{x}_{t-1}^{(i)}, \alpha_{k,t-1}^{(i)}, \mathbf{z}_{k,t}) \\ &\propto \alpha_{k,s_k,t-1}^{(i)} \mathcal{N}(h_{k,s_k,t-1}(\hat{x}_{t|t-1}^{(i)}), S_{k,s_k,t}^{(i)}) \end{aligned} \tag{6}$$

- The optimal proposal distribution for the hyperparameter $\sigma_{k,t-1}^{\alpha}$ is calculated in closed form as

$$\begin{aligned} &q\left(log(\sigma_{k,t}^{\alpha})|\alpha_{k,t}^{(i)}, \alpha_{k,t-1}^{(i)}, \sigma_{k,t-1}^{\alpha(i)}\right) \\ &= \frac{D\left(\alpha_{k,t}^{(i)}; \sigma_{k,t}^{\alpha} \alpha_{k,t-1}^{(i)}\right)}{D\left(\alpha_{k,t}^{(i)}; \sigma_{k,t-1}^{\alpha} \alpha_{k,t-1}^{(i)}\right)} \\ &\times \mathcal{N}\left(log(\sigma_{k,t}^{\alpha}); log(\sigma_{k,t-1}^{\alpha(i)}), \lambda^{\alpha}\right). \end{aligned} \tag{7}$$

We generate samples from the distribution with the Adaptive Rejection Sampling (ARS) method because direct sampling is not feasible [17]. Using ARS, the need for locating the supremum diminishes because the distribution is log-concave. Another advantage of ARS is that it uses recently acquired information to update the envelope and squeezing functions, which reduces the need to evaluate the distribution after each rejection step. Fig. 2 shows an ARS sampling result indicating the effectiveness of the applied sampling method for the proposal distribution. It can be seen in Fig. 2 that a tight piecewise hull has converged to the target distribution after rejection steps, and interior knots are regenerated in the vicinity of the expected values.

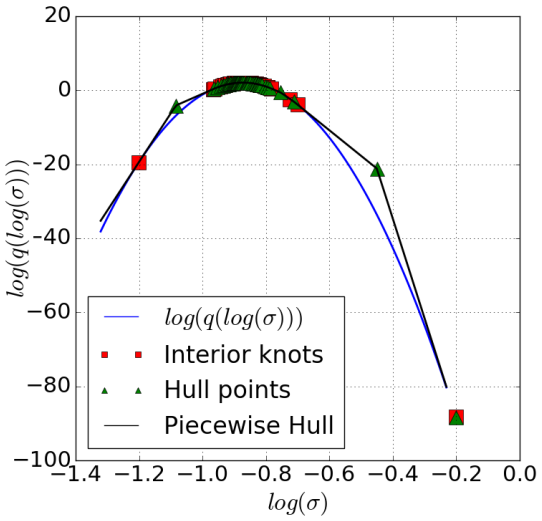


Fig. 2: Example ARS sampling result for $log(\sigma_{k,t})$. The piecewise hull and the generated samples are shown.

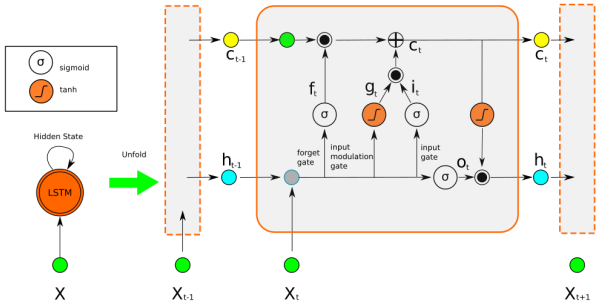


Fig. 3: Information flow through the units of the LSTM [18]

- Considering that the Dirichlet distribution is conjugate to the multinomial distribution, the optimal proposal distribution for the confidence parameter $\alpha_{k,t}$ can be reformulated in closed form as a Dirichlet distribution with a decreasing variance parameter for failure sensor states.

C. RNN-based Kinematics Model

Existing sensor fusion methods based on traditional particle filter and Kalman filter approaches have their limitations

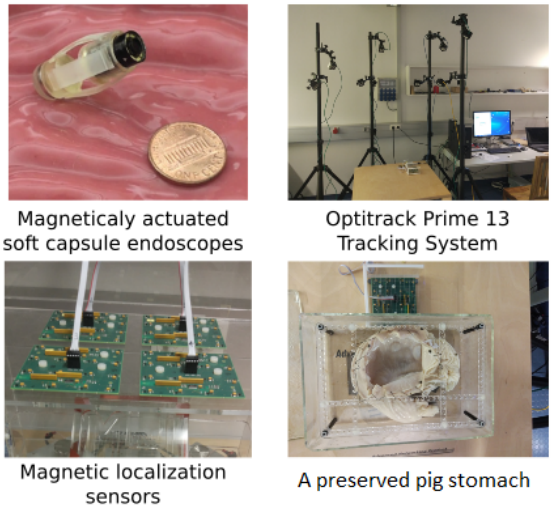


Fig. 4: Experimental setup

when applied to nonlinear dynamic systems. The Kalman filter and extended Kalman filter assume that the underlying dynamic process is well-modeled by linear equations or that these equations can be linearised without a major loss of fidelity. On the other hand, particle filters accommodate a wide variety of dynamic models, allowing for highly complex dynamics in the state variables.

In the last few years, deep learning (DL) techniques have provided solutions to many computer vision and machine learning tasks. Contrary to these high-level tasks, multi-sensory data fusion is mainly concerned with motion dynamics and the relations across sequences of pose observations obtained from sensors, which can be formulated as a sequential learning problem. Unlike traditional feed-forward artificial neural networks, RNNs are very suitable for modelling the dependencies across time sequences and for creating a temporal motion model since they have a memory of hidden states over time and have directed cycles among hidden units, enabling the current hidden state to be a function of arbitrary sequences of inputs. Thus, using an RNN, the pose estimation of the current time step benefits from information encapsulated in previous time steps and is suitable to formulate the state transition functions f and g in Equation 1. A particle filter tracks the 6-DoF pose of the capsule robot using the transition function modelled by the LSTM network. To train the LSTM, the inputs are 6-DoF poses (states) at time step $t-1$, and output labels are 6-DoF poses at time t. In that way, the LSTM learns the dynamic model of the capsule robot.

Long Short-Term Memory (LSTM) is a suitable implementation of RNN to exploit longer trajectories since it avoids the vanishing gradient problem of RNN, resulting in a higher capacity of learning long-term relations among the sequences by introducing memory gates such as input, forget and output gates, and hidden units of several blocks. The information flow of the LSTM is shown in Fig.3. The input gate controls the amount of new information flowing into the current state, the forget gate adjusts the amount of existing information that remains in the memory, and the output gate decides which part of the information triggers the activations. Given the input vector x_k at time k, the output vector h_{k-1} and the cell state vector c_{k-1} of the previous LSTM unit, the LSTM updates at time step k according to the following equations:

$$f_k = \sigma(W_f \cdot [x_k, h_{k-1}] + b_f) \tag{8}$$
$$i_k = \sigma(W_i \cdot [x_k, h_{k-1}] + b_i) \tag{9}$$
$$g_k = \tanh(W_g \cdot [x_k, h_{k-1}] + b_g) \tag{10}$$
$$c_k = f_k \odot c_{k-1} + i_k \odot g_k \tag{11}$$
$$o_k = \sigma(W_o \cdot [x_k, h_{k-1}] + b_o) \tag{12}$$
$$h_k = o_k \odot \tanh(c_k) \tag{13}$$

where σ is sigmoid non-linearity, tanh is hyperbolic tangent non-linearity, W terms denote corresponding weight matrices, b terms denote bias vectors, i_k, f_k, g_k, c_k and o_k are input gate, forget gate, input modulation gate, the cell state and output gate at time k, respectively, and $\odot$ is the Hadamard product [19].

III. Experimental Setup and Dataset

A. Magnetically Actuated Soft Capsule Endoscopes (MASCE)

Our capsule prototype is a magnetically actuated soft capsule endoscope (MASCE) designed for disease detection, drug delivery and biopsy operations in the upper GI-tract. The prototype is composed of an RGB camera, a permanent magnet, an empty space for drug chamber and a biopsy tool (see Figs. 4 and 5 for visual reference). The magnet exerts magnetic force and torque to the robot in response to a controlled external magnetic field [5]. The magnetic torque and forces are used to actuate the capsule robot and to release drugs. Magnetic fields from the electromagnets generate the magnetic force and torque on the magnet inside the MASCE so that the robot moves inside the workspace. Sixty-four three-axis magnetic sensors are placed at the top of the workspace, and nine electromagnets are placed at the bottom [5].

B. Magnetic Localization System

Our 5-DoF magnetic localization system is designed for the position and orientation estimation of untethered meso-scale magnetic robots [10]. 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. The core idea of our localization technique is separation of capsule's magnetic field from actuator's magnetic field. For that purpose, the part of the magnetic field due to the actuators is subtracted from the magnetic field data which is acquired by Hall-effect sensor array. As a further step, second-order directional differentiation is applied to reduce the localization error [10].

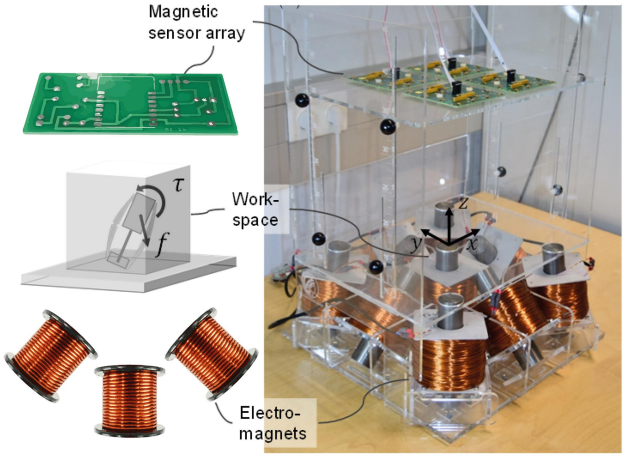


Fig. 5: Actuation system of the MASCE [5], [20]–[23]

C. Monocular Visual Odometry

The visual odometry is performed by minimization of a multi-objective cost function, which includes terms that measure photometric and volumetric correlation. For every input RGB image, we create its depth image using the source code of the perspective shape-from-shading under realistic lighting conditions project [24]. Once the depth map is obtained, the framework uses both RGB and depth map information to jointly estimate camera pose. An energy minimization-based pose estimation technique is applied containing both sparse optical flow (OF) based correspondence establishment, and dense volumetric and photometric alignment [25]. Inspired from the pose estimation strategies proposed by [25], for a parameter vector

$$X = (R_o, t_o, ..., R_{|S|}, t_{|S|})^T \tag{14}$$

for $|S|$ frames, the alignment problem is defined as a variational non-linear least squares minimization problem with the following objective, consisting of the OF based pixel correspondences and dense jointly photometric-geometric constraints [25]. Outliers after OF estimation are eliminated using motion bounds criteria, which removes pixels with a very large displacement and motion vectors too different from neighbouring pixels. The energy minimization equation is as follows:

$$E_{\text{align}}(X) = \omega_{\text{sparse}} E_{\text{sparse}}(X) + \omega_{\text{dense}} E_{\text{dense}}(X) \tag{15}$$

where ω_{sparse} and ω_{dense} are weights assigned to sparse and dense matching terms and E_{sparse} (X) and $E_{\text{dense}}(X)$ are the sparse and dense matching terms, respectively. The sparse matching term is

$$E_{\text{sparse}}(X) = \sum_{\text{i=1}}^{|S|} \sum_{\text{j=1}}^{|S|} \sum_{\text{(k,1)} \in \text{C(i,j)}} ||\tau_i P_{\text{i,k}} - \tau_j P_{\text{j,k}}||^2 \tag{16}$$

Here, $P_{\text{i, k}}$ is the k^{th} detected feature point in the i-th frame. $C(i, j)$ is the set of all pairwise correspondences between the i-th and the j-th frame. The Euclidean distance over all the detected feature matches is minimized once the best rigid transformation τ_i is found. Dense pose estimation is described as follows [25]:

$$E_{\text{dense}}(\tau) = \omega_{\text{photo}} E_{\text{photo}}(\tau) + \omega_{\text{geo}} E_{\text{geo}}(\tau) \tag{17}$$

where,

$$E_{\text{photo}}(X) = \sum_{\text{(i,j)} \in \text{E}} \sum_{\text{k=0}}^{|I_i|} ||I_i(\omega(d_{\text{i,k}})) - I_j(\omega(\tau_j^{-1} \tau_i d_{\text{i,k}}))||_2^2 \tag{18}$$

and,

$$E_{\text{geo}}(X) = \sum_{\text{(i,j)} \in \text{E}} \sum_{\text{k=0}}^{|D_i|} [n_{\text{i,k}}^T (d_{\text{i,k}} - \tau_i^{-1} \tau_j \omega^{-1} (D_j(\omega(\tau_j^{-1} \tau_i d_{\text{i,k}}))))]^2 \tag{19}$$

with τ_i being the rigid camera transformation, $P_{\text{i,k}}$ the k^{th} detected inlier point in i^{th} frame, and $C(i, j)$ being the set of pairwise correspondences between the i^{th} and j^{th} frame. In Equation 15, ω_{dense} is linearly increased; this allows the sparse term to first find a good global structure, which is then refined with the dense term (coarse-to-fine alignment [26]). Using Gauss-Newton optimization, we find the best pose parameters X which minimizes the proposed highly non-linear least squares objective.

D. Dataset

We created our own dataset, which was recorded on five different real pig stomachs. To ensure that our algorithm is not tuned to a specific camera model, four different commercial endoscopic cameras were employed. For each pig stomach-camera combination, 2,000 frames were acquired which makes for four cameras, five pig stomachs, and a total of 40,000 frames. Sample images from the dataset are shown in Fig. 6 for visual reference. An Optitrack motion tracking system consisting of eight Prime-13 cameras and the manufacturer's tracking software was utilized to obtain 6-DoF pose measurements (see Fig. 4) as a ground truth for the evaluations of the pose estimation accuracy. The capsule robot was moved via the magnetic actuation system, with an effort to obtain a large range of poses, during which data was simultaneously recorded from the magnetic localization system, the on-board video camera, and the Optitrack system. We divided our dataset into two groups. A first group consisting of 30,000 frames was used for RNN training purposes, whereas the remaining 10,000 frames were used for testing.

E. LSTM Training

The training data is divided into pose sequences of length 50, which are passed into the LSTM module with the expectation that it predicts the next 6-DoF pose value, i.e. the 51st pose measurement, which was used to compute the cost function for training. The LSTM module was trained using the Keras library with GPU programming and the Theano back-end. Using the back-propagation-through-time method, the weights of hidden units were trained for up to 200 epochs with an initial learning rate of 0.001. Overfitting was prevented using dropout and early stopping techniques. The dropout regularization technique, introduced in [27], is

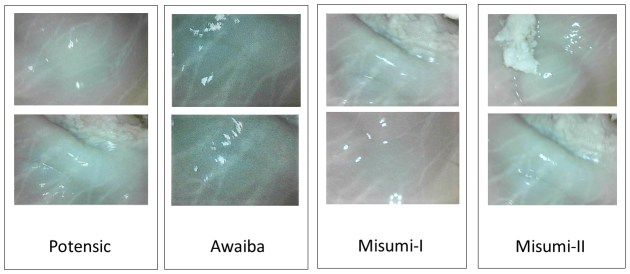


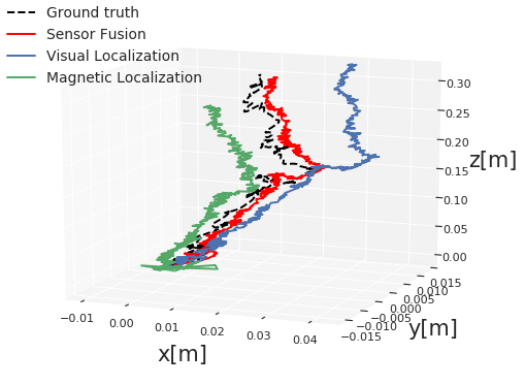
Fig. 6: Sample frames from the dataset used in the experiments.

an extremely effective and simple method to avoid overfitting. It samples a part of the whole network and updates its parameters based on the input data. Early stopping is another widely used technique to prevent overfitting of a complex neural network architecture which was optimized by a gradient-based method.

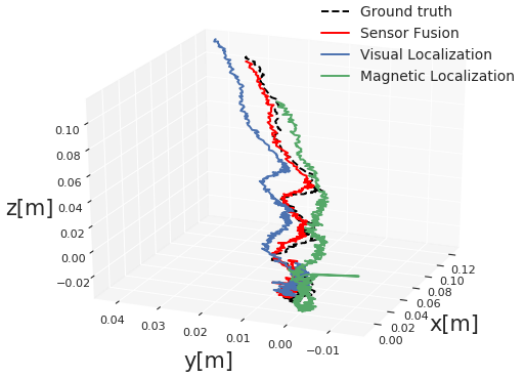
IV. Results and Discussion

The performance of the proposed multi-sensor fusion approach was analysed by examining posterior probabilities of the switch parameters $\mathbf{s}_{k,t}$ (see Fig. 8), the minimum mean square error (MMSE) estimates of $\alpha_{k,t}$ (see Fig. 8) and evolution of the hyper-parameter $\sigma^{\alpha}_{k,t}$ (see Fig. 9) using 200 particles which is determined experimentally. The computational time required to update the state is 26 ms on average. For various trajectories with different complexity of motion, including uncomplicated paths with slow incremental translations and rotations, comprehensive scans with many local loop closures and complex paths with sharp rotational and translational movements, we analysed both the localization accuracy and the fault detection performance of our multi-sensor fusion approach (see Figs. 7 and 10). Additionally, we compared the rotational and translational motion estimation accuracy of the multi-sensor fusion approach with the visual localization and magnetic localization (see Fig. 10) using RMSE.

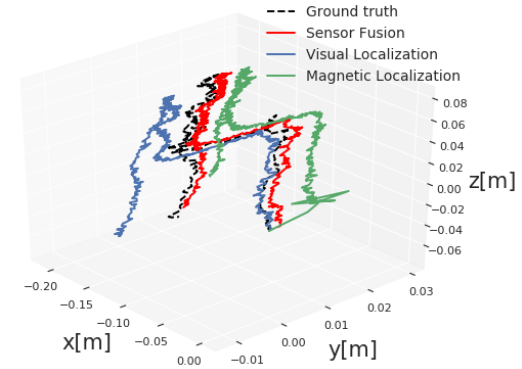
The results in Fig. 8 indicate that the sensor states are accurately estimated. Visual localization failed because of very fast frame-to-frame motions between 14-36 seconds and magnetic sensor failed due to the increased distance of the ring magnet to the sensor array between 57-76 seconds. Both failures are detected successfully, and the MMSE is kept low, thanks to the switching option ability from one observation model to another in case of a sensor failure. In our model, we do not make a Markovian assumption for the switch variable $\mathbf{s}_{k,t}$ but we do for its prior $\alpha_{k,t}$, resulting in a priori dependence on the past trajectory sections, which is more likely for the incremental endoscopic capsule robot motions. Our model thus introduces a memory over the past sensor states rather than simply considering the last state. The length of the memory is tuned by the hyper-parameters $\sigma^{\alpha}_{k,t}$, leading to a long memory for large values and vice-versa. This is of particular interest when considering sensor failures. Our system is designed to automatically



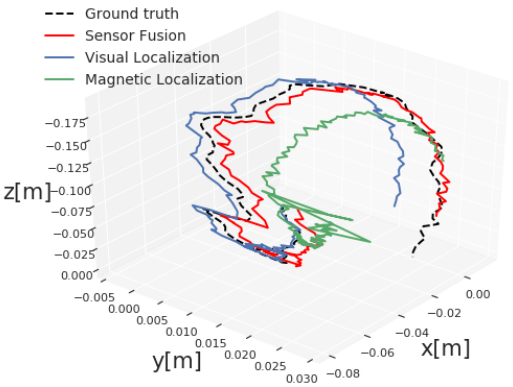
(a) Trajectory 1



(b) Trajectory 2



(c) Trajectory 3



(d) Trajectory 4

Fig. 7: Sample trajectories comparing the multi-sensor fusion result with ground truth and sensor data.

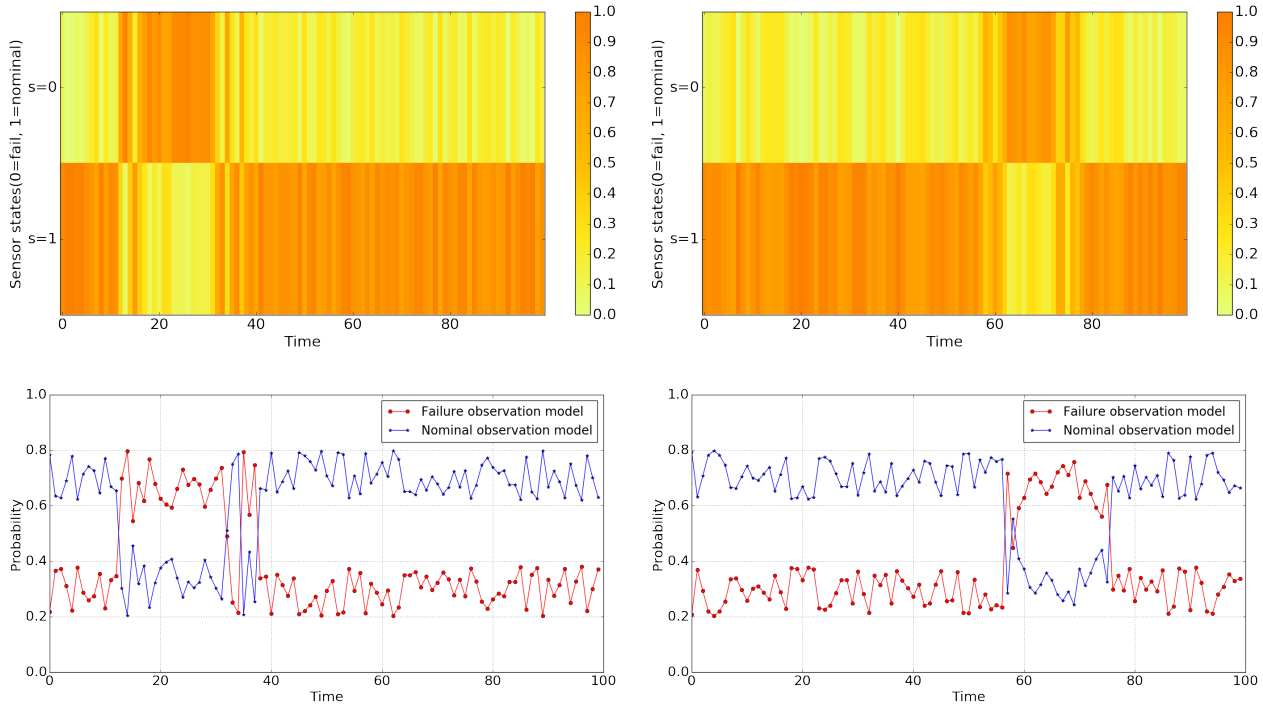


Fig. 8: Top figures: Posterior probability of $\mathbf{s}_{k,t}$ parameter for endoscopic RGB camera (left) and for magnetic localization system (right). Bottom figures: The minimum mean square error (MMSE) of $\alpha_{k,t}$ for endoscopic RGB camera (left) and for magnetic localization system (right). The switch parameter, $\mathbf{s}_{k,t}$, and the confidence parameter $\alpha_{k,t}$ reflect the failure times accurately: Visual localization fails between 14-36 seconds and magnetic sensor fails between 57-76 seconds. Both failures are detected confidentially.

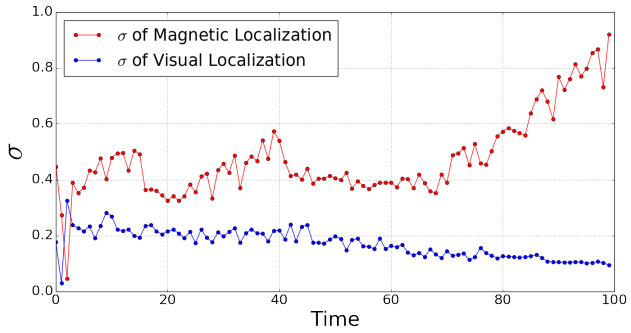


Fig. 9: Evolution of the $\sigma^{\alpha}_{k,t}$ parameter for the sensors. $\sigma^{\alpha}_{k,t}$ does not tend to increase during sensor failure periods.

detect failure states. For example, the confidence in the RGB sensor decreases when visual localization fails recently due to occlusions, fast-frame-to frame changes etc. On the other hand, the confidence in the magnetic sensor decreases if the magnetic localization fails due to magnetic interference from the environment or if the ring magnet has a big distance to the magnetic sensor array.

The results depicted in Figs. 7 indicate that the proposed fusion technique clearly outperforms either the magnetic or visual localization approaches, in terms of both translational and rotational pose estimation accuracy. The multi-sensor

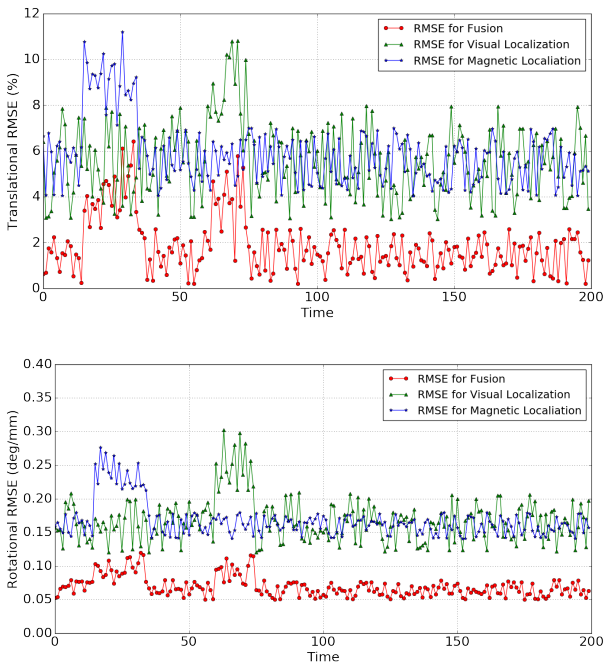


Fig. 10: Translational (top) and rotational (bottom) RMSEs for multi-sensor fusion, visual localization and magnetic localization.

fusion approach is able to stay close to the ground truth pose despite sensor failures. Even for very fast and challenging paths that can be seen in Fig. 7c and 7d, the deviations of the sensor fusion approach from the ground-truth still remain in an acceptable range for medical operations. We presume that the effective use of switching observations and particle filtering with non-linear motion estimation using LSTM enabled learning the motion dynamics very effectively, but this was not explicitly examined.

V. Future clinical applications of actively controlled capsule robot

Capsule endoscopy is primarily utilized to monitor GI tract organs esophagus, stomach, bowels and colon. However, current capsule endoscopy technology is actuated by passive peristaltic motions of the GI tract, which is non-optimal for disease diagnosis and also prevents any type of targeted therapeutic intervention. Recent advances in that field have enabled active manipulation and other therapeutic functionalities such as drug delivery, biopsy operations etc. We envision that the proposed tracking technique would facilitate these advanced functionalities by providing both enhanced situational awareness for the remote operator and more accurate feedback to the magnetic control system.

VI. Conclusions

In this study, we have presented, to the best of our knowledge, the first particle filter-based multi-sensor data fusion approach with sensor failure detection and observation switching capability for endoscopic capsule robot localization. An LSTM architecture was used for non-linear motion model estimation of the capsule robot. The proposed system results in sub-millimetric accuracy for position measurement and sub-degree scale accuracy for orientation measurement. Moreover, it clearly outperforms vison- or magnetic-based tracking alone. As a future step, we plan to integrate a deep learning based noise-variance modelling functionality into our approach to eliminate sensor noise more effectively.

VII. Acknowledgements

H.G. thanks the Alexander von Humboldt Foundation for funding support. This work is funded by the Max Planck Society.

References

[1] 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.

[2] 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, 2017.

[3] M. Turan, Y. Almalioglu, E. Konukoglu, and M. Sitti, "A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots," *arXiv preprint arXiv:1705.05435*, 2017.

[4] F. Munoz, G. Alici, and W. Li, "A review of drug delivery systems for capsule endoscopy," *Advanced drug delivery reviews*, vol. 71, pp. 77–85, 2014.

[5] D. Son, M. D. Dogan, and M. Sitti, "Magnetically actuated soft capsule endoscope for fine-needle aspiration biopsy," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1132–1139.

[6] 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*, pp. 1–15, 2017.

[7] M. Turan, Y. Y. Pilavci, R. Jamiruddin, H. Araujo, 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," *arXiv preprint arXiv:1705.06524*, 2017.

[8] 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.

[9] F. Carpi, N. Kastelein, M. Talcott, and C. Pappone, "Magnetically controllable gastrointestinal steering of video capsules," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 231–234, 2011.

[10] 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.

[11] G. Bao, K. Pahlavan, and L. Mi, "Hybrid localization of microrobotic endoscopic capsule inside small intestine by data fusion of vision and rf sensors," *IEEE Sensors Journal*, vol. 15, no. 5, pp. 2669–2678, 2015.

[12] I. Umay and B. Fidan, "Adaptive magnetic sensing based wireless capsule localization," in *Medical Information and Communication Technology (ISMICT), 2016 10th International Symposium on*. IEEE, 2016, pp. 1–5.

[13] ——, "Adaptive wireless biomedical capsule tracking based on magnetic sensing," *International Journal of Wireless Information Networks*, vol. 24, no. 2, pp. 189–199, 2017.

[14] J. D. Gumprecht, T. C. Lueth, and M. B. Khamesee, "Navigation of a robotic capsule endoscope with a novel ultrasound tracking system," *Microsystem technologies*, vol. 19, no. 9-10, pp. 1415–1423, 2013.

[15] 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.

[16] C. Hue, J.-P. Le Cadre, and P. Perez, "Sequential monte carlo methods for multiple target tracking and data fusion," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 309–325, 2002.

[17] W. R. Gilks and P. Wild, "Adaptive rejection sampling for gibbs sampling," *Applied Statistics*, pp. 337–348, 1992.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[20] 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.

[21] S. Yim, K. Goyal, and M. Sitti, "Magnetically actuated soft capsule with the multimodal drug release function," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 4, pp. 1413–1418, 2013.

[22] S. Yim and M. Sitti, "Shape-programmable soft capsule robots for semi-implantable drug delivery," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1198–1202, 2012.

[23] ——, "Design and rolling locomotion of a magnetically actuated soft capsule endoscope," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 183–194, 2012.

[24] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Metric depth recovery from monocular images using shape-from-shading and specularities," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 25–28.

[25] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph." Robotics: Science and Systems, 2015.

[26] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 3, p. 24, 2017.

[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

Machine Vision and Applications (2018) 29:345–359
https://doi.org/10.1007/s00138-017-0905-8

ORIGINAL PAPER

Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots

Mehmet Turan[1,5] · Yusuf Yigit Pilavci[2] · Ipek Ganiyusufoglu[3] · Helder Araujo[4] · Ender Konukoglu[5] · Metin Sitti[1]

Received: 10 December 2016 / Revised: 11 November 2017 / Accepted: 28 November 2017 / Published online: 27 December 2017

Abstract
Despite significant progress achieved in the last decade to convert passive capsule endoscopes to actively controllable robots, robotic capsule endoscopy still has some challenges. In particular, a fully dense three-dimensional (3D) map reconstruction of the explored organ remains an unsolved problem. Such a dense map would help doctors detect the locations and sizes of the diseased areas more reliably, resulting in more accurate diagnoses. In this study, we propose a comprehensive medical 3D reconstruction method for endoscopic capsule robots, which is built in a modular fashion including preprocessing, keyframe selection, sparse-then-dense alignment-based pose estimation, bundle fusion, and shading-based 3D reconstruction. A detailed quantitative analysis is performed using a non-rigid esophagus gastroduodenoscopy simulator, four different endoscopic cameras, a magnetically activated soft capsule robot, a sub-millimeter precise optical motion tracker, and a fine-scale 3D optical scanner, whereas qualitative ex-vivo experiments are performed on a porcine pig stomach. To the best of our knowledge, this study is the first complete endoscopic 3D map reconstruction approach containing all of the necessary functionalities for a therapeutically relevant 3D map reconstruction.

Keywords Endoscopic capsule robots · 3D map reconstruction · Sparse-then-dense feature tracking

1 Introduction

Many diseases necessitate access to the internal anatomy of the patient for diagnosis and treatment. Since direct access to most anatomic regions of interest is traumatic, and sometimes impossible, endoscopic cameras have become a common method for viewing the anatomical structure. In particular, capsule endoscopy has emerged as a promising new technology for minimally invasive diagnosis and treatment of gastrointestinal (GI) tract diseases. The low invasiveness and high potential of this technology have led to substantial investment in their development by both academic and industrial research groups, such that it may soon be feasible to produce a robotic capsule endoscope with most of the functionality of current flexible endoscopes.

Although robotic capsule endoscopy has high potential of diagnostic and therapeutic capabilities, it continues to face many challenges. In particular, there is no broadly accepted approach for generating a comprehensive and therapeutically relevant 3D map of the organ being investigated. This problem is made more severe by the fact that such a map may require a precise localization method for the endoscope, and such a method will itself require a map of the organ, a classic chicken-and-egg problem [1]. The repetitive texture, lack of distinctive features, and specular reflections characteristic of the GI tract exacerbate this difficulty, and the non-rigid deformations introduced by peristaltic motions further complicate the reconstruction task [2]. Finally, the small size of endoscopic camera systems implies a number of limitations, such as restricted fields of view (FOV), low signal-to-noise ratio, and low frame rate; all of which degrade image quality [3]. These issues, to name a few, make accurate and precise

Mehmet Turan
mturan@student.ethz.ch

1 Physical Intelligence Department, Max-Planck Institute for Intelligent Systems, Stuttgart, Germany

2 Electrical and Electronics Engineering Department, Middle East Technical University, Ankara, Turkey

3 Computer Science and Engineering Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

4 Institute for Systems and Robotics, Universidade de Coimbra, Coimbra, Portugal

5 Computer Vision Laboratory, ETH Zurich, Zürich, Switzerland

localization and reconstruction a difficult problem and can render navigation and control counterintuitive [4].

Despite these challenges, accurate and robust three-dimensional (3D) mapping of patient-specific anatomy remains a difficult goal. Such a map would provide doctors with a reliable measure of the size and location of a diseased area, thus allowing more intuitive and accurate diagnoses. In addition, should next-generation medical devices be actively controlled, a map would dramatically improve the doctors control in diagnostic, prognostic, and therapeutic operations [5]. As such, considerable energy has been devoted to adapt computer vision techniques to the problem of in vivo 3D reconstruction of tissue surface geometry.

Two primary approaches have been pursued as workarounds for the challenges mentioned previously. First, tomographic intra-operative imaging modalities, such as ultrasound (US), intra-operative computed tomography (CT), and interventional magnetic resonance imaging (iMRI), have been investigated for capturing detailed information of patient-specific tissue geometry [5]. However, surgical and diagnostic operations pose significant technological challenges and costs for the use of such devices, due to the need to acquire a high signal-to-noise ratio (SNR) without impediment to the doctor. Another proposal has been to equip endoscopes with alternative sensor systems in the hope of providing additional information; however, these alternative systems have other restrictions that limit their use within the body.

This paper proposes a complete pipeline for 3D visual map reconstruction using only RGB camera images, with no additional sensor information. The pipeline is arranged in a modular form and includes a preprocessing module for removal of specular reflections, vignetting and radial lens distortions, a keyframe selection module, a pose estimation and image stitching module for registration of images, and a shape-from-shading (SfS) module for reconstruction of 3D structures. We provide both qualitative and quantitative analysis of pose estimation and 3D map reconstruction accuracy using a porcine pig stomach, an esophagus gastroduodenoscopy simulator, four different endoscopic camera models, an optical motion tracker, and a 3D optical scanner. In sum, our method proposes a substantial contribution toward a more general, therapeutically relevant, and extensive use of the information that capsule endoscopes may provide.

2 Literature survey

Several studies in the literature have discussed 3D map reconstruction for standard hand-held and passive capsule endoscopes [6–13], etc. These methods may be broken into four major classes, i.e.,

- stereoscopy
- shape from shading (SfS)
- structured light (SL)
- time of flight (ToF)

Structured light and time-of-flight methods require additional sensors, with a concomitant increase in cost and space; as such, they are not covered in this paper. Stereo-based methods use the parallax observed when viewing a scene from two distinct viewpoints to obtain an estimate of the distance from observer to object under observation. Typically, such algorithms have four stages in computing the disparity map [14]: cost computation, cost aggregation, disparity computation and optimization, and disparity refinement.

With multiple algorithms reported per year, computational stereo depth perception has become an extremely researched field. The first work reporting stereoscopic depth reconstruction in endoscopic images was the work done by [6], which implemented a dense computational stereo algorithm. Later, Hager et al. developed a semi-global optimization [7], which was used to register the depth map acquired during surgery to preoperative models [8]. Stoyanov et al. used local optimization to propagate disparity information around feature-matched seed points, and it has also been reported to perform well for endoscopic images. This method was able to handle highlights, occlusions, and noisy regions. Similar to stereo vision, another method that employs epipolar geometry and feature extraction is also proposed in [15]. This work flow starts with camera calibration, and it relies on SIFT extraction and feature description. Finally, the main algorithm calculates the 3D spatial point location using extrinsic parameters, which is calculated from matched features in consecutive frames. Although this system exploits the advantage of sparse 3D reconstruction, the strong dependency on feature extraction causes performance-related issues for endoscopic type of imaging. Despite the variety of algorithms and simplicity of implementation, computational stereo techniques are affected by several important disadvantages. To begin with, stereo reconstruction algorithms generally require two cameras, since the triangulation needs a known baseline between viewpoints. Further, the accuracy of triangulation decreases with distance from the cameras due to the shrinkage of relative baseline between camera centers and reconstructed points. Most endoscopic capsule robots have only one camera, and in those that have more, the diameter of endoscope inherently bounds the baseline. As such, stereo techniques have yet to find wide application in endoscopy.

Due to the difficulty in obtaining stereo-compatible hardware, efforts have been made to adapt passive monocular three-dimensional reconstruction techniques to endoscopic images. These techniques have been focused on research in computer vision for decades and have the distinct advan-

tage of not requiring extra hardware equipment in addition to existing endoscopic devices. Two main methods have emerged as useful in the field of endoscopic images: shape from motion (SfM) and shape from shading (SfS). SfS, which has been studied since the 1970s [16], has demonstrated some suitability for endoscopic image reconstruction. Its primary assumption is that there is a single light source on the scene, of which the intensity and pose relative to the camera are known. Both assumptions are mostly fulfilled in endoscopy [11–13]. Furthermore, the transfer function of the camera can be included in the algorithm to additionally refine estimates [17]. Additional assumptions are that the object reflects light obeying lambertian model and that the object surface has a constant albedo. If these assumptions hold to a degree and the equation parameters are known, SfS can use the brightness of a pixel to estimate the angle between cameras depth axis and the shape normal at that pixel. This has been demonstrated to be effective in recovering details, although global shape recovery often fails.

Both methods have been demonstrated to have disadvantages: SfS often fails in the presence of uncertain information, e.g., bleeding, reflections, noise artifacts, and occlusions; feature tracking-based SfM methods tend to fail in the presence of poorly textured areas and occlusions.

Therefore, many state-of-the-art works are mainly based on the combination of these two techniques: In [18], a pipeline for 3D reconstruction of endoscopy imaging using SfS and SfM techniques is presented. In this work, the pipeline starts with basic preprocessing steps and focuses on 3D map reconstruction, which is independent of light source position and illumination. Finally, the framework ends with frame-to-frame feature matching to solve the scaling issue of monocular images. This paper proposes interesting methods for the difficult task of reconstruction. However, enhanced preprocessing and especially less dependency on feature extraction and matching are still needed. In the recent work of [19], SfS and SfM are fused together to reach a better 3D map accuracy. With SFM, a sparse point cloud is obtained and a dense version of this cloud is generated by means of SFS. For better performance of SFS, they also propose a refined reflectance model. One notable idea based on SfS and SfM fusion is proposed in [20]. This methodology first reconstructs a sparse 3D map using SfM and iteratively refines the final reconstruction using SfS. The approach does not directly address the difficulties caused by the ill-posed illumination and specular reflectance, although the proposed geometric fusion tries to eliminate such issues. And the strong reliance on the establishment of feature correspondence remains unsolved. Attempts to solve the latter problem with template-matching techniques have had some success, but tend to be computationally very complex which makes it unsuitable for real-time performance. In [21], only SFS is used for reconstruction and 2D features are preferred for estimating the transformation. Similarly, [22] and [23] combine SFM and SFS for 3D reconstruction without any preprocessing and with the Lambertian surface assumption. In [24], machine learning algorithms are applied for 3D reconstruction. Basically, training is completed with an artificial dataset and real endoscopy images are used for test data. Another state-of-the-art pipeline is proposed in [25], which presents a workflow combining RGB camera and inertial measurement sensors (IMU). Besides improved results, this hardware makes the overall flow more complex and costly. Moreover, IMU sensors occupy extra place and they are not accurate enough. In addition, they interfere with the magnetic actuation systems which makes them unsuitable for the next generation of actively controllable endoscopic capsule robots. The main common issue remaining for 3D reconstruction of endoscopic-type datasets is the visual complexity of these images. The challenges which we mentioned in the abstract and introduction affect the performance of standard computer vision algorithms. In particular, the proposed method must be robust to specular view-dependent highlights, noise, peristaltic movements, and focus-dependent changes in calibration parameters. Unfortunately, a quantitative measure of algorithm robustness has not been suggested in the literature until today, despite its clear value for the evaluation of algorithmic dependability and precision. Moreover, all of the mentioned methods in that section were developed and evaluated on only one specific camera model, which makes it impossible to justify the robustness of the framework in the case of different camera choices with limited specifications such as lower resolution and image quality.

Our paper proposes a full pipeline consisting of camera calibration, reflection detection and suppression, radial undistortion, de-vignetting, keyframe selection, pose estimation, frame stitching, and SfS to reconstruct a therapeutically relevant 3D map of the organ under observation. Both synthetic and real pig stomachs are used for evaluation. Among other contributions, an extensive quantitative analysis has been proposed and performed to demonstrate the influence of pipeline modules on the accuracy and robustness of the estimated camera pose and reconstructed 3D map. To our knowledge, this is the first such comprehensive quantitative analysis to be enacted in endoscopic type of image processing.

3 Method

This section represents the proposed framework in more depth. Preprocessing steps, keyframe selection, pose estimation, frame stitching, and SfS module will be discussed in detail.

3.1 Preprocessing

The proposed modular endoscopic 3D map reconstruction framework starts with a preprocessing module which performs intrinsic camera calibration, reflection detection and suppression, radial distortion correction, and de-vignetting. Specular reflections are a common problem causing inaccurate depth estimation and map reconstruction. Therefore, eliminating specular artifacts is a fundamental endoscopic image preprocessing step to ensure lambertian surface properties and increase the quality of the 3D map. On the other hand, specularities can deliver useful information for pose estimation, especially orientation information. For the reflection detection task, we propose an original method which determines the reflection regions by making use of geometric and photometric information. To determine the locations of the reflection areas, the gradient map of the input gray-scale image is created and a morphological closing operation is applied to fill the gaps inside reflection-distorted areas. For the closing operation, we used OPENCV function close(). In parallel, a photometric method applies adaptive thresholding determined by the mean and standard deviation of the gray-scale image I to identify the specular regions:

$$Mask_{Illu} = \begin{cases} 0, & I < \mu_I + \sigma_I \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

where μ_I and σ_I are the mean and standard deviation of the intensity levels of the gray-scale image I. The pixel-wise combination of both detection strategies leads to a robust reflection detection approach. Once specular reflection pixels are detected, the inpainting method proposed by [26] is applied to suppress the saturated pixels by replacing the specularity by an intensity value derived from a combination of neighboring pixel values.

As a next step, the Brown-Conrady [27] undistortion technique is applied to handle the radial distortions. Vignetting, referring to an inhomogeneous illumination distribution relative to the image center, primarily caused by camera lens imperfections and light source limitations, is handled by applying a radial gradient symmetry enforcement-based method (Fig. 1). Our framework applies the vignetting correction approach proposed by [28] which de-vignettes the image by enforcing the symmetry of the radial gradient from center to boundaries. An example of input image and vignetting-corrected output image can be seen in Fig. 1. De-vignetting is demonstrated in Fig. 2, where it is clearly observable that the intensity levels of de-vignetted image have a more homogeneous pattern.

3.2 Keyframe selection

Endoscopic videos generally contain thousands of highly overlapping frames (more than %75 overlap) due to slow endoscopic capsule movement during organ exploration. A subset of the most relevant keyframes has to be chosen automatically. The minimum amount of key frames required to recover the entire stomach surface with approximately %50 overlapping area between keyframes is around 300 frames. Thus, at least every tenth frame could be selected as a keyframe. However, since the endoscopic capsule robot motion is not constant during organ exploration, it is not a good practice to blindly assign keyframes with a constant interval. We developed an adaptive keyframe selection method based on Farneback optical flow (OF) estimation between frame pairs. Farneback OF is chosen due to its

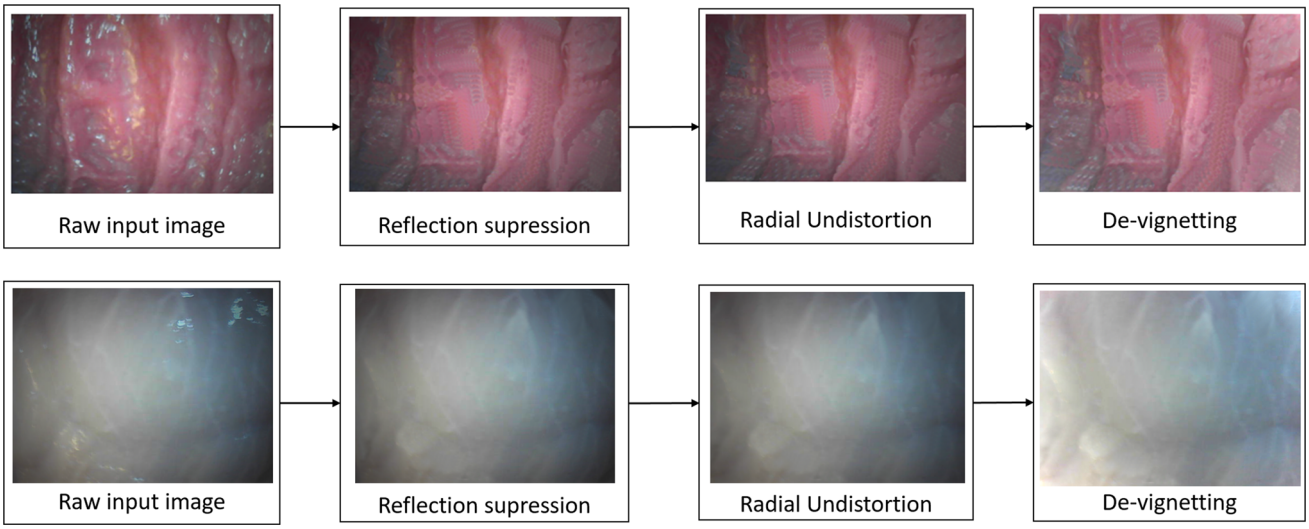


Fig. 1 Preprocessing pipeline: reflection removal, radial undistortion, de-vignetting

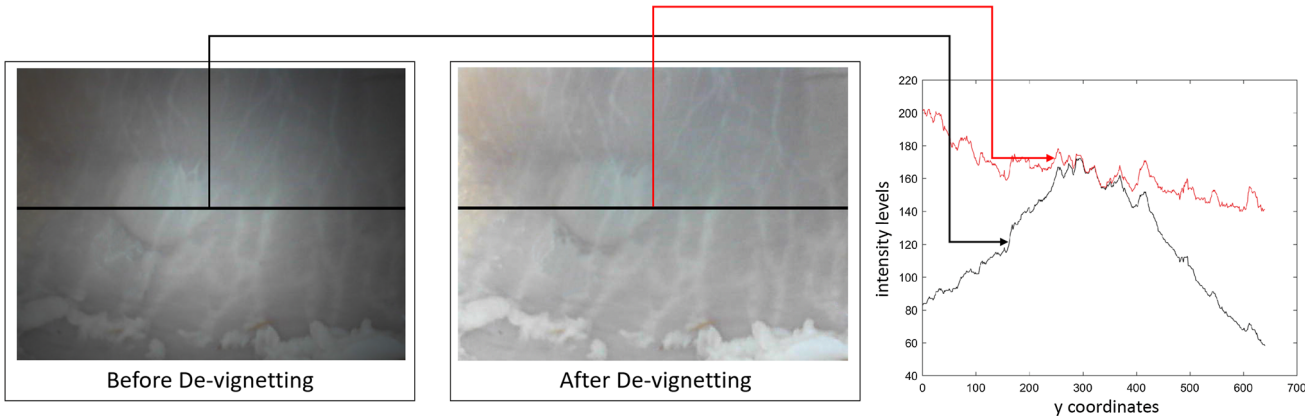


Fig. 2 Demonstration of the de-vignetting process

improved performance relative to other optical flow methods applied to our dataset. We add the magnitudes of optical flow values for each frame pair and normalize the sum by total image resolution. If the normalized sum does not exceed a predefined threshold $\tau = 30$ pixels, the overlap between reference keyframe and keyframe candidate is accepted as being high (more than %70 overlap). In that case, candidate frame fails and the algorithm goes to the next frame. The loop starts again and runs until a keyframe is found. The key frame selection procedure and termination criteria are represented in algorithm 1:

Algorithm 1 Keyframe selection algorithm

1: Extract Farneback optical flow between reference keyframe and candidate keyframe.
2: Sum the magnitude values of the optical flow vectors for each pixel pair.
3: Normalize the sum by total pixel number.
4: If the normalized sum is less than predefined threshold $\tau = 30$ pixels, go to the next frame; else identify the frame as a keyframe and go tho the first step.
5: If fifteen frames failed to fulfill the key frame conditions, and still $\tau = 30$ pixels could not be exceeded, assign the frame with highest τ value among these fifteen frames as a key frame and go to the first step.

3.3 Keyframe stitching

A state-of-the-art image stitching pipeline contains several stages:

- Feature detection, which detects features in input image pair.
- Feature matching, which matches features between input images.
- Homography estimation, which estimates extrinsic camera parameters between the image pairs.
- Bundle adjustment, which is a postprocessing step to correct drifts in a global manner.
- Image warping, which warps the images onto a compositing surface.
- Gain compensation, which normalizes the brightness and contrast of all images.
- Blending, which blends pixels along the stitch seam to reduce the visibility of seams.

Stitching algorithms fall broadly into two categories: direct alignment-based methods and feature-based methods. Direct alignment-based methods attempt to match every pixel between the frame pair using iterative optimization techniques. These methods have the benefit of using all the available data which is a good practice for low-textured images such as endoscopic type of images. However, direct methods require a good initialization so that they do not converge into local minima. Moreover, they are very susceptible to varying brightness conditions. Feature-based methods, on the other hand, first find unique feature points such as corners and try to match them. These methods do not require an initialization, but the features are not easy to detect in low-textured images and detected features can be susceptible to illumination changes, scale changes caused by zoom-in and out and viewpoint changes. Our keyframe stitching technique makes use of both alignment methods in a coarse-to-fine fashion combining Farneback OF-based coarse alignment with patch-wise fine alignment. Farneback OF delivers the initial 2D motion estimation, whereas the SSD-based energy minimization applied to circular regions of interest with a radius of 15 pixels around each inlier point refines this estimation. Patch-wise fine alignment estimates the parameters of affine transformation by minimizing an intensity difference-based energy cost function. The affine transformation maps an image I_1 onto the reference image I_2, where x' , y' represent the transformed and x, y the original pixel coordinates, and a_1, a_2, a_3, a_4, t_x, t_y the parameters of affine transfor-

mation matrix A, respectively. We define a cost function measuring the pixel intensity similarity between the image pair (Eq. 4), which is supposed to be minimized by the corresponding affine transformation parameters.

$$\begin{pmatrix} x_2 \\ y_2 \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & t_x \\ a_3 & a_4 & t_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ y_1 \\ 1 \end{pmatrix} \tag{2}$$

Since the cost function has to ignore the pixels lying outside the circular patches defined around inlier points, a weighting function $w(x, y)$ is defined:

$$\omega(x, y) = \begin{cases} 0, if (x - x_c)^2 + (y - y_c)^2 \geq r^2 \\ 1, if (x - x_c)^2 + (y - y_c)^2 < r^2 \end{cases} \tag{3}$$

where x_c and y_c are the coordinates of inlier point and r the radius of the circular image region around this inlier point center. The resulting cost function has a bias toward smaller overlapping solutions; thus a normalization of it by the overlapping area is necessary, resulting in the mean squared pixel error (MSE):

$$e_{\mathrm{MSE}}(A) = \frac{\sum_i \omega(x_i, y_i)\omega(x_i', y_i')(I_2(x_i', y_i') - I_1(x_i, y_i))^2}{\sum_i \omega(x_i, y_i)\omega(x_i', y_i')}. \tag{4}$$

The affine transformation matrix A is iteratively determined by the image transformation that minimizes e_{MSE} using Gaussian–Newton optimization. CUDA library was utilized to achieve better performance and reduce execution time of GN Optimization through parallelism. The system architecture diagram of the proposed frame stitching algorithm is demonstrated in Fig. 3.

The termination criteria of the Gaussian–Newton optimization were defined by a threshold $\tau = e^{-9}$, whereas the optimization stops when the e_{MSE} drops below the threshold τ or maximum number of iterations have already been reached. Once the optimization has converged and the affine transformation parameters are estimated, bundle adjustment is performed to correct drifts for all the camera parameters jointly and to minimize the accumulative errors. At the next step, all keyframes I_i are transformed into the coordinate system of the anchor keyframe I_A. In areas where several keyframes overlap, corresponding image pixels often do not

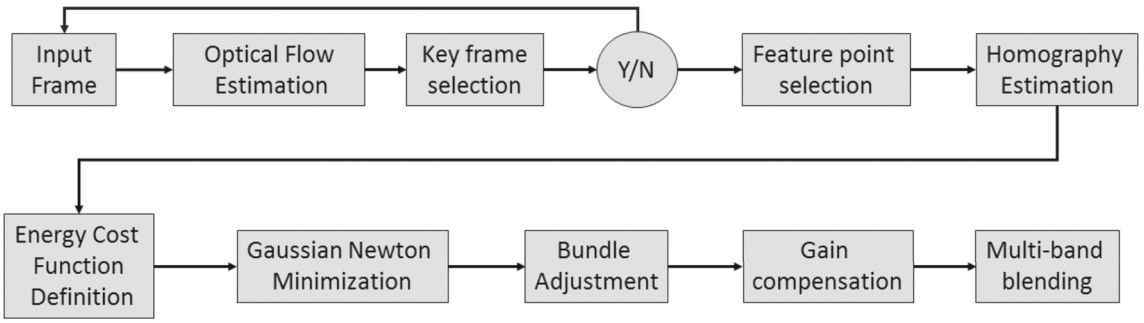


Fig. 3 Image stitching flowchart

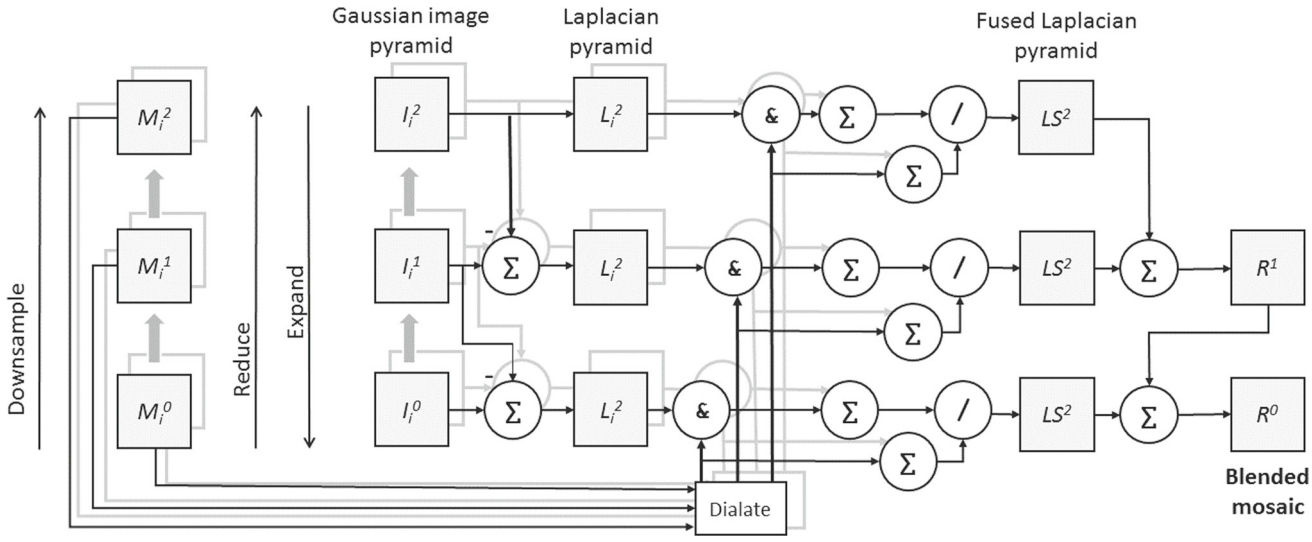


Fig. 4 Multi-band blending flowchart

Fig. 5 Demonstration of the keyframe stitching process for the non-rigid esophagus gastroduodenoscopy simulator (left) and real pig stomach (right)

have the same intensity due to illumination changes, scale changes, and intensity level variations. Multi-band blending method is applied to overcome these issues. The overview of multi-blending approach is shown in Fig. 4. For further details, the reader is referred to the original work of [29]. Algorithm 2 summarizes the steps of keyframe stitching module. Results of the stitching process for the real pig stomach and nonrigid simulator are shown in Fig. 5.

Algorithm 2 Proposed endoscopic keyframe stitching module

1: Identify the next keyframe.
2: Match pixels between the reference keyframe and the identified next keyframe using optical flow estimation.
3: Use RANSAC to detect inlier points.
4: Use optical flow vectors between inlier matches as initialization for the GN optimization.
5: Define circular regions around each inlier point.
6: Calculate the intensity difference-based energy cost function.
7: Execute iterative Gaussian–Newton optimization (GN) to minimize the energy cost function.
8: Perform GPU-based multi-core bundle adjustment to globally optimize all of the camera poses jointly [30].
9: Perform frame warping.
10: Perform gain compensation [31].
11: Perform multi-band blending.

3.4 Deep learning and frame stitching

A major drawback of our frame stitching module is the need for an extensive engineering and implementation effort. To overcome these issues, we investigated the applicability of deep learning techniques to the endoscopic capsule robot pose estimation [2]. Deep learning (DL) has been drawing the attention of the machine learning research community over the last decade. Much of its success roots on having made available models and technologies capable of achieving ground-breaking performances in a variety of traditional fields of application of machine learning, such as machine vision and natural language processing. Admittedly, some of the DL flagships, like NLP and image processing, have their implications in medical fields, e.g., in extracting information from the images taken from patients' records to find anomalous patterns and detect diseases. With that motivation, we are trying to extend the application of DL technology into endoscopic capsule robot localization. The core idea of our DL-based method is the use of deep recurrent convolutional neural networks (RCNNs) for the pose estimation task, where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used for the feature extraction and inference of dynamics across the frames, respectively [2]. Using this pretrained neural network, we are able to achieve pose estimation accuracies comparable our sparse-then-dense pose alignment [2]. Thus, as a future step, we might consider to integrate DL-based pose estimation into our frame stitching module to decrease the complexity of our stitching method and relax the extensive engineering and implementation efforts required in this study. Since DL-based pose estimation is out of scope of this paper, the reader is referred to the original paper [2] for further details.

3.5 Endo-VMFusenet and frame stitching

Even though the proposed sparse-then-dense alignment-based visual pose estimation achieves very promising results for endoscopic capsule robot localization, it fails in case of very fast frame-to-frame motions. This is a common issue of any vision-based odometry algorithm. If the overlap between consecutive frames becomes less than a certain percentage, any vision-based pose estimation approach fails. It can even occur that due to drifts of endoscopic capsule robot, the overlap area between frame pairs decreases drastically, which can even be zero in some cases. To overcome this issue, we developed a supervised sensor fusion approach based on an end-to-end trainable deep neural network consisting of multi-rate long short-term memories (LSTMs) for frequency adjustment between sensors and a core LSTM unit for fusion of the adjusted sensor information. Detailed evaluations indicate

that our pretrained DL-based sensor fusion network detects whether visual odometry fails and instantaneously makes use of magnetic localization until visual odometry path again recovers. The same applies if magnetic sensor-based localization fails. Additionally, monocular cameras suffer with the absence of real depth information which causes any measurements made by them to be recoverable only up to a scale. This condition is known as scale ambiguity. Another contribution of our DL-based sensor fusion approach is the accurate scale estimation by using absolute position information obtained by the magnetic localization system. In that way, doctors will have a 3D map of exactly same size of the explored inner organ, which will not only help the exact estimation of the diseased region size, but also enable biopsy-like treatments or local drug delivery onto the diseased region. Since it is out of scope, for further details of our DL-based sensor fusion approach, the reader is referred to our paper [4].

3.6 Depth image creation

Once the final mosaic image is obtained, the next module creates its depth image using the SfS technique of Tsai and Shah [32]. Tsai–Shah SfS method is based on the following assumptions:

- The object surface is lambertian.
- The light comes from a single-point light source.
- The surface has no self-shaded areas.

Lambertian surface assumption is not obeyed by raw endoscopic images due to the specular reflections inside the organs. We addressed this problem through the reflection suppression technique previously described. Subsequently, the above assumptions allow the image intensities to be modeled by

$$I(x, y) = \rho(x, y, z) \cdot \cos \Theta_i, \tag{5}$$

where I is the intensity value, p is the albedo (reflecting power of surface), and theta is the angle between surface normal N and light source direction S. With this equation, the gray values of an image I are related only to albedo and angle theta. Using these assumptions, the above equation can be rewritten as follows:

$$I(x, y) = \rho \cdot N.S, \tag{6}$$

where (.) is the dot product, N is the unit normal vector of the surface, and S is the incidence direction of the source light. These may be expressed respectively as

$$N = \frac{(-p(x, y), -q(x, y), 1)}{(p^2 + q^2 + 1)^{(1/2)}} \tag{7}$$

$$S = (\cos \tau \cdot \sin \sigma, \sin \tau \cdot \sin \sigma, \cos \sigma) \tag{8}$$

where (τ) and (σ) are the slant and tilt angles, respectively, and p and q are the x and y gradients of the surface Z:

$$p(x, y) = \frac{\partial Z(x, y)}{\partial x} \tag{9}$$

$$q(x, y) = \frac{\partial Z(x, y)}{\partial y}. \tag{10}$$

The final function then takes the form

$$\begin{aligned} &I(x, y) \\ &\quad = \rho \cdot \frac{(\cos \sigma + p(x, y) \cdot \cos \tau \cdot \sin \sigma + q(x, y) \cdot \sin \tau \cdot \sin \sigma)}{((p(x, y))^2 + (q(x, y))^2 + 1)^{(1/2)}} \\ &\quad = R(p_{\mathrm{x,y}}, q_{\mathrm{x,y}}). \end{aligned} \tag{11}$$

Solving this equation for p and q essentially corresponds to the general problem of SfS. The approximations and solutions for p and q yield the reconstructed surface map Z. The necessary parameters are tilt, slant, and albedo, and can be estimated as proposed in [33]. The unknown parameters of the 3D reconstruction are the horizontal and vertical gradients of the surface Z, p, and q. With discrete approximations, they can be written as follows:

$$p(x, y) = Z(x, y) - Z(x - 1, y) \tag{12}$$

$$q(x, y) = Z(x, y) - Z(x, y - 1), \tag{13}$$

where $Z(x, y)$ is the depth value of each pixel. From these approximations, the reflectance function $R(p_{\mathrm{x,y}}, q_{\mathrm{x,y}})$ can be expressed as

$$R(Z(x, y) - Z(x - 1, y), Z(x, y) - Z(x, y - 1)). \tag{14}$$

Using equations 12, 13, and 14, the reflectance equation may also be written as

$$\begin{aligned} &f(Z(x, y), Z(x, y - 1), Z(x - 1, y), I(x, y)) \\ &\quad = I(x, y) - R(Z(x, y) - Z(x - 1, y), \\ &\quad Z(x, y) - Z(x, y - 1)) = 0. \end{aligned} \tag{15}$$

Tsai and Shah proposes a linear approximation using a first-order Taylor series expansion for function f and for depth map Z^{n-1}, where Z^{n-1} is the recovered depth map after $n-1$ iterations. The final equation is

$$Z^n(x, y) = Z^{(n-1)}(x, y) - \frac{f(Z^{(n-1)}(x, y))}{\frac{\mathrm{d}(f(Z^{(n-1)}(x, y))}{\mathrm{d}(Z(x, y))}}, \tag{16}$$

where f is a predefined function, constrained by

$$\frac{\mathrm{d}f(Z^{(n-1)}(x, y))}{\mathrm{d}Z(x, y)}(1 + i_x^2 + i_y^2)) \tag{17}$$

and

$$i_x = \cos\tau \cdot \frac{\sin\sigma}{\cos\sigma} \tag{18}$$

$$i_y = \sin\tau \cdot \frac{\sin\sigma}{\cos\sigma}. \tag{19}$$

The nth depth map Z^n is calculated by using the estimated slant, tilt, and albedo values.

4 Evaluation

We evaluate the performance of our system both quantitatively and qualitatively in terms of pose estimation and surface reconstruction. We also report the computational complexity of the proposed framework.

4.1 Dataset

We created our own dataset from a real pig stomach and from a non-rigid open GI tract model EGD (esophagus gastroduodenoscopy) surgical simulator LM-103 (Figs. 6, 7). The EGD surgical simulator was used for quantitative analyses, and the real pig stomach for qualitative evaluations. Synthetic stomach fluid was applied to the surface of the EGD simulator to imitate the mucosa layer of the inner tissue. To ensure that our algorithm is not tuned to a specific camera model, four different commercially available endoscopic cameras were employed for the video capture varying in their resolution, pixel size, depth of focus, and image quality. A total of 17010 endoscopic frames were acquired by these four camera models which were mounted on our robotic magnetically actuated soft capsule endoscope prototype (MASCE) (Fig. 8, [34,35]). The first sub-dataset, consisting of 4230 frames, was acquired with an Awaiba NanEye camera (Table

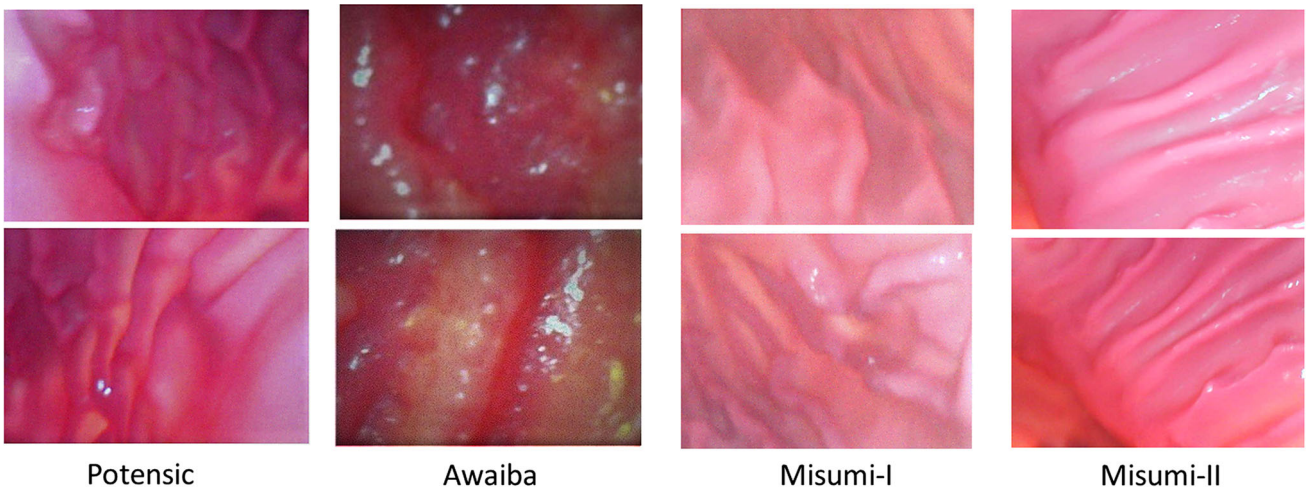


Fig. 6 Non-rigid esophagus gastroduodenoscopy simulator dataset overview for different endoscopic cameras

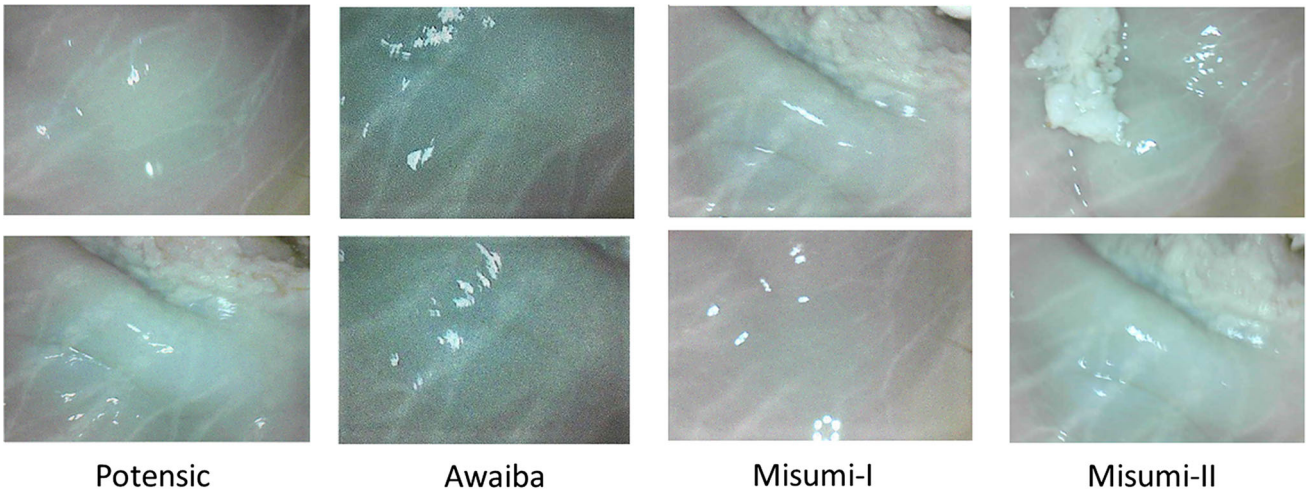


Fig. 7 Real pig stomach dataset overview for different endoscopic cameras

Fig. 8 Robotic magnetically actuated soft capsule endoscopes (MASCE) [34,35]

Table 1 Awaiba Naneye monocular endoscopic camera

Resolution	250 × 250 pixel
Footprint	2.2 × 1.0 × 1.7 mm
Pixel size	$3 \times 3\,\mu m^2$
Pixel depth	10 bit
Frame rate	44 fps

Table 2 Misumi-V3506-2ES monocular camera

Resolution	400 × 400 pixel
Diameter	8.2 mm
Pixel size	$5.55 \times 5.55\,\mu m^2$
Pixel depth	10 bit
Frame rate	30 fps

Table 3 Misumi-V5506-2ES monocular camera

Resolution	640 × 480 pixel
Diameter	8.6 mm
Pixel size	$6.0 \times 6.0\,\mu m^2$
Pixel depth	10 bit
Frame rate	30 fps

Table 4 Potensic monocular mini camera

Resolution	1280 × 720 pixel
Diameter	8.8 mm
Pixel size	$10.0 \times 10.0\,\mu m^2$
Pixel depth	10 bit
Frame rate	30 fps

1). The second sub-dataset, consisting of 4340 frames, was acquired by the Misumi V3506-2ES endoscopic camera with the specification shown in Table 2. The third sub-dataset of 4320 frames was obtained by the Misumi V5506-2ES endoscopic camera with the specification shown in Table 3. Finally, the fourth sub-dataset of 4120 frames was obtained by the Potensic mini camera with the specification shown in Table 4. We scanned the open stomach simulator using the 3D Artec Space Spider image scanner and used this 3D scan as the ground truth for the 3D map reconstruction framework (Fig. 9). Even though our focus and ultimate goal is an accurate and therapeutically relevant 3D map reconstruction, we also evaluated the pose estimation accuracy of the proposed framework quantitatively since a precise pose estimation is a prerequisite for an accurate 3D mapping. Thus, an Optitrack motion-tracking system consisting of eight Prime-13 cameras and a tracking software was utilized to obtain a 6-DoF localization ground truth data of the endoscopic capsule motion with a sub-millimeter precision (Fig. 9).

4.2 Trajectory estimation

To evaluate the pose estimation performance, we tested our system on different trajectories of various difficulty levels. The absolute trajectory (ATE) root-mean-square error metric (RMSE) is used for quantitative pose accuracy evaluations. The absolute trajectory (ATE) root-mean-square error metric measures the root-mean-square of Euclidean distances between the estimated endoscopic capsule robot poses and the ground truth poses estimated by the motion capture system. Table 5 shows the results of the trajectory estimation for six different trajectories. Trajectory 1 is an uncomplicated path with very slow incremental translations and rotations. Trajectory 2 follows a comprehensive scan of the stomach with many local loop closures. Trajectory 3 contains an extensive scan of the stomach with more complicated local loop closures. Trajectory 4 consists of more challenge motions including fast rotational and translational frame-to-frame motions. Trajectory 5 is the same of trajectory 4, but included synthetic noise to evaluate the robustness of system against noise effects. Before capturing trajectory 6, we added more synthetic stomach oil into the simulator tissue to have heavier reflection conditions. Similar to the trajectory 5, trajectory 6 consists of very loopy and complex motions. As seen in Table 5, the system performs very robust and accurate in terms of trajectory tracking in all of the challenge datasets. Tracking accuracy is only decreased for very fast frame-to-frame movements, motion blur, noise, or heavy spectral reflections occurring frequently in last trajectories especially.

RMSE results for pose estimation before and after application of reflection suppression, de-vignetting, and radial undistortion were evaluated and compared to quantitatively

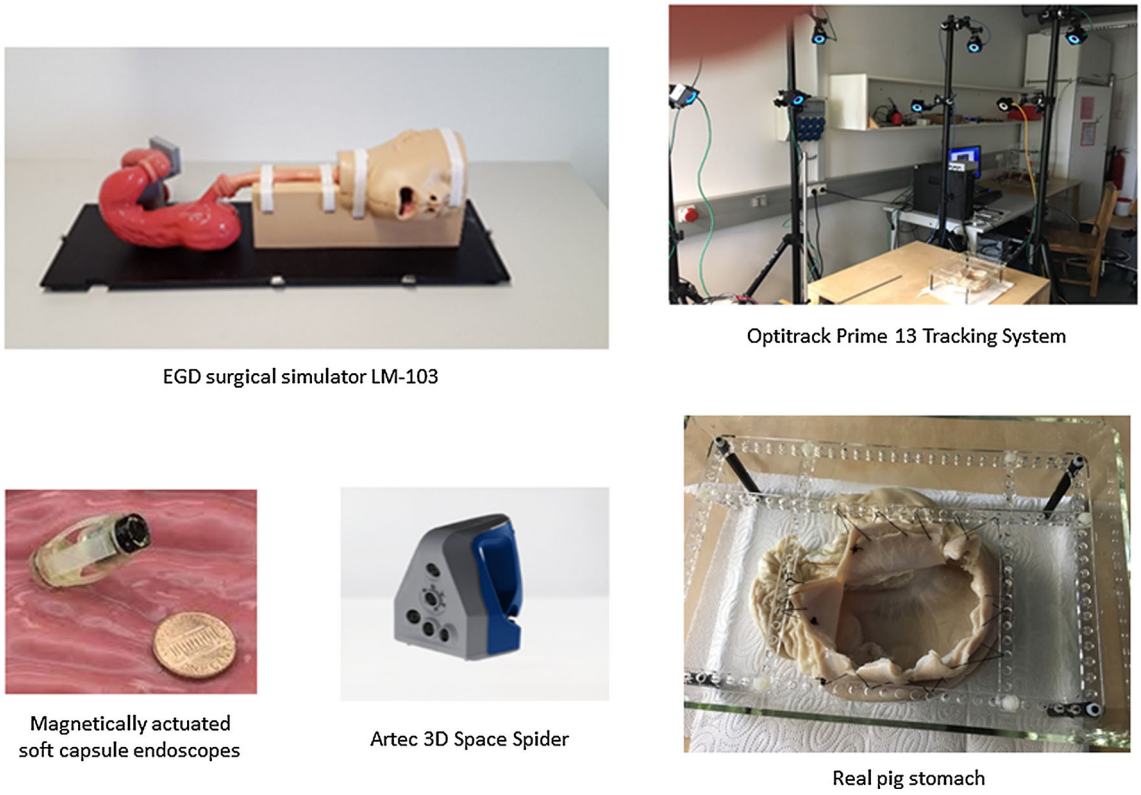


Fig. 9 Schematics of the experimental setup for 3D visual map reconstruction: a real pig stomach, an esophagus gastroduodenoscopy simulator for surgical training, 3D image scanner, Optitrack system, endoscopic camera, and active robotic capsule endoscope

Table 5 Comparison of ATE RMSE for different trajectories and cameras

	Length in cm	Potensic	Misumi-I	Misumi-II	Awaiba
Traj 1	123.5	4.10	4.23	4.17	6.93
Traj 2	132.4	4.14	4.45	4.32	7.12
Traj 3	124.6	5.23	5.54	5.43	7.42
Traj 4	128.2	5.53	5.67	5.47	7.51
Traj 5	128.2	6.32	5.45	5.32	8.32
Traj 6	123.1	7.73	6.72	6.51	8.73

analyze their effects in terms of pose estimation accuracy. Results shown in Table 6 for Misumi camera-II indicate that reflection suppression leads to a decrease in pose estimation performance. This decrease might be related to the fact that such saturated peak values contain orientation information. Thus, in consideration of pose estimation, reflection suppression should be avoided. On the other hand, radial undistortion and de-vignetting operations both increase pose estimation accuracy of the framework as expected.

4.3 Surface reconstruction

We evaluated the surface reconstruction accuracy of our system on the same dataset that we used for the trajectory estimation framework as well. We scanned the open non-rigid esophago-gastroduodenoscopy (EGD) simulator to obtain the ground truth 3D data using a highly accurate com-

Table 6 Comparison of ATE RMSE for MISUMI-II camera and different combinations of preprocessing operations

	RS	NRS	RS+RUD	RS+RUD+DV
Traj 1	5.45	4.12	4.01	4.03
Traj 2	6.44	4.23	4.07	4.04
Traj 3	6.57	5.13	4.97	4.98
Traj 4	7.55	5.34	5.16	5.08
Traj 5	8.43	5.43	5.14	5.02
Traj 6	8.69	5.64	5.25	5.12

NPR No preprocessing applied, *RS* reflection suppression applied, *RUD* radial undistortion applied, *DV* de-vignetting applied

Table 7 Comparison of surface reconstruction accuracy results on the evaluated datasets

	Depth	Potensic	Misumi-I	Misumi-II	Awaiba
Traj 1	63.42	2.82	2.32	2.14	3.42
Traj 2	63.45	2.56	2.45	2.16	4.14
Traj 3	63.41	3.16	2.76	2.45	4.45

Quantities shown are the mean distances from each point to the nearest surface in the ground truth 3D model in cm

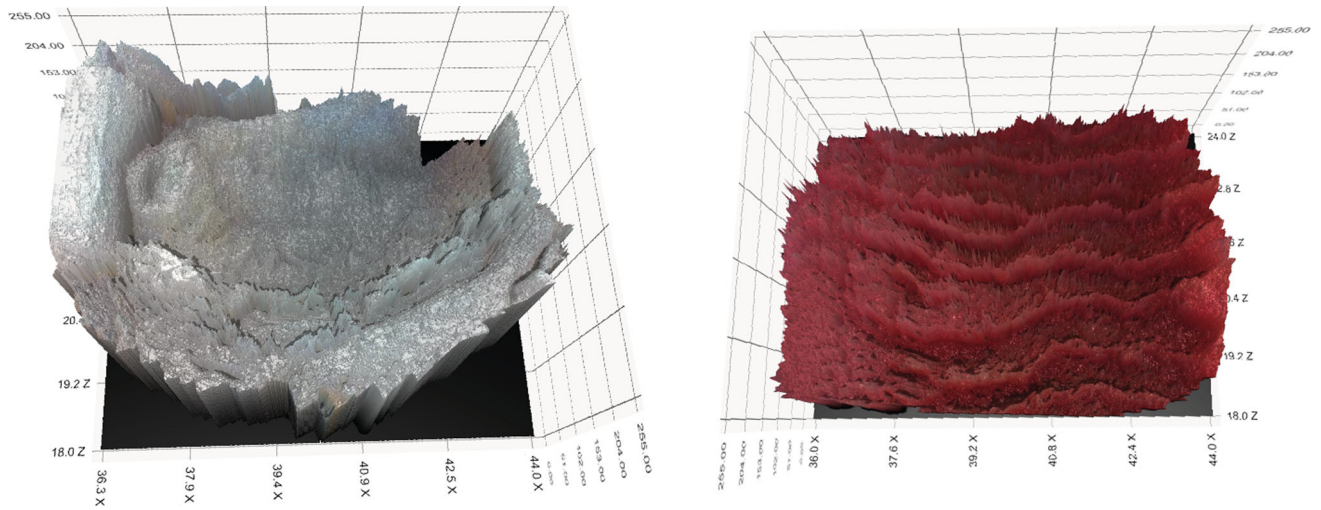


Fig. 10 Qualitative 3D reconstructed map results for different cameras [(real pig stomach (left), synthetic human stomach (right)]

Table 8 Comparison of ATE RMSE for different trajectories and combinations of preprocessing operations on the evaluated dataset

	NPR	RSM	RSPM	RSPM+RUD	RSPM+RUD+DV
Traj 1	5.45	3.65	3.42	2.02	2.14
Traj 2	6.44	3.91	3.71	2.08	2.16
Traj 3	6.54	4.23	3.94	2.27	2.45
Traj 4	7.25	4.53	4.14	3.02	3.14
Traj 5	8.35	4.95	4.63	3.34	3.52
Traj 6	8.95	5.55	5.14	3.55	3.82

Quantities shown are the mean distances from each point to the nearest surface in the ground truth 3D model in cm

NPR No preprocessing applied, *RSPM* reflection suppression applied for both pose estimation and map reconstruction, *RSM* reflection suppression applied only for map reconstruction, *RUD* radial undistortion applied, *DV* de-vignetting applied, MISUMI-II camera were used

mercial 3D scanner (Artec 3D Space Spider). The final 3D map of the stomach model obtained by the proposed framework and the ground truth scan were aligned using iterative closest point algorithm (ICP). The absolute depth (ADE) RMSE was used to evaluate the performance of map reconstruction approach, which measured the root-mean-square of Euclidean distances between estimated depth values and the corresponding ground truth depth values. A lowest RMSE of 2.14 cm (Table 7) proves that our system can achieve very high map accuracies. Even in more challenge trajectories such as trajectory 3, our system is still capable of providing an acceptable 3D map of the explored inner organ tissue. Three-dimensional reconstructed maps of real pig stomach and synthetic human stomach are represented in Fig. 10 for visual reference.

To evaluate the contributions of each preprocessing module on the map reconstruction accuracy, we tested the approach with leave-one out strategy leaving one module each time. As shown in Table 8, each preprocessing operation has a certain influence on the RMSE results. One important observation is that even though pose accuracy increases with existence of reflection points, these saturated pixels have negative influence on the map accuracy, as expected. Therefore, disabling reflection suppression during pose estimation and enabling it for map reconstruction are the best option to follow.

4.4 Computational performance

To analyze the computational performance of the proposed framework, we determined the average frame pair processing time across the trajectory sequences. The test platform was a desktop PC with an Intel Xeon E5-1660v3-CPU at 3.00, 8 cores, 32GB of RAM, and an NVIDIA Quadro K1200 GPU with 4GB of memory. Three-dimensional reconstruction of 100 frames took 80.54 s to process, whereas processing of 200 frames took 180.83 s, and processing of 300 frames 290.12 s, respectively. That indicates an average frame pair processing time of 919.15 ms, implying that our pipeline needs to be accelerated using more effective parallel computing and GPU power in order to reach real-time performance. To achieve this, we developed a RGB-Depth SLAM

method, which is capable of capturing comprehensive and globally dense surfel-based maps of the inner organs in real time, by using joint photometric–volumetric pose alignment, dense frame-to-model camera tracking, and frequent model refinement through non-rigid surface deformations [1]. The execution time of the RGB-Depth SLAM is dependent on the number of surfels in the map, with an overall average of 48 ms per frame scaling to a peak average of 53 ms, implying a worst case processing frequency of 18 Hz. Even though RGB-Depth SLAM is much faster than our sparse-then-dense alignment-based 3D reconstruction method, the map quality decreases due to the use of surfel elements. Moreover, the joint photometric–volumetric pose alignment is prone to converge into local minima in low-textured areas. For further details of our RGB Depth SLAM method, the reader is referred to our paper [1].

4.5 Conclusion

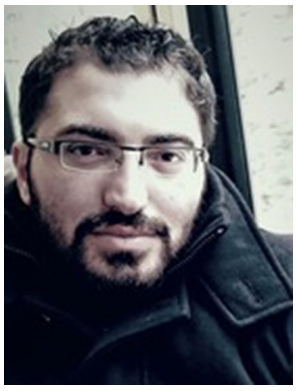
In this study, we proposed a therapeutically relevant and very detailed 3D map reconstruction approach for endoscopic capsule robots consisting of preprocessing, key frame selection, a sparse-then-dense pose estimation, frame stitching, and shading-based 3D reconstruction. Detailed quantitative and qualitative evaluations show that the proposed system achieves sub-millimeter precision for both 3D map reconstruction and pose estimation. In future, we aim to achieve real-time operation for the proposed framework so that it can be used for active navigation of the robot during endoscopic operations, as well. Moreover, we plan to incorporate magnetic localization and scale estimation module into our method to develop even more robust endoscopic reconstruction tools.

Acknowledgements Open access funding provided by Max Planck Society.

References

1. Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., Sitti, M.: A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots. Int. J. Intell. Robot. Appl. (2017a)
2. Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., Sitti, M.: Deep EndoVO: a recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. Neurocomputing (2017b)
3. Sitti, M., Ceylan, H., Hu, W., Giltinan, J., Turan, M., Yim, S., Diller, E.: Biomedical applications of untethered mobile milli/microrobots. Proceedings of the IEEE **103**(2), 205–224 (2015)
4. Turan, M., Almalioglu, Y., Gilbert, H. , Sari, A.E. Soylu, U., Sitti, M.: Endo-VMFuseNet: deep visual-magnetic sensor fusion approach for uncalibrated, unsynchronized and asymmetric endoscopic capsule robot localization data. arXiv:1709.06041 [cs.RO] (2017c)
5. Turan, M., Shabbir, J., Araujo, H., Konukoglu, E., Sitti, M.: A deep learning based fusion of RGB camera information and magnetic localization information for endoscopic capsule robots. J. Intell. Robot. Appl, Int (2017). https://doi.org/10.1007/s41315-017-0039-1
6. Devernay, F., Mourgues, F., Coste-Manire, É.: Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In: International Workshop on Medical Imaging and Augmented Reality (MIAR), pp. 16–20 (2001)
7. Hager, G., Vagvolgyi, B., Yuh, D.: Stereoscopic video overlay with deformable registration. In: Medicine Meets Virtual Reality (MMVR) (2007)
8. Su, L.M., Vagvolgyi, B.P., Agarwal, R., Reiley, C.E., Taylor, R.H., Hager, G.D.: Augmented reality during robot-assisted laparoscopic partial nephrectomy: toward real-time 3D-CT to stereoscopic video registration. Urology **73**, 896–900 (2009)
9. Stoyanov, D., Scarzanella, M., Pratt, P., Yang, G.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention MICCAI, pp. 275–282 (2010)
10. Stoyanov, D., Mylonas, G., Deligianni, F., Darzi, A., Yang, G.: Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 3759, pp. 114–121 (2005)
11. Wu, C., Narasimhan, S.G., Jaramaz, B.: A multi-image shape-from-shading framework for near-lighting perspective endoscopes. Int. J. Comput. Vis. **86**, 211–228 (2010)
12. Yeung, S., Tsui, H., Yim, A.: Global shape from shading for an endoscope image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 318–327 (1999)
13. Okatani, T., Deguchi, K.: Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. Comput. Vis. Image Underst. **66**, 119–131 (1997)
14. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. **47**, 7–42 (2002)
15. Fan, Y., Meng, MQ-H., Li, B.: 3D reconstruction of wireless capsule endoscopy images. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (2010)
16. Horn, B.: Shape from shading. Cambridge: Massachusetts Institute of Technology. Int. J. Comput. Vis. **5**(1), 37–75 (1970)
17. Rai, L., Higgins, W.E.: Method for radiometric calibration of an endoscopes camera and light source. In: SPIE Medical Imaging: Visualization, Image-Guided Procedures, and Modeling, pp. 691–813 (2008)
18. Visentini-Scarzanella, M., Stoyanov, D., Yang, G.-Z.: Metric depth recovery from monocular images using shape-from-shading and specularities. IEEE International Conference on Image Processing (ICIP), Orlando, FL (2012)
19. Wang, R, et al.: Improving 3D surface reconstruction from endoscopic video via fusion and refined reflectance modeling. (2017)
20. Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: The Endoscopogram: A 3D model reconstructed from endoscopic video frames. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016. MICCAI

2016. Lecture Notes in Computer Science, vol. 9900. Springer, Cham (2016)
21. Kaufman, A., Wang, J.: 3d Surface Reconstruction from Endoscopic Videos, Visualization in Medicine and Life Sciences, pp. 61–74. Springer, Berlin (2008)
22. Malti, A., Bartoli, A.: Combining conformal deformation and cooktorrance shading for 3-D reconstruction in laparoscopy. IEEE Trans. Biomed. Eng. **61**(6), 1684–1692 (2014)
23. Malti, A., Bartoli, A., Collins, T.: Template-based conformal shape-from-motion-and-shading for laparoscopy. In: International Conference on Information Processing in Computer-Assisted Interventions. Springer, Berlin (2012)
24. Nadeem, S., Kaufman, A.: Depth reconstruction and computer-aided polyp detection in optical colonoscopy video frames. arXiv preprint arXiv:1609.01329 (2016)
25. Abu-Kheil Y, Ciuti G, Mura M, Dias J, Dario P, Seneviratne L: Vision and inertial-based image mapping for capsule endoscopy. In: 2015 International Conference on Information and Communication Technology Research (ICTRC) (2015)
26. Telea, Alexandru: An image inpainting technique based on the fast marching method. J. Graph GPU Game Tools **9**, 23–34 (2004)
27. Conrady, A.: Decentering lens systems. Mon. Not. R. Astron. Soc. **79**, 384–390 (1919)
28. Zheng, Y., Yu, J., Kang, S.B., Lin, S., Kambhamettu, C.: Single-image vignetting correction using radial gradient symmetry. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 562–576 (2008)
29. Burt, P.J., Adelson, E.H.: A multi-resolution spline with application to image mosaics. ACM Trans. Graph. (TOG). https://dl.acm.org (1983)
30. Wu, C., Agarwal, S., Curless, B., Seitz, S.M. Multicore bundle adjustment. In: CVPR (2011)
31. Brown, M., Loewe, D.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vision **74**(1), 59–73 (2007)
32. Ping-Sing, T., Shah, M.: Shape from shading using linear approximation. Image Vis. Comput. **12**(8), 487–498 (1994)
33. Elhabian, S.Y.: Hands on shape from shading. SCI Home Technical Report, Spring (2008)
34. Yim, S., Sitti, M.: Design and rolling locomotion of a magnetically actuated soft capsule endoscope. IEEE Trans. Robot. **28**, 183–194 (2012)
35. Yim, S., Goyal, K., Sitti, M.: Magnetically actuated soft capsule with multi-modal drug release function. IEEE/ASME Trans. Mechatron. **18**, 1413–1418 (2013)



Mehmet Turan received his Diploma Degree from the Information technology and Electronics engineering department of RWTH Aachen, Germany in 2012. He was a research scientist at UCLA (University of California Los Angeles) between 2013 and 2014 and a research scientist at the Max Planck Institute for Intelligent Systems between 2014-present. He is currently enrolled as a Ph.D. Student at the ETH Zurich, Switzerland. He is also affiliated with Max Planck-ETH Center for Learning Systems, the first joint research center of ETH Zurich and the Max Planck Society. His research interests include SLAM (simultaneous localization and mapping) techniques for milli-scale medical robots and deep learning techniques for medical robot localization and mapping. He received DAAD fellowship between years 2005–2011 and Max Planck Fellowship between 2014-present. He has also received MPI-ETH Center fellowship between 2016-present.

Yusuf Yigit Pilavci received his Bachelor Degree from Electrical and Electronics Engineering of Middle East Technical University, Ankara in 2017. He worked as an undergraduate researcher focusing on image processing, computer vision, machine learning and artificial intelligence. Currently, he pursues his master degree in Computer Science and Engineering of Politecnico di Milano, Italy. Additionally, he is working on graph signal processing and domain adaptation problems.

Ipek Ganiyusufoglu pursues B.Sc. degree in Department of Computer Science, Sabanci University, Turkey. Besides smaller projects, her current interests include computer graphics, interaction and vision, which she plans to focus on further when doing masters.

Helder Araujo is a Professor at the Department of Electrical and Computer Engineering of the University of Coimbra. His research interests include Computer Vision applied to Robotics, robot navigation and visual servoing. In the last few years he has been working on non-central camera models, including aspects related to pose estimation, and their applications. He has also developed work in Active Vision, and on control of Active Vision systems. Recently he has started work on the development of vision systems applied to medical endoscopy.

Ender Konukoglu Ph.D., finished his Ph.D. at INRIA Sophia Antipolis in 2009. From 2009 till 2012 he was a post-doctoral researcher at Microsoft Research Cambridge. From 2012 till 2016 he was a junior faculty at the Athinoula A. Martinos Center affiliated to Massachusetts General Hospital and Harvard Medical School. Since 2016 he is an Assistant Professor of Biomedical Image Computing at ETH Zurich. His is interested in developing computational tools and mathematical methods for analysing medical images with the aim to build decision support systems. He develops algorithms that can automatically extract quantitative image-based measurements, statistical methods that can perform population comparisons and biophysical models that can describe physiology and pathology.

Dr. Metin Sitti received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1992 and 1994, respectively, and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. He was a research scientist at UC Berkeley during 1999–2002. He has been a professor in the Department of Mechanical Engineering and Robotics Institute at Carnegie Mellon University, Pittsburgh, USA since 2002. He is currently a director at the Max Planck Institute for Intelligent Systems in Stuttgart. His research interests include small-scale physical intelligence, mobile microrobotics, bio-inspired materials and miniature robots, soft robotics, and micro-/nanomanipulation. He is an IEEE Fellow. He received the SPIE Nanoengineering Pioneer Award in 2011 and NSF CAREER Award in 2005. He received many best paper, video and poster awards in major robotics and adhesion conferences. He is the editor-in-chief of the Journal of Micro-Bio Robotics.

INVITED PAPER

Biomedical Applications of Untethered Mobile Milli/Microrobots

This paper reviews the current advances in biomedical untethered mobile millirobots and microrobots.

By Metin Sitti, *Fellow IEEE*, Hakan Ceylan, Wenqi Hu, *Student Member IEEE*, Joshua Giltinan, *Student Member IEEE*, Mehmet Turan, Sehyuk Yim, *Student Member IEEE*, and Eric Diller, *Member IEEE*

ABSTRACT | Untethered robots miniaturized to the length scale of millimeter and below attract growing attention for the prospect of transforming many aspects of health care and bioengineering. As the robot size goes down to the order of a single cell, previously inaccessible body sites would become available for high-resolution *in situ* and *in vivo* manipulations. This unprecedented direct access would enable an extensive range of minimally invasive medical operations. Here, we provide a comprehensive review of the current advances in biomedical untethered mobile milli/microrobots. We put a special emphasis on the potential impacts of biomedical microrobots in the near future. Finally, we discuss the existing challenges and emerging concepts associated with designing such a miniaturized robot for operation inside a biological environment for biomedical applications.

KEYWORDS | Biomedical engineering; medical robots; microrobots; minimally invasive surgery

I. INTRODUCTION

One of the highest potential scientific and societal impacts of small-scale (millimeter and submillimeter size) untethered mobile robots would be their healthcare and bioengineering applications. As an alternative to existing tethered medical devices such as flexible endoscopes and catheters, mobile medical milli/microrobots could access complex and small regions of the human body such as gastrointestinal (GI), brain, spinal cord, blood capillaries, and inside the eye while being minimally invasive and could even enable access to unprecedented submillimeter size regions inside the human body, which have not been possible to access currently with any medical device technology [1], [2].

As an alternative to tethered flexible endoscopes used in the GI tract, untethered pill-size, swallowable capsule endoscopes with an on-board camera and wireless image transmission device have been commercialized and used in hospitals (FDA approved) since 2001, which has enabled access to regions of the GI tract that were impossible to access before, and has reduced the discomfort and sedation related work loss issues [3]–[7]. However, capsule endoscopy is limited to passive monitoring of the GI tract via optical imaging as clinicians have no control over the capsule's position, orientation, and functions. Several groups have been proposing active, robotic capsule endoscopes within the last decade where such devices could be remotely controlled to achieve active imaging and have other medical functions [8]–[13]. In bioengineering, mobile microrobots, due to their ability to manipulate individual biological microentities with high precision repeatedly, could be used as a new scientific study or prototyping tool for tissue engineering (e.g., assembling and controlling the building blocks of regenerated tissues) and cellular biology

Manuscript received October 20, 2014; revised November 26, 2014; accepted December 17, 2014. Date of current version March 23, 2015. This work was supported by the NIH R01-NR014083 grant, the NSF Cyber Physical Systems Program (CNS-1135850), and the NSF National Robotics Initiative Program (NRI-1317477).
M. Sitti and **J. Giltinan** are with Max-Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany, and also are with Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15238 USA (e-mail: sitti@is.mpg.de; giltinan@is.mpg.de).
H. Ceylan, **W. Hu**, and **M. Turan** are with Max-Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany (e-mail: ceylan@is.mpg.de; wenqi@is.mpg.de).
S. Yim is with Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: sam.sehyuk@gmail.com).
E. Diller is with Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S3G8, Canada (e-mail: ediller@mie.utoronto.ca).
Digital Object Identifier: 10.1109/JPROC.2014.2385105

such as single cell studies by manipulating single non-motile or motile cells.

Reported small-scale biomedical robot sizes use range from tens of micrometers to several centimeters. We can classify such different length scale miniature robots as millirobots and microrobots. We define a mobile microrobot as a mobile robotic system where its untethered mobile component has all dimensions less than 1 mm and larger than 1 μm and its mechanics is dominated by microscale physical forces and effects. Thus, for microrobots, bulk forces such as inertial forces and buoyancy are negligible or comparable to surface area and perimeter related forces such as surface tension, adhesion, viscous forces, friction, and drag. In millirobots, their untethered mobile components have all dimensions less than palm size and larger than 1 mm and macroscale forces such as bulk forces dominate their mechanics. On-board components for milli/microrobots must have overall sizes much smaller than the given robot overall size. Therefore, all on-board robot components such as mechanisms, tools, actuators, sensors, power source, electronics, computation, and wireless communication must be miniaturized down to micron scale. Moreover, for milli/microrobots, such components need to be fabricated by micro/nanofabrication methods, which are different from conventional macroscale machining techniques.

There are two main approaches of designing, building, and controlling mobile medical small-scale robots:

- *On-board approach*: Similar to a typical macroscale mobile robot, the untethered, self-contained and self-propelled miniature robot has all on-board components to operate autonomously or with a remote control.
- *Off-board approach*: The mobile, untethered component of the milli/microrobotic system is externally (off-board) actuated, sensed, controlled, or powered.

Since various commercial on-board components exist for millirobots, on-board approach is possible for millirobots while such components are not readily available for microrobots. Thus, most of the current mobile microrobotics studies in literature have been using the off-board approach, and therefore our microrobotics definition also covers such studies.

In addition to the on-board and off-board approaches, milli/microrobots can be also classified as synthetic and biohybrid. In the former case, the milli/microrobot is made of fully synthetic materials such as polymers, magnetic materials, silicon, composites, elastomers, and metals, while the latter is made of both biological and synthetic materials. biohybrid milli/microrobots are typically integrated with muscle cells such as cardiomyocytes or microorganisms such as bacteria, algae, spermatozoids, and protozoa, and powered by the chemical energy inside the cell or in the environment [14]. They harvest the efficient and robust propulsion, sensing, and control capabilities of biological cells or tissues. Such cells could propel the robot in a given physiologically compatible environment, and sense environmental stimuli to control the robot motion by diverse mechanisms such as chemotaxis, magnetotaxis, galvanotaxis, phototaxis, thermotaxis, and aerotaxis.

Advances in and increased use of microelectromechanical systems (MEMS) since the 1990s have driven the development of untethered milli/microrobots. MEMS fabrication methods allow for precise features to be made from a wide range of materials, which can be useful for functionalized microrobots. There has been a surge in microrobotics work in the past few years, and the field is relatively new and is growing fast [1], [15]. Fig. 1 presents an overview of a few of the new microrobotic technologies, which have been published, along with their approximate overall size scale.

The first miniature machines were conceived by Feynman in his lecture on "There's Plenty of Room at the Bottom" in 1959. In popular culture, the field of milli/microrobotics is familiar to many due to the 1966 sci-fi movie Fantastic Voyage, and later the 1987 movie Innerspace. In these films, miniaturized submarine crews are injected inside the human body and perform noninvasive surgery. The first studies in untethered robots using principles which would develop into milli/microrobot actuation principles were only made recently, such as a magnetic stereotaxis system [16] to guide a tiny permanent magnet inside the human body and a magnetically driven screw which moved through tissue [20]. At the millimeter and centimeter size scale, advances in such millirobots have brought crawling, flying, and swimming devices with increased interest over the last decade. While many developments in millirobots are not directly relevant to biomedical applications, the technologies developed can be used in biomedical millirobots. One major milestone was the creation of centimeter-scale crawling robot with on-board power and computation in 1999 [51]. Micromechanical flying insect robots were first introduced in 2000 [19]. A solar powered crawling robot was introduced in 2004 [21]. Centimeter-scale compliant running robots with on-board power, actuation, and control were advanced with compliant mechanisms in 2008 [26]. Free flight (but with off-board power delivered via wires) mechanical insect-inspired robot was demonstrated in 2013 [46]. The first capsule endoscopes for medical use were used clinically in 2001 under FDA approval. Additional milestones for capsule endoscopy has been the introduction of a crawling mechanism [52] and the introduction of on-board drug delivery mechanism [53].

At the submillimeter scale, other significant milestone studies in untethered microrobotics include a study on bacteria-inspired swimming propulsion [54], bacteria-propelled beads [23], [55], steerable electrostatic crawling microrobots [30], catalytic self-propelled microtubular swimmers [24], laser-powered microwalkers [31], magnetic resonance imaging (MRI) device-driven magnetic beads [29], and magnetically driven millimeter-scale nickel robots [56]. These first studies have been followed by other

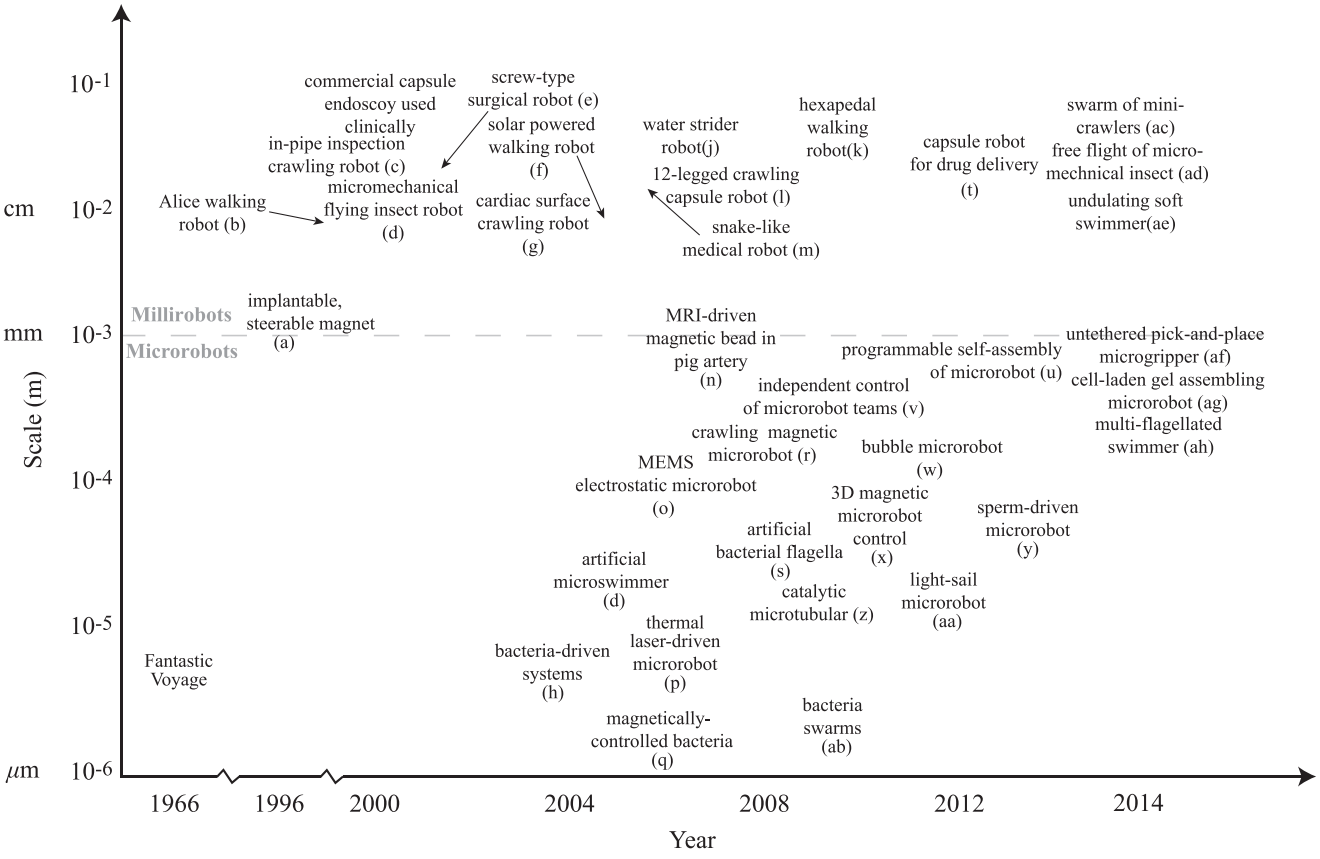


Fig. 1. ***Approximate timeline showing the emerging new milli/microrobot systems with their given overall size scale as significant milestones. (a) Implantable tiny permanent magnet steered by external electromagnetic coils [16]. (b) Alice 1 cm³ walking robot [17]. (c) In-pipe inspection crawling robot [18]. (d) Micromechanical flying insect robot [19]. (e) Screw-type surgical millirobot [20]. (f) Solar powered walking robot [21]. (g) Cardiac surface crawling medical robot [22]. (h) Bacteria-driven biohybrid microrobots [23]. (i) biohybrid magnetic microswimmer [24]. (j) Water strider robot [25]. (k) Hexapedal compliant walking robot [26]. (l) 12-legged crawling capsule robot [27]. (m) Snake-like medical robot [28]. (n) Magnetic bead driven by a Magnetic Resonance Imaging device in pig artery [29]. (o) MEMS electrostatic microrobot [30]. (p) Thermal laser-driven microrobot [31]. (q) Magnetically controlled bacteria [32]. (r) Crawling magnetic microrobot [33]. (s) Magnetic microswimmer inspired by bacterial flagella [34], [35]. (t) Flexible capsule endoscope with drug delivery mechanism [36]. (u) Programmable self-assembly of microrobots [37]. (v) Independent control of microrobot teams [38]. (w) Bubble microrobot [39]. (x) 3D magnetic microrobot control [40]. (y) Sperm-driven biohybrid microrobot [41]. (z) Catalytic microtubular [42]. (aa) Light-sail microrobot [43]. (ab) Bacteria swarms as microrobotic manipulation systems [44]. (ac) Swarm of mini-crawlers [45]. (ad) Free flight of micromechanical insect [46]. (ae) Undulating soft swimmer [47]. (af) Untethered pick-and-place microgripper [48]. (ag) Cell-laden gel assembling microrobot [49]. (ah) Multiflagellated swimmer [50].***

novel actuation methods such as helical propulsion [34], [57], stick-slip crawling microrobots [33], magnetotactic bacteria swarms as microrobots [58], optically driven bubble microrobots [39], and microrobots driven directly by the transfer of momentum from a directed laser spot [43], among others. Figs. 2 and 3 shows a number of the existing approaches to microrobot mobility in the literature for motion in two-dimensions (2D) and three-dimensions (3D). Most of these methods belong to the off-board (remote) microrobot actuation and control approach, and will be discussed in detail later. It is immediately clear that actual microrobots do not resemble the devices shrunk down in popular microrobotics depictions.

In this review paper, first, existing and potential biomedical applications of mobile millirobots and microrobots are described including a brief case study in each application category, if available. Next, challenges and emerging concepts in miniaturized biomedical robots are presented. Finally, Section IV provides the conclusions and future directions. The material covered in the paper is outlined in schematic form in Fig. 4.

II. CURRENT AND POTENTIAL BIOMEDICAL APPLICATIONS OF MILLI/MICROROBOTS

A. Active Visual Imaging for Disease Diagnosis

Active visual (optical) imaging such as endoscopic and laparoscopic techniques is one of the most significant

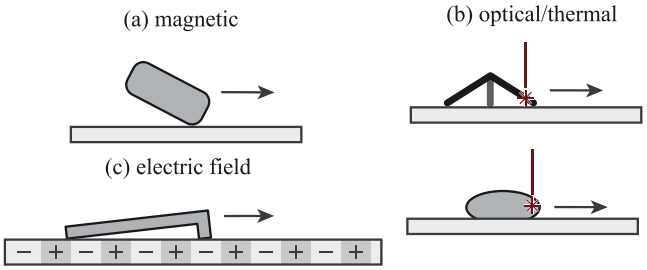


Fig. 2. *Some existing off-board approaches to mobile microrobot actuation and control in 2D. (a) Magnetically driven crawling robots include the Mag-μBot [33], the Mag-Mite magnetic crawling microrobot [59], the magnetic microtransporter [60], rolling magnetic microrobot [61], the diamagnetically-levitating mm-scale robot [62], the self-assembled surface swimmer [63], and the magnetic thin-film microrobot [64]. (b) Thermally driven microrobots include the laser-activated crawling microrobot [31], microlight sailboat [43], and the optically controlled bubble microrobot [39]. (c) Electrically driven microrobots include the electrostatic scratch-drive microrobot [65] and the electrostatic microbiorobot [60]. Other microrobots which operate in 2D include the piezoelectric-magnetic microrobot MagPieR [66] and the electrowetting droplet microrobot [67].*

methods to diagnose diseases. While flexible endoscopes and catheters provide visual disease diagnosis currently, they can be invasive and are only for short duration screening purposes. For minimally invasive and implantable

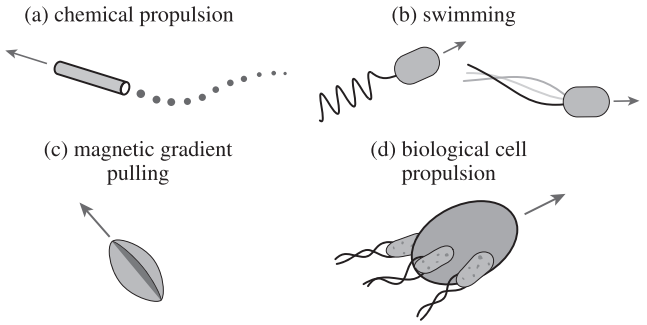


Fig. 3. *Some existing off-board and on-board approaches to mobile milli/microrobot actuation and control in 3D. (a) Chemically propelled designs include the microtubular jet microrobot [42] and the electro-osmotic swimmer [68]. (b) Swimming milli/microrobots include the colloidal magnetic swimmer [24], the magnetic thin-film helical swimmer [69], the micron-scale magnetic helix fabricated by glancing angle deposition [35], the microhelix microrobot with cargo carrying cage, fabricated by direct laser writing [70] and the microhelix microrobot with magnetic head, fabricated as thin-film and rolled using residual stress [34]. (c) Milli/microrobots pulled in 3D using magnetic field gradients include the nickel microrobot capable of five-degrees-of-freedom (DOF) motion in 3D using the OctoMag system [40] and the MRI-powered and imaged magnetic bead [71]. (d) Cell-actuated biohybrid approaches include the artificially-magnetotactic bacteria [72], the cardiomyocyte driven microswimmers [73], the chemotactic steering of bacteria-propelled microbeads [74], sperm-driven and magnetically steered microrobots [41], and the magnetotactic bacteria swarm manipulating microscale bricks [44].*

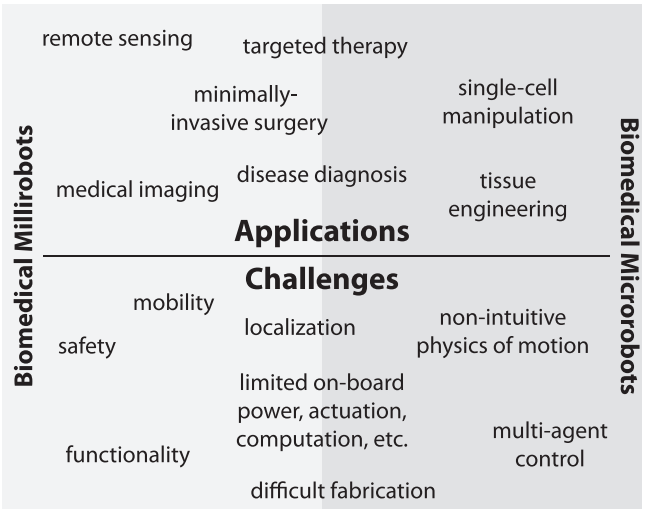


Fig. 4. *Applications and challenges for biomedical milli/microrobots.*

(long-duration) visual imaging and accessing small spaces that were not possible to reach before (e.g., small intestines), existing pill-size capsule endoscopes have been becoming a significant alternative [4], [7], [75], [76]. Such commercial pill-size capsule cameras have an on-board camera, a wireless transmission device, and a battery to just take images and send them to an external recording device. Turning such passive imaging devices into capsule millirobots would enable untethered active imaging of hard-to-access areas minimally invasively and for long durations. Therefore, many groups have been proposing robotic capsule millirobots for active imaging using different approaches. Using an on-board actuation approach, miniature motors based on leg or fin mechanisms were used to propel capsule robots inside the GI tract in a controlled manner. Through off-board actuation approach, many groups used remote magnetic actuation to stop, propel, or navigate capsule millirobots in the GI tract [77]. The former approach does not require bulky external devices for actuation while motors consume too much power compared to imaging, which reduces the imaging duration from hours to several minutes. However, external actuation or power transfer does not have such issue while they require bulky equipment around the patient, which would limit her/his motion capability and could be more expensive.

During active imaging, it is important to know the exact 3D location (and orientation) of the millirobot to enable more localized diagnosis and new advanced methods such as 3D visual mapping of the GI tract such as stomach by combining the 3D position information with the 2D camera images. For the localization of millirobots inside the GI tract, as the first approach, medical imaging devices such as fluoroscopy, which uses low-dose X-rays to image the capsule region at 1–2 frames per second [77] ultrasonic imaging [77]–[80], positron emission tomography (PET) [77], [81] and magnetic resonance imaging (MRI) [77]. As

an alternative approach, a radio transmitter has been placed on the commercial passive capsule endoscopes, and by placing multiple receiver antennas around the patient, an average position error of approximately 38 mm has been realized [77]. Moreover, by placing a small magnet inside the millirobot, hall-effect sensor arrays outside the patient have been used to localize the device [77], [82]. However, for magnetically actuated capsule robots, hall-effect sensor-based methods get more challenging due to the interference of the magnetic field from the actuating external magnet or electromagnetic coils and the magnet on the capsule robot on the sensor. Several studies addressed this problem and could still enable 3D localization using hall-effect sensors on the capsule [83] or outside the patient's body [84]. Also, magnetically actuated soft capsule robots with a hall-effect sensor could be localized using the shape change information of the capsule due to the external magnet position [85].

As an example, capsule millirobot, magnetically actuated soft capsule endoscope (MASCE) with an integrated CMOS camera (see Fig. 5 and Table 1) was proposed to actively image stomach type of 3D surfaces using remote magnetic control [53]. Soft design of the capsule body enabled safe operation (i.e., no damage to the tissue due to high stresses), extra degree-of-freedom actuation, and shape changing capability. After swallowing the MASCE and reaching to stomach in several seconds, an external magnet was used to roll it inside stomach for navigation and position control via the two tiny permanent magnets embedded inside it. Several localization methods [84], [85] were proposed to know the 3D position and 2D orientation of the robot precisely during imaging. Inside a surgical phantom stomach model, the feasibility of active imaging using such millirobot was demonstrated *in vitro*.

Since the currently available smallest CMOS camera with its lens from Awaiba GmbH with reasonable resolution (62,500-pixels) is 1 mm × 1 mm × 1 mm current active imaging functions are only [86] available for milliscale medical robots. Future lower resolution smaller cameras with integrated lighting and lens could enable mobile microrobots to actively image new smaller spaces inside the human body such as bile duct, spinal cord fluid, and brain lobes.

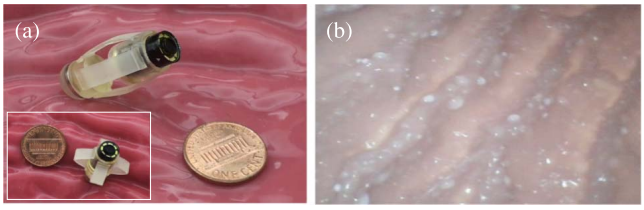


Fig. 5. *(a) Photograph of the prototype (left picture) of an example magnetically actuated capsule millirobot for active imaging inside stomach. A CMOS camera and LED lighting were integrated to the soft capsule robot, which can axially deform due to external magnetic actuation control. (b) An active imaging example (right picture) snapshot of the surgical stomach model from the CMOS camera during its active orientation control by an external magnet.*

Table 1 Specifications of the Example Magnetically Actuated Soft Capsule Millirobot Shown in Fig. 5

Diameter	10 mm
Length (min/max)	24 mm / 30 mm
Weight	6.1 g
Internal magnets	8 mm diameter x 1 mm long cylindrical NdFeB
Body material	Polyurethane elastomer

B. Mobile *In Situ* Sensing for Disease Diagnosis and Health Monitoring

Current passive biomedical sensors can be implanted inside or located outside the human body for continuous monitoring of a patient's or healthy person's health condition. Such sensors could measure or detect glucose, pH, temperature, oxygen, viral or bacterial activity, body motion (inertia), balance, blood pressure, respiration, muscle activity, neural activity, pulse rate, etc., *in situ* to diagnose and inform any abnormal medical condition or pathological activity. Adding remote or on-board mobility and control capability to such sensors by having them on medical milli/microrobots could enable a future mobile medical sensor network inside the human body for active health monitoring. Thus, various mobile sensors could be concurrently deployed with the purpose of patrolling inside the different body parts in a minimally invasive manner. Such important biomedical application of milli/microrobots (other than visual monitoring as given in Section II-A) has not been explored much yet. As a preliminary study, Ergeneman *et al.* [87] proposed a magnetically controlled untethered magnetic microrobot that could achieve optical oxygen sensing for intraocular measurements inside the eye.

C. Targeted Therapy

Targeted therapy is able to enrich the local concentration of therapeutics such as drugs, mRNA, genes, radioactive seeds, imaging contrast agents, stem cells, and proteins in a specific targeted region inside the body while maintaining minimal side effects in the rest of the body. Moreover, controlling the release kinetics can also modulate the concentration of the drug at the therapeutic window, and thereby prolonging the effect of single dose administration. Mobile milli/microrobots can release such therapeutic biological and chemical substances in a specific target location in precise and controlled amounts so that potential side effects are minimized and stronger amounts of the substances could be delivered for faster and better recovery.

As the main targeted therapy application, small-scale mobile robots have been used for targeted drug delivery in

the GI tract, blood vessels, etc. At the millimeter scale, active capsule endoscopes have been used to deliver drugs in the GI tract using passive or active drug release methods [8]. Typical drug delivery capsules use a remotely controlled triggering to move a mechanism that could eject the drug actively for one time in a controlled amount into the target location. Triggering of the drug release mechanism can be achieved by visible light, near-infrared light, ultrasound, or magnetic fields [88]. Also, the Joule electrical heating of a shape memory alloy wire could be used to trigger a drug mechanism [89]. A piston mechanism in a capsule robot was moved by a micromotor based actuation method [90] and by a remotely triggered ignition of the propellant based microthruster [91] for single-use ejection of drugs. An axial compression of a magnetically actuated soft capsule millirobot also enabled controlled ejection of liquid drugs for multiple times inside stomach [53], [92]. Moreover, the same soft capsule robot could change into a spherical like shape inside stomach so that it could stay there for a long time to deliver drugs by passive diffusion [36] as a semi-implantable drug delivery platform. After the drug delivery operation was over, the capsule was taken out by changing back its shape from a spherical shape to a cylindrical one, which enables its disposal naturally by peristalsis. As a specific example, Fig. 6 shows the active ejection of a liquid drug from the soft capsule robot using remote magnetic actuation control.

At the micron scale, there have been some preliminary studies to use untethered mobile microrobots to deliver drugs or other agents in the vascular system and eye [93]. In relatively larger human arteries with a dimension from 4 to 25 mm with a blood flow velocity from 100 to 400 mm/s, milliscale robots can be pulled or pushed around using magnetic field gradients [94]. Martel *et al.* proposed the magnetic resonance navigation to actuate a 1.5 mm diameter spherical magnet in swine carotid artery [29]. And the similar system was later used by Pouponneau *et al.* to deliver doxorubicin through rabbit hepatic artery [95]. In contrast to the system actuated by the spherical permanent magnet, this magnetic navigation system has larger switching rate enabling a closed-loop control [94].

To be able to access to the vessels smaller than arterioles ($< 150\ \mu$m), rotating magnetic microswimmers with a helical tail, inspired by flagella swimming of bacteria, were proposed for efficient swimming locomotion in low Reynolds number [35], [50], [57]. Such microswimmers can be coated with drugs and deliver them in a target location using passive diffusion [96] or potentially by an active release mechanism. Moreover, several studies proposed biohybrid microrobots where bacteria attached to a cargo such as drug particles or molecules transported the cargo to a desired location [14], [93] by remote control or bacterial sensing of the environment. Here, bacteria behave as on-board microactuators using the chemical energy inside the cell or in the environment and also as on-board microsensors detecting chemical, pH, oxygen, and temperature gradients in the environment [17]. A magnetotactic unipolar MC-1 bacterium could transport up to 70 sub-200 nm diameter liposomes, which encapsulate drugs, without a significant impact to the bacteria's swimming velocity using the remote magnetic steering control [97]. Also, Carlsen *et al.* [98] used many chemotactic bacteria to transport potential drug microparticles with embedded superparamagnetic nanoparticles while using remote magnetic fields to control the motion direction of the microparticles to reach to targeted regions before releasing the potential drug cargo. Swimming speed of such bacteria-propelled microparticles with 6 μm diameter was up to 7.3 μm/s under homogenous < 10 mT magnetic fields. Such biohybrid microrobots could be manufactured in large numbers cost effectively and fast, which could enable future targeted drug delivery applications using microrobot swarms (see Fig. 7).

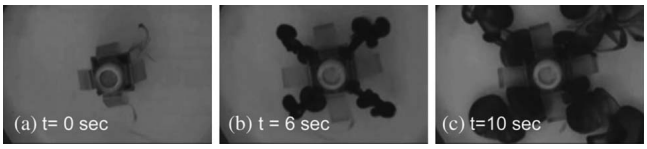


Fig. 6. *Active drug delivery demonstration of a soft capsule millirobot (see Table 1 for its specifications) inside stomach. (a)–(c) Time snapshots of the drug diffusion during the active compression of the drug chamber with the remote magnetic actuation [92].*

D. Minimally Invasive Surgery

In addition to diagnostic and therapeutic applications of milli/microrobots, next level of their medical use could be minimally invasive surgery inside the body. Such surgical operations or functions could be opening clogged vessels or other channels, cauterization, hyperthermia, biopsy, occlusion, electrical stimulation, injection, cutting, drilling, biomaterial removal, or addition at a given target, etc. Only several of these potential applications have been studied before. Many groups proposed integrated biopsy tools for capsule millirobots to collect tissue samples for further disease diagnosis. Kong *et al.* designed a rotational biopsy device designed to scratch the epithelial tissue [99]. Park *et al.* proposed a spring-driven biopsy microdevice with microspikes [100], [101]. Simi *et al.* created a biopsy capsule with a rotational razor that can be activated by a magnetic torsion spring mechanism [102]. These preliminary biopsy capsules have common drawbacks of inaccurate targeting of a certain area and inability of conducting biopsy for multiple times. On the other hand, in their soft capsule millirobot, Yim *et al.* [12] could release hundreds of untethered microgrippers that could grab tissue stochastically by self-folding due to the increased body temperature, and retrieve the microgrippers with their grabbed tissues inside stomach *ex vivo* for further genetic analysis. Next, inside the eye, Ullrich *et al.* tried to puncture a blood

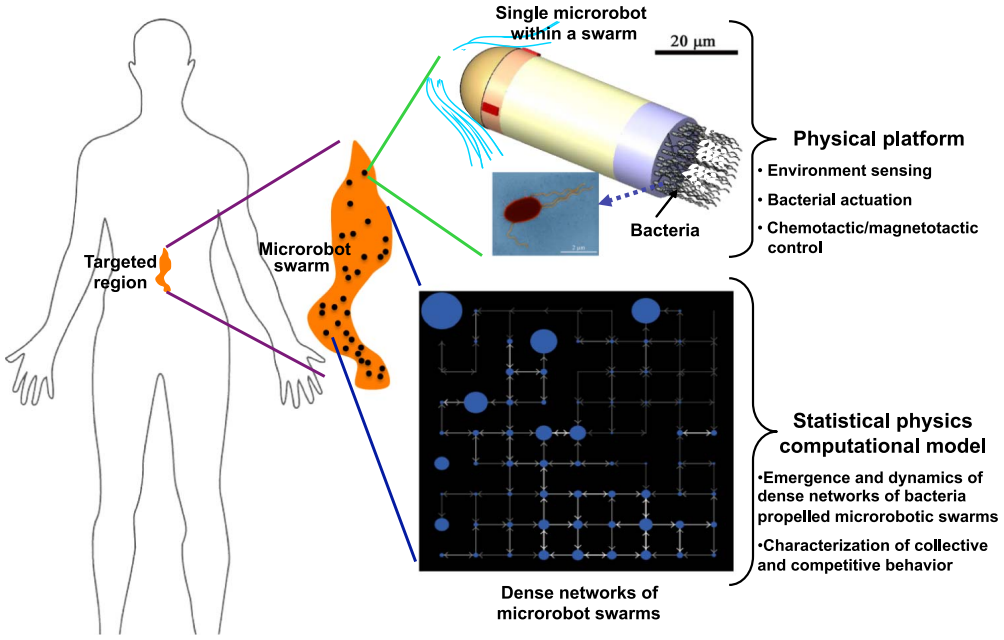


Fig. 7. ***Conceptual sketch of a bacteria-propelled biohybrid microrobot swarm, as a dense stochastic network, transporting and delivering drugs on targeted regions inside the stagnant fluid regions of the human body.***

vessel close to the retina using the rotational motion of a magnetic millirobot with a sharp tip [103]. Yu *et al.* [104] and Miloro *et al.* [105] proposed magnetic millirobots that could be spun remotely by remote rotating magnetic fields to potentially open clogs in blood vessels. In overall, there are only few preliminary minimally invasive surgery studies, which could be extended significantly with many new potential applications inside the circulatory system, brain, spinal cord, and other organs.

E. Tissue Engineering

Many diseases could be treated by precisely delivering the differentiated stem cells and regenerating tissues at the pathological sites [106]. Preliminary research has been done by Kim *et al.* who designed a cage shape microrobot which is fabricated by stereolithography of negative tone photoresist [107]. Coating the developed polymer structures with Ni/Ti bilayer rendered the microrobot steerable by the magnetic field. By coating the microrobot further with poly-L-lysine, the author could culture human embryonic kidney cell (HEK293) in 3D inside the microrobot, showing the possibility of using it as bioscaffold to support tissue regeneration [2]. Alternatively, artificial tissues can also be constructed *in vitro* first and then replace its malfunction *in vivo* counterparts, and thereby provide a new source for medical transplantation [108]–[110]. One way to achieve artificial tissues is by arranging microscale hydrogels (microgel) laden with different cells into predefined geometries [111]–[113]. For example, Tasoglu *et al.* [114] functionalized microgel with radical solution in a high magnetic gradient to make it paramagnetic. This enables microgels to be self-assembled into desired shapes under the influence of a uniform magnetic field. After the assembly, the magnetization of microgel could be disabled by vitamin E, so that the free radicals could be eliminated to ensure the proliferation of cells throughout the hydrogel scaffold [114].

As a more general way, the microrobot can also directly manipulate the non-functionalized microgels into desired geometry. For example, Tasoglu *et al.* [49] used a crawling magnetic microrobot ($750 \times 750 \times 225\ \mu m^3$) to push cell laden microgels made of either polyethylene glycol dimethacrylate (PEGDMA) or gelatin methacrylate (GelMA). As shown in Fig. 8(b), the assembly on the upper layer was aided by a microfabricated ramp to elevate the microrobot. In contrast to the conventional manipulation by optical tweezers [115] and dielectrophoresis force [116], this microrobotic approach distinguishes itself by minimally relying on the property of the microobjects. Thus, many different materials could be transported and integrated into tissue construct [49]. This is especially helpful in testing various combinations of different materials to figure out the optimal solution for constructing a specific tissue.

However, it has to be noticed that the microfabricated ramp used could limit the maximum layers of the assembly. To interface microrobot with the conventional tissue culturing dish with flat bottom, the microrobot has to pick up and drop the microgel on top of each other. Diller *et al.* addressed this by reshaping the magnetic microrobot into a gripper, as shown in Fig. 8(c) [48]. The microgripper jaw was remotely controlled by the magnetic field to clamp and

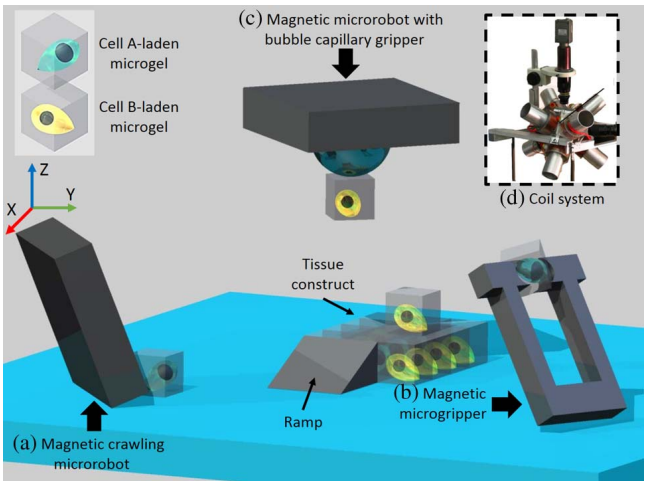


Fig. 8. *3D assembly of cell-laden microgels by different microrobots. (a) Magnetic crawling microrobot [49]. The microgel is pushed by the microrobot to the desired position. A microfabricated ramp is used to elevate the microrobot to higher layer of the tissue constructs. (b) Magnetic microgripper [48]. The jaw is opened and closed by external magnetic field to pick up and release the microobject. (c) Magnetic microrobot with bubble capillary gripper [117]. Changing the pressure inside the working environment can extend and retract the bubble to pick up and release the microobject. (d) Magnetic coil system for microrobot control.*

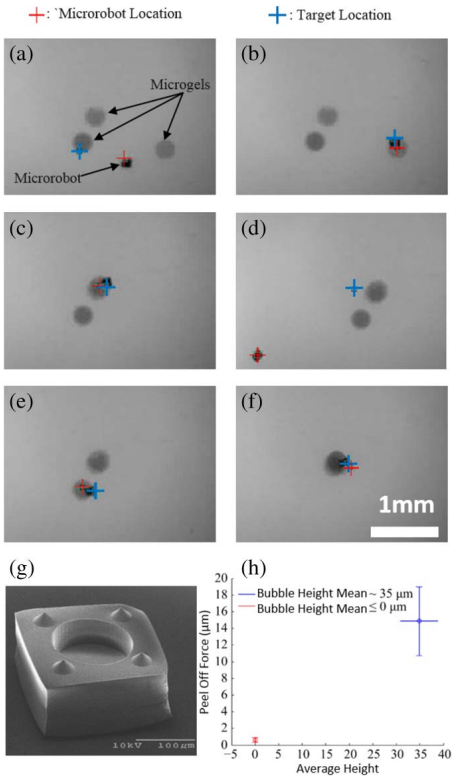


Fig. 9. *(a)–(f) Capillary gripping microrobot manipulating hydrogels into a stack as shown from the top-down view. (a) The microrobot position is given by the red cross and the desired position is given by the blue cross. microrobot position control is achieved by a PID controller used to determine the applied magnetic force. The hydrogels are the three circular disks (diameter ~ 350 μm) and the microrobot is a capillary gripping microrobot with a side dimension of 150 μm. (b) The microrobot is directed above the hydrogel and the bubble is drawn out of the cavity by a negative applied pressure in the microrobot workspace. The microrobot is then lowered onto the hydrogel. (c) The microrobot with the hydrogel positions itself over the center hydrogel and comes into contact. (d) The microrobot detaches from the stack of two hydrogels. (e), (f) The process is repeated for the left hydrogel, resulting in a three-hydrogel stack. Scale bar is 1 mm. (g) Example magnetic microrobot with a cavity for bubble-based capillary gripping. The four cones ensure surface contact is minimized when releasing parts. (h) Peel off force versus the average bubble height. The peel off force is calculated as the equivalent force acting on the center of the microrobot due to the applied magnetic torque. The magnetic torque is calculated from the applied uniform magnetic field and the known magnetization of the microrobot. The bubble height is measured from the cavity opening to the highest point of the bubble when it is not in contact with the test substrate. The height of 0 indicates the bubble is completely inside the cavity and there should be no capillary attachment force and is considered to be in the "release" state. Any positive non-zero bubble height will be considered the "pick" state. On a test silicon substrate, the best current work shows an attachment switching ratio of peel off force in the "pick" state to the peel off force in the "release" state of 25 : 1.*

release the microgel. In another work by Giltinan *et al.* [117], the microobject was picked up by the capillary force on a microbubble nested in the cavity of the magnetic microrobot, as shown in Fig. 8(d). Increasing the pressure inside the working environment could retract the bubble and release the microgel.

As a specific example, a top-down view of the microrobot manipulating microgel into a stack is shown in Fig. 9(a)–(f). Here, the force required to peel off a silicon substrate using magnetic torque was used as the metric of effectiveness for the capillary gripping magnetic microrobot. While many variables can affect the force required to peel the bubble from the test substrate, Fig. 9(h) shows the peel off forces when the bubble height, measured before the experiment, is less than 0, indicating the bubble is in the cavity, and when the bubble height is approximately 35 μm for a cavity radius of 75 μm. The minimum peel off force average of 0.6 μN and maximum peel off force of 14.9 μN indicate a switching ratio of approximately 25 : 1. The peel off force minimization was aided by surface contact minimizing features, shown on an example microrobot in Fig. 9(g).

F. Cell Manipulation

The biomedical analysis of single cells can differentiate genetic, metabolic and behavior heterogeneity, which pushes the microbiology research to an unprecedented resolution [118], [119]. The single cell manipulation is conventionally done by a micromanipulator, which is a microscale end effector connected to macroscale actuator. This design restricts its access to open channels such as a petri dish [120]. In contrast, untethered microrobots can

Table 2 Single Cell Manipulation Studies Conducted by Untethered Microrobots

Reference	Microrobot Agent	Target Cell	Application
Steager *et al.* [121]	Magnetic microstructure	Mouse embryo Trypsinized neuron	Cell transportation Microgel transportation
Hu *et al.* [122]	Bubble driven hydrogel disk	Mouse fibroblast	Cell transportation
Kwon *et al.* [123]	Magnetic microstructure attached with oscillating bubble	Fish egg	Cell transportation
Ye and Sitti [124]	Rotating magnetic bead	*S. marcescens* bacterium	Cell transportation
Schurle *et al.* [125]	Magnetic bead	Macrophage	Test engulfing
Hagiwara *et al.* [126], Feng *et al.* [127]	Magnetically driven microtool	Bovine oocyte	Cell transportation Cell cutting
Kawahara *et al.* [128]	Magnetically driven microtool	*P. Laevis*	Physical stimulation
Zhang *et al.* [129]	Ni nanowire	Epidermal cell Blood cell Microorganism with flagella	Cell transportation Cargo delivery
Petit *et al.* [130]	Self-assembled superparamagnetic microsphere doublet	*E. coli* bacterium	Cell transportation

manipulate cells in enclosed spaces such as microfluidic or other biological chips. Up till now, many different single cell manipulation tasks have been realized by untethered microrobots and are summarized in Table 2. Among these manipulations, microtransportation is the most common operation. Through microtransportation, either single cell can be isolated from its culture for later analysis [4] or drugs can be precisely delivered to a cell network to modulate the intracellular communication [121]. Moreover, random distributed cells can also be re-arranged into desired spatial geometry for the research such as observation of the cancer cell progression [122].

While manipulation of immotile cell is relative easy, manipulation of flagellated bacteria is much more challenging, which is conventionally done by optical tweezers with the cell threaten by photo damage [131], [132]. To address this, Ye and Sitti used the rotational flows around the rotating magnetic microparticle to selectively trap *S. marcescens* bacterium [124]. The authors showed that a uniform magnetic field smaller than 3.5 mT was enough to drive the microparticle and translate it with at a speed up to 100 μm/s.

Besides microtransportation, several other cell manipulations could be achieved by untethered microrobots. For example, a microrobot with force sensor was designed by Kawahara *et al.* to mechanically stimulate and investigate the *P. Laevis* response [128]. In the future, such sensor could be used to distinguish abnormal cell by its mechanical properties [133]. Furthermore, Hagiwara *et al.* proposed a magnetically driven microtools that was able to orient, position and cut single cell [126]. These functions were used to enucleate the oocyte [127]. The authors argued that this method was significantly faster than the conventional mechanical micromanipulators and also caused less damage to the oocyte.

As the next approach, the cellular level manipulation capabilities of the untethered microrobot could be further strengthened to realize more applications as envisioned in Fig. 10. Control method, either on-board or off-board, could be introduced to render the microrobot to be a complete autonomous agent. In this case, a large number of microrobots could be released into the biomedical sample to finish predefined applications such as detecting circulating tumor cells [134] and systematically probing the cellular communication [135].

III. CHALLENGES AND EMERGING CONCEPTS IN MINIATURIZED BIOMEDICAL ROBOTS

To enable high-impact biomedical applications of miniaturized mobile robots, many fundamental challenges need to be addressed. As the functional robot size goes down to the millimeter scale and below, design, fabrication, and control of these systems require design principles which greatly differ from that of macro scale robotics. Moreover, medical activities inside the human body will require additional tasks such as feedback from the environment

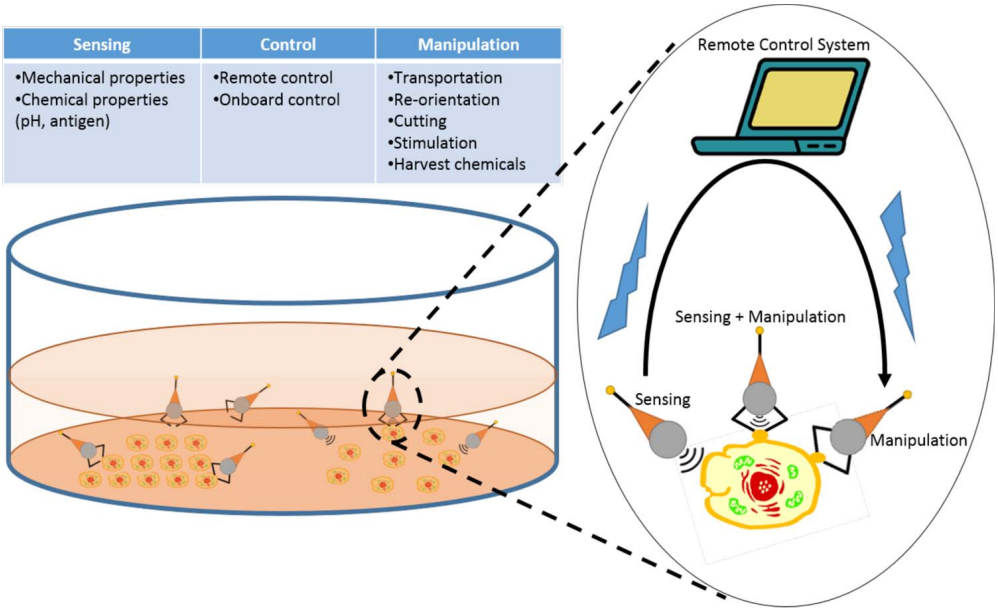


Fig. 10. ***Conceptual figure/illustration showing all potential applications of microrobotic cell manipulation.***

and communication with the operator. In this section, we discuss the challenges associated with miniaturization of untethered biomedical robots from their initial design to the preclinical testing steps. We also provide a future outlook toward a solution in light of the recent advances addressing some of these challenges.

A. Design and Modeling

How can we design a mobile milli/microrobot for a specific biomedical task to achieve optimal operational performance, such as the shortest operation duration, minimum power consumption, and largest area coverage, while constrained by software, hardware, manufacturing, motion, control, lifetime, and safety? Given that biological environments are remarkably crowded, the design of physically and chemically, robust, and flexible milli/microrobots is of paramount importance. Such a design requires an integrated strategy where components, locomotion principles, materials, and power sources are considered altogether for functioning via a real-time closed-loop control system (Fig. 11).

These design problems can be addressed in many different perspectives. One primary design variable is the number of milli/microrobots: a single multitasking robot versus a team [48], [59], [122], [136] or a swarm [137] of robots with parallel and distributed functions. Considering the potential size of a human tissue or organ, a single microrobot would be insufficient for enough theranostic effect in a given operation, while a microrobot team functioning in a concerted manner could significantly amplify the expected throughput. In the multi-robot perspective, each individual robot could be either identical (i.e., homogeneous) with the same functions or different (i.e., heterogeneous) with varying functionality [48]. The team could move deterministically or stochastically using on-board or off-board (remote) actuation methods [48], [136], [138]–[140]. As locomotion, they could swim, crawl, roll, spin, or hop [30],

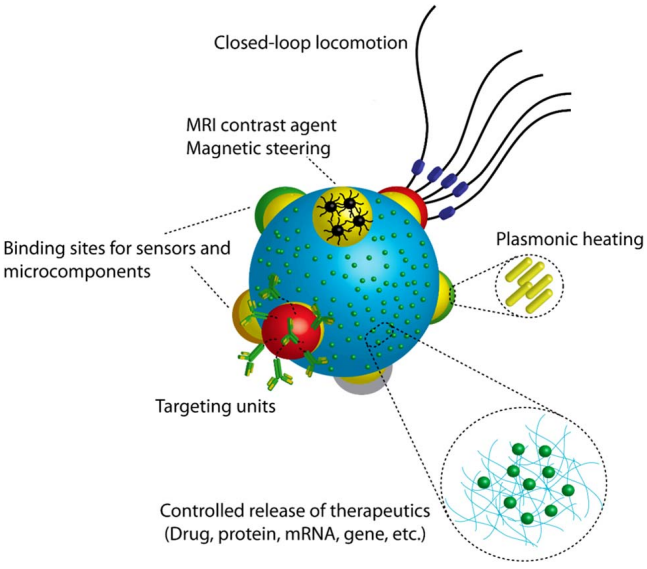


Fig. 11. ***Visionary design of a soft, modular microrobot with spatio-selective functionalization. Each functional component is assembled on a main board. The main board further serves as a large depot for therapeutics to launch controlled release at the site of action. A closed-loop autonomous locomotion (e.g., a biohybrid design) couples environmental signals to motility. Targeting units enable reaching and localization at the intended body site. MRI contrast agents loaded on the microrobot enables visualization as well as manual steering on demand. Gold nanorods enable plasmonic heating to decompose a tumor tissue.***

[33], [50], [61], [136], [138], [141]. They could have integrated micro/nano-sensors, microactuators, and other components such as microcontrollers, power source, wireless communication, etc. [39], [49], [142]–[145].

Programming individual components to spontaneously assemble into fully equipped multifunctional microrobots is a promising design strategy (Fig. 11) [146]. Reprogrammable self-assembly of small components into larger, complex structures is a universal route of material fabrication by biological organisms [147]. Individual components, called building blocks, carry the necessary information for structural integration/disintegration as well as specific biological functions. Despite the complexity of the final ensemble, reprogrammable assembly is a simple and robust strategy for rapid adaptation of the organism to dynamic changes in the environment. Such level of intelligence in biological systems provides a powerful source of inspiration for making similarly complex, synthetic designs, which should be functionally capable of multitasking and autonomously responding to changes in the environmental conditions. Modular assembly of individual micro and nanocomponents could therefore enable flexible customization of optimally working milli/microrobots, which could be manufactured in large quantities in a feasible and reliable way. However, the reprogrammable material concept is still at its infancy, and there is a need for thoroughly understanding and controlling assembly and related processes using simple and robust strategies. A major challenge in macroscopic self-assembly is coding information in individual building blocks. Recent work has addressed this issue by designing self-assembling soft building blocks in various size and shapes [49], [114], [148]–[151]. For example, colloidal patchy particles, which can form directional and programmable interactions in 3D, are among the state-of-the-art examples [152]. By spatioselective surface modification of individual building blocks, which range in 0.1 to a few micrometers, anisotropic and heterogeneous configuration could inspire similar robot designs based on the self- assembly concept (Fig. 11). In this regard, a similar approach would be useful for larger, i.e., 10 μm–1 mm, building blocks for manufacturing a microrobot. On the other hand, increased particle size creates many non-specific interaction sites, leading to the loss of the directionality and destabilized structural coherence. To this end, high fidelity directional bonding among the building blocks with high overall assembly yield remain as the major challenges to solve. One alternative to this would be remotely picking and then placing individual building blocks to assemble into 2D and 3D structures by the aid of a human operator [49], [114]. However, with this way, interactions between the individual building blocks usually remain weak, which does not support the overall structural integrity and the ensemble tends to fall apart. To surmount this, a secondary covalent cross-linking step is needed [49]. On the other hand, covalent cross-linking is an irreversible process that completely eliminates the intrinsic reprogrammable nature of the final ensemble. Therefore, another future task is to provide bonding stability while maintaining the dynamic nature of the self-assembly and ensure bonding directionality for building prescribed manufacturing of microrobots.

At the system level, real-time interactions and feedback among individual components of a milli/microrobot are essential for proper functioning. For an ideally autonomous microrobot, continuous sensing of the surrounding environment needs to be functionally coupled to mobility, cargo release, powering, and other operational components. Therefore, novel sensing mechanisms that modulate robot behavior would conditionally be able to activate operations. For example, sensing the location of a tumor site and subsequent taxis of microrobots to that location is crucial for carrying out a noninvasive medical operation. However, the major challenge of continuous sensing in the living environment is the unreliable biological signals that might cause false positive or false negatives, thereby leading to unintended microrobot activations. To surmount this problem, molecular logic gates sensing for multiple markers on a conditional basis would enable more accurate operational evaluations by milli/microrobots [153], [154]. In overall, there are alternative design approaches and variables one needs to select correctly for a given application. After developing approximate models of such milli/microrobot systems, rigorous numerical design optimization methods using evolutionary algorithms need to be developed as a significant future challenge.

B. Materials and Fabrication

Robots designed to be operating at the small scale is essentially a materials science problem because intelligence of such robots would mainly come from their physical material, structure, mechanism, and design properties. For any material coming into contact with biological fluids need to be resistant to corrosion, as highly saline aqueous environment could easily cause leaching hazardous products from robots as well as causing irreversible robotic malfunctions. Mechanical resilience and durability of milli/microrobots are also highly critical, particularly in large vessels and load-bearing tissues. Inside arteries, for example, high blood flow rate and shear forces can easily disintegrate tiny robots or prevent their motion control [2]. One bioinspired solution toward overcoming that issue might be recapitulation of erythrocyte deformability in milli/microrobots. Erythrocytes can change shape under applied stress without undergoing plastic deformation. There has already been an ongoing effort for developing injectable, shape memory polymers for tissue engineering applications [155], [156]. These materials can be compressed under large mechanical force and then completely recover repeatedly. Such a design could greatly help robust locomotion in blood vessels with changing diameter. For multicomponent systems, surface bonds should also be stable as these interconnections sites are the weakest points under mechanical stress. On top of all of these, robots interacting with biological tissues or working inside the

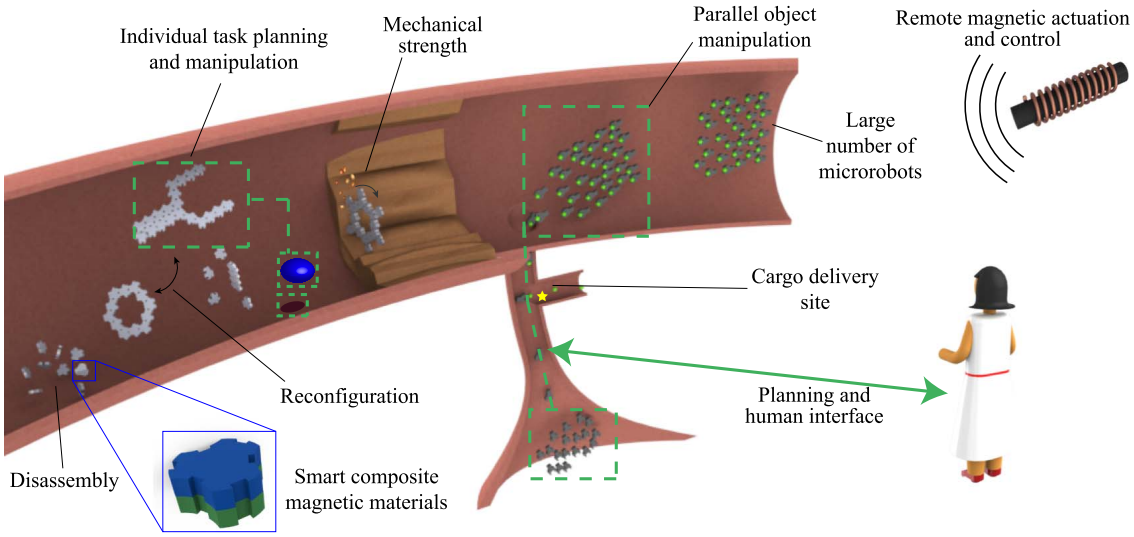


Fig. 12. ***Conceptual sketch of a large number of microrobots made of smart materials that can be remotely actuated and controlled inside the human body with a user interface to achieve different biomedical functions.***

human body must be biocompatible while most of the existing microrobots are made of materials that are not biocompatible. In most biomedical applications, it is crucial to also have novel materials that are soft, biodegradable, multi-functional, smart, and compatible to existing micro and nanofabrication processes. On the other hand, current robot materials are typically rigid, non-biodegradable, and have single function. Creating milli/microrobots from such novel materials require many custom and novel micro/nanoscale fabrication and prototyping tools in 2D and 3D that could be based on optical lithography, two-photon stereo-lithography, self-folding thin-films, micro/nanomachining, micro/nano-imprinting, and micro/nanomolding [49], [138], [151], [157], [158]. Finally, it is crucial to manufacture these robots in large numbers for their potential medical use (Fig. 12). Robot mass-production at the micro/nanoscale is integral for their future commercial applications using roll-to-roll, directed self-assembly, and programmable self-folding methods.

C. Functionality

In the millimeter scale, although active imaging is possible with current capsule millirobots, this function is primarily used for post-procedure diagnosis. In the future, it is imperative to go beyond this to advanced image processing for diagnosis of visually undetectable disease [159], to map the 3D environment of the given organ using visual simultaneous localization and mapping (SLAM) [160] or optical flow based advanced motion detection algorithms to predict the capsule motion precisely [161], and to propose new active focusing and 3D illumination methods to improve the imaging quality and diagnosis precision [162].

On the micron scale, the only practically available site for microrobot functionalization is its surface. Porous soft materials can also allow cargo encapsulation inside their 3D body. This would be a very useful strategy as it allows higher amount of cargo loading compared to 2D surface. There has been extensive experience over drug encapsulation and release for targeted therapy and controlled-release applications, which might be directly transferred to microrobotic applications [163]–[165]. For this purpose, a whole microrobot can be fabricated as a big cargo depot, which will significantly prolong the impact of single dose administration. In accordance with the special medical requirement, microrobot surface can be modified with operational microtools enabling the sensing of disease diagnosis, therapeutic functions, e.g., targeted drug or gene delivery, and surgical functions, e.g., cauterization and clearing clogged blood vessels. In this sense, mechanical microgrippers could be promising microtools for ablation and biopsy as well as drug/gene delivery [48], [117]. Similar microtools for drilling and heating local tissue sites could profoundly improve non-invasive surgical operations, particularly for removing tumor in deep tissue sites. High throughput or organized operations could find pervasive use in biomedicine. A typical example of microrobot swarms piece-by-piece building tissue scaffolds could revolutionize tissue engineering.

D. Mobility

For the capsule robots, there have been many 2D and 3D locomotion methods proposed. However, there are still many open challenges such as increasing the locomotion precision and speed for accurate and shorter operations, minimizing the power consumption during locomotion, increased safety for not damaging any tissue or not creating any negative reaction from the body, and robust operation

against the relative organ motion such as respiration, heart beating, and peristalsis. Also, every person with different age, gender, and race has a different scale and property of biological tissues. Therefore, the given locomotion method could be adapted to such variations robustly.

Possible locomotion modes of untethered biomedical microrobots are swimming in 3D liquid environment and walking, crawling, sliding, spinning, hopping, and rolling on 2D surfaces. Using such locomotion modes, microrobots should be able to navigate in hard-to-reach regions of the human body with high degree of mobility i.e., 6-DOF actuation and high steering capability [166], speed (achieving the tasks in reasonable durations for realistic clinical use), range of motion, penetration depth (i.e., reaching to the deep regions of the body), precision, and autonomy in teams or large numbers. Depending on its given task, a microrobot can have either or both 3D and 2D mobility to reach a specific site inside body. In body sites with low velocity or stagnant fluid flows, swimming and remote directing would be more efficient and faster, whereas in solid tissues or organs, 2D mobility might be the best option to penetrate into deep regions. To this end, autonomously switchable locomotion modes by sensing the body environment are significant challenges. Even so, minimum interaction with solid tissue surface would be desirable to avoid potential irritation and injury-related side effects. Speed control of a mobile microrobot is another critical factor for timely achievement of a given medical task. Synthetic micropropellers harvesting energy from a local source are far from providing a useful locomotion speed even in the unrestricted liquid environment, i.e., without the limitation of a solid tissue barrier. To the best of our knowledge, no microp propeller system has been demonstrated that can move against the blood stream in large vessels due to high-speed blood flow. Despite the fact that biological microorganisms can reach faster speed than synthetic and biomimetic microp propellers, none of the available sources (either natural or synthetic) has inspired for a practically useful speed for biomedical applications. For example, average swimming speed of a flagellum-carrying *E. coli* is 30 μm s^{-1} inside water [167]. A biohybrid design involving remote mobilization of magnetotactic bacteria was demonstrated to reach a maximum swimming speed of 200 μm s^{-1} [137].

Speed control is important for reaching to target site and completing the medical operation. In order to speed up in low Reynolds number, forces acting on the microrobot should be higher. Therefore, there is room for novel micromotor designs that will elevate the efficiency of harvesting local energy source by increasing the micromotor speed. For remotely controlled microrobots, the remote actuation torque or field gradient can simply be tuned to adjust speed [34] while there is a maximum speed limit in magnetic microrobots due to the roll-off behavior depending on the rotational drag properties of the robot.

E. Powering

One of the most significant bottlenecks of untethered mobile milli/microrobots is powering their mobility, sensing, communication, tools, and computation for long enough durations required for a given medical task. Capsule millirobots are powered by silver-oxide coin batteries inside the capsule shell that provide for approximately from 1 min to 8 hours of operation; for example, on-board actuated capsule can last for 1 min when they are actuated all the time, and just on-board imaging and data transfer can last up to 8 hours or so. There is always need for high power density power sources for longer operation durations. On the other hand, minimizing the energy consumption for sensing, locomotion, data transfer, and computation would help such grand challenge. As an alternative solution, wireless power transmission techniques such as inductive powering and radio frequency, microwave radiation, and piezoelectric ultrasound systems are promising options because they are off-board providing space for other modules on the capsule and increasing the operation duration [101]. However, when you scale down the capsule robot size significantly or increase the distance of the device from the power transmitter, such wireless power transfer efficiency goes down exponentially, reducing the provided average power numbers to approximately 1–20 mW.

On the micron scale, especially mobility requires significant amount of power as the motion at low Reynolds numbers could be significantly affected by the viscous drag on the robot body. Moreover, high mechanical power is needed for stable mobility control inside pulsating blood flow. At the sub-millimeter scale, storing, harvesting, and transmitting power is not feasible in the conventional sense we are used to in our macroscopic world. Therefore, a significant effort has been concentrated on various power sources, including remote magnetic, electrical, acoustic, and optical actuation and self-powering, including self-electrophoresis, self-diffusiophoresis, and self-thermophoresis, for microrobot locomotion [24], [168]–[173].

Biological systems have adapted to living in this size domain by storing energy in the form of chemical energy, which is then converted to mechanical motion, sensing, communication and reproduction. Similarly, autonomous microrobots should be powered by available local chemical energy inside the human body. To this end, a proof-of-concept gold-platinum bimetallic nanorod was demonstrated to autonomously move via self-electrophoresis in the presence of 2–3 vol.% H_2O_2 as the fuel [172]. Translating this technology to the micrometer scale, platinum nanoparticle catalyzed generation of oxygen gas drove motion of polymer stomatocytes at as low as 0.3 vol.% H_2O_2. Similar conceptual designs were shown to be operational in other liquids containing N,N-dimethyl hydrazine or methanol, though, none of which is close to a biologically relevant environment [174], [175]. To overcome this, a strategy that harnesses locally available sources is crucial. Mano and Heller's strategy of reacting glucose and oxygen was

promising to drive locomotion, though it is only operational at water-oxygen interface and requires high oxygen pressures. Recently, enzyme-powered micropumps have been shown to be viable source of motion in biologically relevant conditions [176].

Energy conversion efficiency is another concept that has so far received little attention. Energy conversion efficiency of microrobots can be described as the ratio of the mechanical power output to the overall work done to drive the motion. The efficiency of the synthetic micro and nano-propellers remain around 1%, significantly lower than macroscopic motors [167]. This might be a limiting step for the overall success of robotic operations.

Altogether, despite some solid progress in self-powering methods for microrobots harvesting the environmental liquids and flows and for sub-millimeter scale robots are still primitive and not directly applicable inside biological environment. It is therefore a great challenge to achieve remote or autonomous microrobot actuation for long durations in a wide range of mobility and inside deep regions of the human body. Maximizing the power efficiency and minimizing the power consumption of microrobotic systems are crucial for long-term medical operations, which could be enabled by optimal design of microrobot's mobility, sensing, and control methods.

F. Robot Localization

Determining the location and orientation of a medical robot in 3D is crucial for precise and safe motion control inside the human body. Many successful localization methods are available for millirobots [77], [78], [81]–[85], while localization of micron scale medical robots is a great challenge due to their much smaller size [94]. Thus, it is better to design such microrobot systems as swarms and facilitate stronger collective imaging signal [137]. Medical imaging systems such as MRI [71], [137], fluoroscopy [177], PET [178], NIR [179], and ultrasound [178] are possible candidates for microrobot localization. Under these systems, the localization could be registered with the medical images to plan and achieve medical tasks safely. At last, having multi-modal localization methods could enable more precise and safer medical operations [179], [180]. Even very early attempts towards precise localization of inside body will have profound impact in the field.

G. Communication

While many commercial transceivers are available for capsule millirobots, no one has tackled yet the challenge of wireless communication with microrobots inside the human body or communication among large number of microrobots, which could be crucial for data or information transfer from the robots to the doctor and vice versa and microrobot control and coordination. Magnetic actuation was proposed as a promising wireless strategy for cooperative [59], [70], [136] and distributed [48], [136] microrobotic tasks. However, effectiveness of distributed operations via magnetic actuation drastically diminishes with increase in the number of microrobots in the team. Further, magnetic actuation is an open-loop controller, lacking of autonomous decision-making based on real-time sensing of changes in the environment and state of individual microrobots. In this regard, principles that govern the social behaviors of biological microorganisms could be a valuable source of inspiration to address control and coordination of microrobot swarms. Microscopic species exhibit collective behaviors in response to environmental stimuli, which are sensed and transmitted among individual species by physical interactions and/or chemical secretions [181], [182]. *Dictyostelium discoideum* is a well-known example of such microorganisms, which, upon self-organization into a hierarchical colony with up to 105 residents, can reconfigure itself and migrate as a single unit [183]. Quorum sensing is another cell-to-cell communication process used in bacteria for sharing information among the population and eliciting a collective reaction [184]. An intriguing property of quorum sensing is that the population density is monitored in real-time by the whole colony and a communal response is elicited as a result [184]. This strategy is particularly inspirational for developing a population density-driven switch for microrobot operation inside body. microrobots gathering inside a specific body site and operating only after their population reaches a particular size would be a highly effective strategy.

H. Safety

It is mandatory to guarantee the safety of biomedical milli/microrobots while they are deployed, operated, extracted inside, and removed out of the human body. Such safety is only possible by designing and selecting proper materials and methods for fabrication, actuation, and powering from the very beginning of the system design and integration. Therefore, any robotic component, remote magnetic or other autonomous actuation or sensing methods should be within the FDA limits so that they don't cause any discomfort, damage, or pain to the patients; synthetic microrobots should be made of biocompatible and biodegradable soft materials; biohybrid (e.g., muscle-cell- or bacteria- actuated) microrobots should not be pathogenic or not create any immunological negative response. Immunogenicity concerns of muscle-cell-actuated microrobots could be successfully evaded by producing functional cells from patient-derived induced pluripotent stem cells (iPSC) [185]. On the other hand, during microrobot fabrication, biohybrid constructs are highly prone to microbial contamination, which should be given a special emphasis [186], [187]. Bacteria-propelled micro robots must be sterilized from any sort of pathogenicity. One safest way is genetically engineering these organisms, so that their proliferation and hazardous by-products are eliminated [188], [189].

While the magnetic strength of the microrobot itself will not present an issue, the magnetic fields used to actuate the microrobot need to be considered [190]. The

FDA currently classifies devices with static fields less than 8 T to be of nonsignificant risk. Current medical trials have shown fields upwards of 9.4 T to be safe, not affecting vital signs or cognitive ability [191]. A DC field of 16 T was shown to levitate a frog and other objects due to the weak diamagnetic properties of living tissue, with no observable negative effects [192]. Blood, which is electrically conductive, moving through a static field will generate a back electromotive force (EMF). A field of 10 T is calculated to reduce blood volume flow by 5% due to the effects of the induced voltage, possibly hazardous to susceptible patients [193]. The resulting current is expected to generate the upper limit of safe static fields [190]. However, the fringing fields at the end of an MRI device or solenoid can lead to large spatial magnetic gradients. A time-varying magnetic field or moving conductor will generate an induced current. The spatial magnetic gradient is used to push or pull magnetic microrobots. These spatial gradients will not harm the patient, however any movement will turn these gradients into time-varying magnetic gradients. On the other hand, if the magnetic microrobot is being precisely controlled, the spatial gradients will change with the control of the microrobot, causing time-varying gradients. Spatial gradients are reported for the patient accessible volume of MRI machines to reach several Tesla per meter, but are typically found outside the central bore and the procedures to measure the maximum possible spatial gradient are not well defined [194].

As discussed in Section III-F, there are several possible localization techniques, some of which may pose a risk to the patient. Using MRI to localize the medical microrobot has the same considerations as those above for actuating the microrobot. Imaging techniques based on ionizing radiation, such as fluoroscopy and positron emission tomography are only used when necessary. Fluoroscopy is limited by a patient dose to 88 mGy per minute by the FDA, and can even pose hazards to the operators [191]. Limitations on PET are already set for staff preparing the tracer nucleotide as well as during patient care. Exposure of 10–30 mSv have been reported for patients, and patients which underwent the procedure showed a higher incidence of cancer [195]. Ultrasonic radiation, while generally considered safe, is able to heat tissue and induce cavitation of gas bubbles. The FDA has set limits for beam intensity dependent on the frequency, pulse length, and number of pulses, ranging upwards to 2 W/cm^2 for pulsed-averaged intensities [196].

I. Preclinical Assessment Models

For gaining mechanistic insight into behaviors of miniaturized robots in a complex living environment, realistic *in vitro* medical models/phantoms or freshly acquired organs or tissues are essential. For cargo delivery and controlled release applications, existing tissue engineering models could be adapted for proof-of-concept investigations. In this regard, organ-on-chip technologies could be a valuable platform as the clinical and physiological mimetic of human body environment [197]. In addition to such *in vitro* testing, it would be crucial to have *in vivo* small animal proof-of-concept tests to show the preclinical feasibility of the proposed novel concepts.

IV. CONCLUSION

Small-scale untethered mobile robots have a promising future in healthcare and bioengineering applications [198]. They are unrivalled for accessing into small, highly confined and delicate body sites, where conventional medical devices fall short without an invasive intervention. Reconfigurable and modular designs of these robots could also allow for carrying out multiple tasks such as theranostic, i.e., both diagnostic and therapeutic, strategies. Notwithstanding, mobility, powering, and localization are the cardinal challenges that significantly limit the transition of viable robotic designs from *in vitro* to preclinical stage. An ideal self-powered microrobot that can be actuated autonomously, targeting a specific location to carry out a programmed function by real-time reporting to an outside operator would truly trigger a paradigm shift in clinical practice. Besides, individual robots that can form swarm-like assemblies for parallel and distributed operations would dramatically amplify their expected clinical outcome. Design and fabrication of miniaturized robots, particularly at the submillimeter scale, require a fundamentally different strategy than the existing macroscale manufacturing. Because surface-surface interactions predominate inertial forces, design and manufacturing at this size domain requires an interdisciplinary effort, particularly the involvement of robotic researchers, chemists, biomedical engineers, and materials scientists. Overall, even the currently presented primitive examples of untethered mobile milli/microrobots have opened new avenues in biomedical applications paving the way for minimally invasive and cost-effective strategies, thereby leading to fast recovery and increased quality of life of patients. ■

REFERENCES

[1] M. Sitti, "Miniature devices: Voyage of the micro-robots," *Nature*, vol. 458, pp. 1121–1122, Apr. 2009.

[2] B. J. Nelson, I. K. Kaliakatsos, and J. J. Abbott, "Micro-robots for minimally invasive medicine," *Annu. Rev. Biomed. Eng.*, vol. 12, pp. 55–85, Aug. 2010.

[3] M. F. Hale, R. Sidhu, and M. E. McAlindon, "Capsule endoscopy: Current practice and future directions," *World J. Gastroenterol.*, vol. 20, pp. 7752–7759, Jun. 2014.

[4] Z. Liao, R. Gao, C. Xu, and Z. S. Li, "Indications and detection, completion, retention rates of small-bowel capsule endoscopy: A systematic review," *Gastrointest. Endosc.*, vol. 71, pp. 280–286, Feb. 2010.

[5] Z. Li, Z. Liao, and M. McAlindon, *Handbook of Capsule Endoscopy*. Dordrecht, The Netherland: Springer Netherlands, 2014.

[6] M. K. Goenka, S. Majumder, and U. Goenka, "Capsule endoscopy: Present status and future expectation," *World J. Gastroenterol.*, vol. 20, pp. 10024–10037, Aug. 2014.

[7] T. Nakamura and A. Terano, "Capsule endoscopy: Past, present, future," *J. Gastroenterol.*, vol. 43, pp. 93–99, 2008.

[8] F. Munoz, G. Alici, and W. Li, "A review of drug delivery systems for capsule endoscopy," *Adv. Drug. Deliv. Rev.*, vol. 71, pp. 77–85, May 2014.

[9] F. Carpi, N. Kastelein, M. Talcott, and C. Pappone, "Magnetically controllable gastrointestinal steering of video capsules," *IEEE Trans. Biomed. Eng.*, vol. 58, pp. 231–234, Feb. 2011.

[10] H. Keller *et al.*, "Method for navigation and control of a magnetically guided capsule endoscope in the human stomach," in *Proc. 2012 4th IEEE RAS EMBS Int. Conf. Biomed. Robot. Biomechatron. (BioRob)*, Rome, Italy, 2012, pp. 859–865.

[11] A. W. Mahoney, S. E. Wright, and J. J. Abbott, "Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet," in *Proc. 2013 IEEE Int. Conf. Robot. Automation (ICRA)*, Karlsruhe, 2013, pp. 5366–5371.

[12] S. Yim, E. Gultepe, D. H. Gracias, and M. Sitti, "Biopsy using a magnetic capsule endoscope carrying, releasing, retrieving untethered microgrippers," *IEEE Trans. Biomed. Eng.*, vol. 61, pp. 513–521, Feb. 2014.

[13] A. J. Petruska and J. J. Abbott, "An omnidirectional electromagnet for remote manipulation," in *Proc. 2013 IEEE Int. Conf. Robot. Automation (ICRA)*, Karlsruhe, 2013, pp. 822–827.

[14] R. W. Carlsen and M. Sitti, "Biohybrid cell-based actuators for microsystems," *Small*, vol. 10, pp. 3831–3851, Oct. 2014.

[15] M. Sitti, "Microscale and nanoscale robotics systems [Grand Challenges of Robotics]," *IEEE Robot. Autom. Mag.*, vol. 14, pp. 53–60, 2007.

[16] D. C. Meeker, E. H. Maslen, R. C. Ritter, and F. M. Creighton, "Optimal realization of arbitrary forces in a magnetic stereotaxis system," *IEEE Trans. Magn.*, vol. 32, pp. 320–328, 1996.

[17] G. Caprari, P. Balmer, R. Piguet, and R. Siegwart, "The autonomous micro robot 'alice': A platform for scientific and commercial applications," in *Proc. 1998 Int. Symp. Micromechatron. Human Sci. (MHS'98)*, 1998, pp. 231–235.

[18] N. Kawahara, T. Shibata, and T. Sasaya, "In-pipe wireless micro-robot," in *Proc. Photonics East'99*, Boston, MA, USA, 1999, pp. 166–171.

[19] R. S. Fearing *et al.*, "Wing transmission for a micromechanical flying insect," in *Proc. 2000 IEEE Int. Conf. Robot. Automation (ICRA)*, San Francisco, CA, USA, 2000, pp. 1509–1516.

[20] K. Ishiyama, M. Sendoh, A. Yamazaki, and K. I. Arai, "Swimming micro-machine driven by magnetic torque," *Sensor. Actuat. A-Phys.*, vol. 91, pp. 141–144, 2001.

[21] S. Hollar, A. Flynn, C. Bellew, and K. Pister, "Solar powered 10 mg silicon robot," in *Proc. 2003 IEEE Sixteenth Annu. Int. Conf. Micro Electro Mechanical Systems (MEMS)*, Kyoto, Japan, 2003, pp. 706–711.

[22] N. A. Patronik, M. A. Zenati, and C. N. Riviere, "Crawling on the heart: A mobile robotic device for minimally invasive cardiac interventions," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2004*. New York, NY, USA: Springer, 2004, pp. 9–16.

[23] N. Darnton, L. Turner, K. Breuer, and H. C. Berg, "Moving fluid with bacterial carpets," *Biophys. J.*, vol. 86, pp. 1863–1870, Mar. 2004.

[24] R. Dreyfus *et al.*, "Microscopic artificial swimmers," *Nature*, vol. 437, pp. 862–865, Oct. 2005.

[25] Y. S. Song and M. Sitti, "Surface-tension-driven biologically inspired water strider robots: Theory and experiments," *IEEE Trans. Robot.*, vol. 23, pp. 578–589, 2007.

[26] A. M. Hoover, E. Steltz, and R. S. Fearing, "RoACH: An autonomous 2.4 g crawling hexapod robot," in *Proc. IEEE/RSJ 2008 Int. Conf. Intell. Robots Syst. (IROS)*, Nice, France, 2008, pp. 26–33.

[27] M. Quirini, R. Webster, A. Menciassi, and P. Dario, "Design of a pill-sized 12-legged endoscopic capsule robot," in *Proc. 2007 IEEE Int. Conf. Robot. Automation (ICRA)*, Rome, Italy, 2007, pp. 1856–1862.

[28] A. Degani, H. Choset, A. Wolf, and M. A. Zenati, "Highly articulated robotic probe for minimally invasive surgery," in *Proc. 2006 IEEE Int. Conf. Robot. Automation (ICRA)*, Orlando, FL, USA, 2006, pp. 4167–4172.

[29] S. Martel *et al.*, "Automatic navigation of an untethered device in the artery of a living animal using a conventional clinical magnetic resonance imaging system," *Appl. Phys. Lett.*, vol. 90, p. 114105, 2007.

[30] B. R. Donald, C. G. Levey, C. D. McGray, I. Paprotny, and D. Rus, "An untethered, electrostatic, globally controllable MEMS micro-robot," *J. Micromech. Microeng.*, vol. 15, pp. 1–15, 2006.

[31] O. J. Sul, M. R. Falvo, R. M. Taylor, S. Washburn, and R. Superfine, "Thermally actuated untethered impact-driven locomotive microdevices," *Appl. Phys. Lett.*, vol. 89, p. 203512, 2006.

[32] S. Martel, "Magnetotactic bacteria as controlled functional carriers in microsystems, microelectronic circuits and interconnections," in *Proc. 16th Eur. Microelectron. Packaging Conf. (EMPC)*, Qulu, 2007.

[33] C. Pawashe, S. Floyd, and M. Sitti, "Modeling and experimental characterization of an untethered magnetic micro-robot," *Int. J. Robot. Res.*, vol. 28, pp. 1077–1094, 2009.

[34] L. Zhang *et al.*, "Characterizing the swimming properties of artificial bacterial flagella," *Nano. Lett.*, vol. 9, pp. 3663–3667, Oct. 2009.

[35] A. Ghosh and P. Fischer, "Controlled propulsion of artificial magnetic nanostructured propellers," *Nano. Lett.*, vol. 9, pp. 2243–2245, Jun. 2009.

[36] Y. Sehyuk and M. Sitti, "Shape-programmable soft capsule robots for semi-implantable drug delivery," *IEEE Trans. Robot.*, vol. 28, pp. 1198–1202, 2012.

[37] S. Miyashita, E. Diller, and M. Sitti, "Two-dimensional magnetic micro-module reconfigurations based on inter-modular interactions," *Int. J. Robot. Res.*, vol. 32, pp. 591–613, 2013.

[38] C. Pawashe, S. Floyd, and M. Sitti, "Multiple magnetic micro-robot control using electrostatic anchoring," *Appl. Phys. Lett.*, vol. 94, pp. 164108–164108-3, 2009.

[39] W. Hu, K. S. Ishii, and A. T. Ohta, "Micro-assembly using optically controlled bubble micro-robots," *Appl. Phys. Lett.*, vol. 99, p. 094103, 2011.

[40] M. P. Kummer *et al.*, "OctoMag: An electromagnetic system for 5-DOF wireless micromanipulation," *IEEE Trans. Robot.*, vol. 26, pp. 1006–1017, 2010.

[41] V. Magdanz, S. Sanchez, and O. G. Schmidt, "Development of a sperm-flagella driven micro-bio-robot," *Adv. Mater.*, vol. 25, pp. 6581–6588, Dec. 2013.

[42] A. A. Solovev, Y. Mei, E. Bermudez Urena, G. Huang, and O. G. Schmidt, "Catalytic microtubular jet engines self-propelled by accumulated gas bubbles," *Small*, vol. 5, pp. 1688–1692, Jul. 2009.

[43] A. Búzás *et al.*, "Light sailboats: Laser driven autonomous micro-robots," *Appl. Phys. Lett.*, vol. 101, p. 041111, 2012.

[44] S. Martel and M. Mohammadi, "Using a swarm of self-propelled natural micro-robots in the form of flagellated bacteria to perform complex micro-assembly tasks," in *Proc. 2014 IEEE Int. Conf. Robot. Automation (ICRA)*, Hong Kong, 2010, pp. 500–505.

[45] M. Rubenstein, A. Cornejo, and R. Nagpal, "Programmable self-assembly in a thousand-robot swarm," *Science*, vol. 345, pp. 795–799, 2014.

[46] K. Y. Ma, P. Chirarattananon, S. B. Fuller, and R. J. Wood, "Controlled flight of a biologically inspired, insect-scale robot," *Science*, vol. 340, pp. 603–607, May 3, 2013.

[47] E. Diller, J. Zhuang, G. Z. Lum, M. R. Edwards, and M. Sitti, "Continuously distributed magnetization profile for millimeter-scale elastomeric undulatory swimming," *Appl. Phys. Lett.*, vol. 104, p. 174101, 2014.

[48] E. Diller and M. Sitti, "Three-dimensional programmable assembly by untethered magnetic robotic micro-grippers," *Adv. Funct. Mater.*, vol. 24, pp. 4397–4404, 2014.

[49] S. Tasoglu, E. Diller, S. Guven, M. Sitti, and U. Demirci, "Untethered micro-robotic coding of three-dimensional material composition," *Nat. Commun.*, vol. 5, p. 3124, Jan. 2014.

[50] Y. Zhou, S. Regnier, and M. Sitti, "Rotating magnetic miniature swimming robots with multiple flexible flagella," *IEEE Trans. Robot.*, vol. 30, pp. 3–13, 2014.

[51] G. Caprari, T. Estier, and R. Siegwart, "Fascination of down scaling-Alice the sugar cube robot," *J. Micromechatron.*, vol. 1, pp. 177–189, 2001.

[52] M. Quirini, A. Menciassi, S. Scapellato, C. Stefanini, and P. Dario, "Design and fabrication of a motor legged capsule for the active exploration of the gastrointestinal tract," *IEEE/ASME Trans. Mechatron.*, vol. 13, pp. 169–179, 2008.

[53] S. Yim and M. Sitti, "Design and rolling locomotion of a magnetically actuated soft capsule endoscope," *IEEE Trans. Robot.*, vol. 28, pp. 183–194, 2012.

[54] J. Edd, S. Payen, B. Rubinsky, M. L. Stoller, and M. Sitti, "Biomimetic propulsion for a swimming surgical micro-robot," in *Proc. 2003 IEEE/RSJ Int. Conf. Intelligent Robots Syst. (IROS)*, Las Vegas, NV, USA, 2003, pp. 2583–2588.

[55] B. Behkam and M. Sitti, "Bacterial flagella-based propulsion and on/off motion control of microscale objects," *Appl. Phys. Lett.*, vol. 90, p. 023902, 2007.

[56] K. B. Yesin, "Modeling and control of untethered biomicro-robots in a fluidic environment using electromagnetic fields," *Int. J. Robot. Res.*, vol. 25, pp. 527–536, 2006.

[57] K. E. Peyer, S. Tottori, F. Qiu, L. Zhang, and B. J. Nelson, "Magnetic helical micromachines," *Chemistry*, vol. 19, pp. 28–38, Jan. 2013.

[58] S. Martel, C. C. Tremblay, S. Ngakeng, and G. Langlois, "Controlled manipulation and actuation of micro-objects with

magnetotactic bacteria," *Appl. Phys. Lett.*, vol. 89, p. 233904, 2006.

[59] D. R. Frutiger, K. Vollmers, B. E. Kratochvil, and B. J. Nelson, "Small, fast, under control: Wireless resonant magnetic micro-agents," *Int. J. Robot. Res.*, vol. 29, pp. 613–636, Apr. 2009.

[60] M. S. Sakar, E. B. Steager, A. Cowley, V. Kumar, and G. J. Pappas, "Wireless manipulation of single cells using magnetic microtransporters," in *Proc. 2011 IEEE Int. Conf. Robot. Automation (ICRA)*, Shanghai, China, 2011, pp. 2668–2673.

[61] G. L. Jiang *et al.*, "Development of rolling magnetic micro-robots," *J. Micromech. Microeng.*, vol. 20, p. 085042, 2010.

[62] R. Pelrine *et al.*, "Diamagnetically levitated robots: An approach to massively parallel robotic systems with unusual motion properties," in *Proc. 2012 IEEE Int. Conf. Robot. Automation (ICRA)*, St. Paul, MN, USA, 2012, pp. 739–744.

[63] A. Snezhko and I. S. Aranson, "Magnetic manipulation of self-assembled colloidal asters," *Nat. Mater.*, vol. 10, pp. 698–703, Sep. 2011.

[64] W. Jing *et al.*, "A magnetic thin film micro-robot with two operating modes," in *Proc. 2014 IEEE Int. Conf. Robot. Automation (ICRA)*, Hong Kong, 2011, pp. 96–101.

[65] B. R. Donald, C. G. Levey, and I. Paprotny, "Planar microassembly by parallel actuation of MEMS micro-robots," *J. Micromech. Microeng.*, vol. 17, pp. 789–808, 2008.

[66] I. A. Ivan *et al.*, "First experiments on MagPieR: A planar wireless magnetic and piezoelectric micro-robot," in *Proc. 2014 IEEE Int. Conf. Robot. Automation (ICRA)*, Hong Kong, 2011, pp. 102–108.

[67] E. Schaler, M. Tellers, A. Gerratt, I. Penskiy, and S. Bergbreiter, "Toward fluidic micro-robots using electrowetting," in *Proc. 2012 IEEE Int. Conf. Robot. Automation (ICRA)*, St. Paul, MN, USA, 2012, pp. 3461–3466.

[68] G. Hwang *et al.*, "Electro-osmotic propulsion of helical nanobelt swimmers," *Int. J. Robot. Res.*, vol. 30, pp. 806–819, 2011.

[69] A. Yamazaki *et al.*, "Wireless micro swimming machine with magnetic thin film," *J. Magn. Magn. Mater.*, vol. 272–276, pp. E1741–E1742, 2004.

[70] S. Tottori *et al.*, "Magnetic helical micromachines: Fabrication, controlled swimming, cargo transport," *Adv. Mater.*, vol. 24, pp. 811–816, Feb. 2012.

[71] S. Martel *et al.*, "MRI-based medical nanorobotic platform for the control of magnetic nanoparticles and flagellated bacteria for target interventions in human capillaries," *Int. J. Rob. Res.*, vol. 28, pp. 1169–1182, Sep. 2009.

[72] D. H. Kim, P. S. S. Kim, A. A. Julius, and M. J. Kim, "Three-dimensional control of Tetrahymena pyriformis using artificial magnetotaxis," *Appl. Phys. Lett.*, vol. 100, p. 053702, 2012.

[73] B. J. Williams, S. V. Anand, J. Rajagopalan, and M. T. Saif, "A self-propelled biohybrid swimmer at low Reynolds number," *Nat. Commun.*, vol. 5, p. 3081, 2014.

[74] D. H. Kim, P. S. S. Kim, A. A. Julius, and M. J. Kim, "Three-dimensional control of engineered motile cellular micro-robots," in *Proc. 2012 IEEE Int. Conf. Robot. Automation (ICRA)*, St. Paul, MN, USA, 2012.

[75] G. Ciuti, A. Menciassi, and P. Dario, "Capsule endoscopy: From current achievements to open challenges," *IEEE Rev. Biomed. Eng.*, vol. 4, pp. 59–72, 2011.

[76] G. Pan and L. Wang, "Swallowable wireless capsule endoscopy: Progress and technical challenges," *Gastroenterol Res. Pract.*, vol. 2012, p. 841691, 2012.

[77] T. D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 2387–2399, Sep. 2012.

[78] M. Fluckiger and B. J. Nelson, "Ultrasound emitter localization in heterogeneous media," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Medicine Biology Society*, Lyon, France, 2007, pp. 2867–2870.

[79] J. M. Rubin *et al.*, "Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans," *J. Ultrasound. Med.*, vol. 25, pp. 1179–1186, Sep. 2006.

[80] K. Kim *et al.*, "Noninvasive ultrasound elasticity imaging (UEI) of Crohn's disease: Animal model," *Ultrasound Med. Biol.*, vol. 34, pp. 902–912, Jun. 2008.

[81] T. D. Than *et al.*, "An effective localization method for robotic endoscopic capsules using multiple positron emission markers," *IEEE Trans. Robot.*, vol. 30, pp. 1174–1186, 2014.

[82] C. Di Natali, M. Beccani, and P. Valdastri, "Real-time pose detection for magnetic medical devices," *IEEE Trans. Magn.*, vol. 49, pp. 3524–3527, 2013.

[83] C. Hu, M. Q. H. Meng, and M. Mandal, "Efficient magnetic localization and orientation technique for capsule endoscopy," *Int. J. Inform. Acquisition*, vol. 2, pp. 23–36, 2005.

[84] D. Soon, 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," submitted for publication.

[85] S. Yim and M. Sitti, "3-D localization method for a magnetically actuated soft capsule endoscope and its applications," *IEEE Trans. Robot.*, vol. 29, pp. 1139–1151, Jun. 2013.

[86] AIWABA. [Online]. Available: http://www.awaiba.com/product/naneye/

[87] O. Ergeneman, G. Dogangil, M. P. Kummer, J. J. Abbott, M. K. Nazeeruddin, and B. J. Nelson, "A magnetically controlled wireless optical oxygen sensor for intraocular measurements," *IEEE Sensors J.*, vol. 8, pp. 29–37, 2008.

[88] B. P. Timko, T. Dvir, and D. S. Kohane, "Remotely triggerable drug delivery systems," *Adv. Mater.*, vol. 22, pp. 4925–4943, Nov. 2010.

[89] S. P. Woods and T. G. Constandinou, "Wireless capsule endoscope for targeted drug delivery: Mechanics and design considerations," *IEEE Trans. Biomed. Eng.*, vol. 60, pp. 945–953, Apr. 2013.

[90] J. Cui, X. Zheng, W. Hou, Y. Zhuang, X. Pi, and J. Yang, "The study of a remote-controlled gastrointestinal drug delivery and sampling system," *Telemed. J. E Health*, vol. 14, pp. 715–719, Sep. 2008.

[91] P. Xitian, L. Hongying, W. Kang, L. Yulin, Z. Xiaolin, and W. Zhiyu, "A novel remote controlled capsule for site-specific drug delivery in human GI tract," *Int. J. Pharm.*, vol. 382, pp. 160–164, Dec. 2009.

[92] S. Yim, K. Goyal, and M. Sitti, "Magnetically actuated soft capsule with the multimodal drug release function," *IEEE ASME Trans. Mechatron.*, vol. 18, pp. 1413–1418, Jan. 2013.

[93] S. Martel, "Bacterial microsystems and micro-robots," *Biomed. Microdevices*, vol. 14, pp. 1033–1045, Dec. 2012.

[94] S. Martel, "micro-robotics in the vascular network: Present status and next challenges," *J.Micro-Bio Robot.*, vol. 8, pp. 41–52, 2013.

[95] P. Pouponneau, J. C. Leroux, G. Soulez, L. Gaboury, and S. Martel, "Co-encapsulation of magnetic nanoparticles and doxorubicin into biodegradable microcarriers for deep tissue targeting by vascular MRI navigation," *Biomaterials*, vol. 32, pp. 3481–3486, May 2011.

[96] S. Fusco, G. Chatzipirpiridis, K. M. Sivaraman, O. Ergeneman, B. J. Nelson, and S. Pane, "Chitosan electrodeposition for micro-robotic drug delivery," *Adv. Healthc. Mater.*, vol. 2, pp. 1037–1044, Jul. 2013.

[97] S. Martel, S. Taherkhani, M. Tabrizian, M. Mohammadi, D. de Lanauze, and O. Felfoul, "Computer 3D controlled bacterial transports and aggregations of microbial adhered nano-components," *J. Micro-Bio Robot.*, vol. 9, pp. 23–28, 2014.

[98] R. W. Carlsen, M. R. Edwards, J. Zhuang, C. Pacoret, and M. Sitti, "Magnetic steering control of multi-cellular biohybrid microswimmers," *Lab Chip*, vol. 14, pp. 3850–3859, Oct. 2014.

[99] K. Kong, J. Cha, D. Jeon, and D. Cho, "A rotational micro biopsy device for the capsule endoscope," in *Proc. 2005 IEEE/RSJ Int. Conf. Intell. Robots Syst., 2005 (IROS 2005)*, Edmonton, AB, Canada, 2005, pp. 1839–1843.

[100] S. Park *et al.*, "A novel micro-biopsy actuator for capsular endoscope using LIGA process," in *Proc. 2007 Solid-State Sensors, Actuators and Microsystems Conf.*, Lyon, 2007, p. 209.

[101] S. Park *et al.*, "A novel microactuator for microbiopsy in capsular endoscopes," *J. Micromech. Microeng.*, vol. 18, 2008.

[102] M. Simi, G. Gerboni, A. Menciassi, and P. Valdastri, "Magnetic torsion spring mechanism for a wireless biopsy capsule," *J. Med. Devices*, vol. 7, p. 041009, 2013.

[103] F. Ullrich *et al.*, "Mobility experiments with micro-robots for minimally invasive intraocular surgery," *Invest. Ophthalmol. Vis. Sci.*, vol. 54, pp. 2853–2863, Apr. 2013.

[104] C. Yu *et al.*, "Novel electromagnetic actuation system for three-dimensional locomotion and drilling of intravascular micro-robot," *Sensor. Actuat. A-Phys.*, vol. 161, pp. 297–304, 2010.

[105] P. Miloro, E. Sinibaldi, A. Menciassi, and P. Dario, "Removing vascular obstructions: A challenge, yet an opportunity for interventional microdevices," *Biomed. Microdevices*, vol. 14, pp. 511–532, Jun. 2012.

[106] I. J. Fox, G. Q. Daley, S. A. Goldman, J. Huard, T. J. Kamp, and M. Trucco, "Stem cell therapy. Use of differentiated pluripotent stem cells as replacement therapy for treating disease," *Science*, vol. 345, p. 1247391, Aug. 2014.

[107] S. Kim *et al.*, "Fabrication and characterization of magnetic micro-robots for three-dimensional cell culture and targeted transportation," *Adv. Mater.*, vol. 25, pp. 5863–5868, Nov. 2013.

[108] A. Khademhosseini, R. Langer, J. Borenstein, and J. P. Vacanti, "Microscale technologies for tissue engineering and biology," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 103, pp. 2480–2487, Feb. 2006.

[109] S. Masuda, T. Shimizu, M. Yamato, and T. Okano, "Cell sheet engineering for heart tissue repair," *Adv. Drug. Deliv. Rev.*, vol. 60, pp. 277–285, Jan. 2008.

[110] D. L. Elbert, "Bottom-up tissue engineering," *Curr. Opin. Biotechnol.*, vol. 22, pp. 674–680, Oct. 2011.

[111] Y. Ling *et al.*, "A cell-laden microfluidic hydrogel," *Lab Chip*, vol. 7, pp. 756–762, Jun. 2007.

[112] P. Zorlutuna *et al.*, "Microfabricated biomaterials for engineering 3D tissues," *Adv. Mater.*, vol. 24, pp. 1782–1804, Apr. 2012.

[113] J. S. Liu and Z. J. Gartner, "Directing the assembly of spatially organized multicomponent tissues from the bottom up," *Trends Cell Biol*, vol. 22, pp. 683–691, Dec. 2012.

[114] S. Tasoglu *et al.*, "Guided and magnetic self-assembly of tunable magnetoceptive gels," *Nat. Commun.*, vol. 5, p. 4702, 2014.

[115] U. Mirsaidov *et al.*, "Live cell lithography: Using optical tweezers to create synthetic tissue," *Lab Chip*, vol. 8, pp. 2174–2181, Dec. 2008.

[116] D. R. Albrecht, G. H. Underhill, T. B. Wassermann, R. L. Sah, and S. N. Bhatia, "Probing the role of multicellular organization in three-dimensional microenvironments," *Nat. Methods*, vol. 3, pp. 369–375, May 2006.

[117] J. Giltinan, E. Diller, C. Mayda, and M. Sitti, "Three-dimensional robotic manipulation and transport of micro-scale objects by a magnetically driven capillary micro-gripper," in *Proc. IEEE Int. Conf. Robot. Automation (ICRA)*, Hong Kong, 2014, pp. 2077–2082.

[118] D. Di Carlo, H. T. K. Tse, and D. R. Gossett, "Introduction: Why Analyze Single Cells?" in *Single-Cell Analysis*, vol. 853, S. Lindström and H. Andersson-Svahn, Eds. Upper Saddle River, NJ, USA: Humana Press, 2012, pp. 1–10.

[119] B. F. Brehm-Stecher and E. A. Johnson, "Single-cell microbiology: Tools, technologies, applications," *Microbiol. Mol. Biol. Rev.*, vol. 68, pp. 538–559, Sep. 2004.

[120] Z. Lu, C. Moraes, G. Ye, C. A. Simmons, and Y. Sun, "Single cell deposition and patterning with a robotic system," *PLoS One*, vol. 5, p. e13542, 2010.

[121] E. B. Steager *et al.*, "Automated biomanipulation of single cells using magnetic micro-robots," *Int. J. Robot. Res.*, vol. 32, pp. 346–359, 2013.

[122] W. Hu, K. S. Ishii, Q. Fan, and A. T. Ohta, "Hydrogel micro-robots actuated by optically generated vapour bubbles," *Lab Chip*, vol. 12, pp. 3821–3826, Oct. 2012.

[123] J. O. Kwon, J. S. Yang, J. B. Chae, and S. K. Chung, "Micro-object manipulation in a microfabricated channel using an electromagnetically driven micro-robot with an acoustically oscillating bubble," *Sensor. Actuat. A-Phys.*, vol. 215, pp. 77–82, 2014.

[124] Z. Ye and M. Sitti, "Dynamic trapping and two-dimensional transport of swimming microorganisms using a rotating magnetic micro-robot," *Lab Chip*, vol. 14, pp. 2177–2182, Jul. 2014.

[125] S. Schurle *et al.*, "Three-dimensional, automated magnetic biomanipulation with subcellular resolution," in *Proc. 2013 IEEE Int. Conf. Robot. Automation (ICRA)*, Karlsruhe, Germany, pp. 1452–1457.

[126] M. Hagiwara *et al.*, "On-chip magnetically actuated robot with ultrasonic vibration for single cell manipulations," *Lab Chip*, vol. 11, pp. 2049–2054, Jun. 2011.

[127] L. Feng, M. Hagiwara, A. Ichikawa, and F. Arai, "On-chip enucleation of bovine oocytes using micro-robot-assisted flow-speed control," *Micromachines*, vol. 4, pp. 272–285, 2013.

[128] T. Kawahara *et al.*, "On-chip micro-robot for investigating the response of aquatic microorganisms to mechanical stimulation," *Lab Chip*, vol. 13, pp. 1070–1078, Mar. 2013.

[129] L. Zhang, T. Petit, K. E. Peyer, and B. J. Nelson, "Targeted cargo delivery using a rotating nickel nanowire," *Nanomedicine*, vol. 8, pp. 1074–1080, Oct. 2012.

[130] T. Petit, L. Zhang, K. E. Peyer, B. E. Kratochvil, and B. J. Nelson, "Selective trapping and manipulation of microscale objects using mobile microvortices," *Nano. Lett.*, vol. 12, pp. 156–160, Jan. 2012.

[131] G. Thalhammer, R. Steiger, S. Bernet, and M. Ritsch-Marte, "Optical macro-tweezers: Trapping of highly motile micro-organisms," *J. Optics*, vol. 13, p. 044024, 2011.

[132] H. Zhang and K. K. Liu, "Optical tweezers for single cells," *J. R. Soc. Interface.*, vol. 5, pp. 671–690, Jul. 2008.

[133] D. H. Kim, P. K. Wong, J. Park, A. Levchenko, and Y. Sun, "Microengineered platforms for cell mechanobiology," *Annu. Rev. Biomed. Eng.*, vol. 11, pp. 203–233, 2009.

[134] L. Hajba and A. Guttman, "Circulating tumor-cell detection and capture using microfluidic devices," *TrAC Trend. Anal. Chem.*, vol. 59, pp. 9–16, 2014.

[135] C. H. Collins, "Cell-cell communication special issue," *ACS Synth. Biol.*, vol. 3, pp. 197–198, Apr. 2014.

[136] T. Y. Huang *et al.*, "Cooperative manipulation and transport of microobjects using multiple helical microcarriers," *RSC Adv.*, vol. 4, pp. 26771–26771, 2014.

[137] S. Martel, M. Mohammadi, O. Felfoul, Z. Lu, and P. Pouponneau, "Flagellated magnetotactic bacteria as controlled MRI-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature," *Int. J. Rob. Res.*, vol. 28, pp. 571–582, Apr. 2009.

[138] M. A. Zeeshan *et al.*, "Hybrid helical magnetic micro-robots obtained by 3D template-assisted electrodeposition," *Small*, vol. 10, pp. 1284–1288, 2014.

[139] S. Bergbreiter and K. S. J. Pister, Eds., "Design of an Autonomous Jumping micro-robot," in *Proc. IEEE Int. Conf. Robot. Automation (ICRA)*, Rome, Italy, 2007.

[140] W. A. Churaman, L. J. Currano, C. J. Morris, J. E. Rajkowski, and S. Bergbreiter, "The first launch of an autonomous thrust-driven micro-robot using nanoporous energetic silicon," *J. Microelectromech. Syst.*, vol. 21, pp. 198–205, 2012.

[141] W. Jing, N. Pagano, and D. J. Cappelleri, "A novel micro-scale magnetic tumbling micro-robot," *J. of Micro-Bio Robotics*, vol. 8, pp. 1–12, Feb. 2013.

[142] B. H. McNaughton, J. N. Anker, and R. Kopelman, "Magnetic microdrill as a modulated fluorescent pH sensor," *J. Magn. Magn. Mater.*, vol. 293, pp. 696–701, 2005.

[143] O. Ergeneman *et al.*, "Cobalt-nickel microcantilevers for biosensing," *J. Intell. Mater. Syst. Struct.*, vol. 24, pp. 2215–2220, Oct. 2012.

[144] W. Jing and D. J. Cappelleri, "Incorporating in-situ force sensing capabilities in a magnetic micro-robot," in *Proc. 2014 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2014)*, Chicago, IL, USA, 2014, pp. 4704–4709.

[145] O. Ergeneman *et al.*, "*In vitro* oxygen sensing using intraocular micro-robots," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 3104–3109, Nov. 2012.

[146] G. M. Whitesides and B. Grzybowski, "Self-assembly at all scales," *Science*, vol. 295, pp. 2418–2421, Mar. 2002.

[147] D. J. Kushner, "Self-assembly of biological structures," *Bacteriol. Rev.*, vol. 33, pp. 302–345, 1969.

[148] S. Tasoglu *et al.*, "Paramagnetic levitational assembly of hydrogels," *Adv. Mater.*, vol. 25, pp. 1137–1143, 2013.

[149] F. Xu *et al.*, "The assembly of cell-encapsulating microscale hydrogels using acoustic waves," *Biomaterials*, vol. 32, pp. 7847–7855, Jul. 2011.

[150] F. Xu *et al.*, "Three-dimensional magnetic assembly of microscale hydrogels," *Adv. Mater.*, vol. 23, pp. 4254–4260, 2011.

[151] H. Qi *et al.*, "DNA-directed self-assembly of shape-controlled hydrogels," *Nat. Commun.*, vol. 4, Sep. 2013.

[152] G. R. Yi, D. J. Pine, and S. Sacanna, "Recent progress on patchy colloids and their self-assembly," *J. Phys.: Condensed Matter*, vol. 25, p. 193101, 2013.

[153] P. A. de Silva, N. H. Q. Gunaratne, and C. P. McCoy, "A molecular photoionic AND gate based on fluorescent signalling," *Nature*, vol. 364, pp. 42–44, 1993.

[154] S. Ozlem and E. U. Akkaya, "Thinking outside the silicon box: Molecular and logic as an additional layer of selectivity in singlet oxygen generation for photodynamic therapy," *J. Amer. Chem. Soc.*, vol. 131, pp. 48–49, Jan. 2009.

[155] S. A. Bencherif *et al.*, "Injectable preformed scaffolds with shape-memory properties," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 109, pp. 19 590–19 595, Nov. 2012.

[156] M. Behl and A. Lendlein, "Shape-memory polymers," *Materials Today*, vol. 10, pp. 20–28, 2007.

[157] K. P. Rajurkar *et al.*, "Micro and nano machining by electro-physical and chemical processes," *CIRP Annals—Manufacturing Technology*, vol. 55, pp. 643–666, 2006.

[158] J. Shao *et al.*, "Generation of fully-covering hierarchical micro-/nano- structures by nanoimprinting and modified laser swelling," *Small*, vol. 10, pp. 2595–2601, 2014.

[159] H. Messmann, R. Knuchel, W. Baumler, A. Holstege, and J. Scholmerich, "Endoscopic fluorescence detection of dysplasia in patients with Barrett's esophagus, ulcerative colitis, or adenomatous polyps after 5-aminolevulinic acid-induced protoporphyrin IX sensitization," *Gastrointest. Endoscopy*, vol. 49, pp. 97–101, Jan. 1999.

[160] P. Mountney and G. Z. Yang, "Motion compensated SLAM for image guided surgery," *Med. Image Comput. Computat. Assist. Interv.*, vol. 13, pp. 496–504, 2010.

[161] H. G. Lee, M. K. Choi, and S. C. Lee, "Motion analysis for duplicate frame removal in wireless capsule endoscope," *Medical Imaging 2011: Image Processing*, vol. 7962, 2011.

[162] S. Yoshizaki, A. Serb, L. Yan, and T. G. Constandinou, "Octagonal CMOs image sensor with strobed RGB LED illumination for wireless capsule endoscopy," in *Proc. 2014 IEEE Int. Symp. Circuits Syst. (ISCAS)*, Melbourne, Australia, 2014, pp. 1857–1860.

[163] P. L. Lam and R. Gambari, "Advanced progress of microencapsulation technologies:

In vivo and *in vitro* models for studying oral and transdermal drug deliveries," *J. Control. Release*, vol. 178, pp. 25–45, Mar. 2014.

[164] M. V. Kiryukhin, "Active drug release systems: Current status, applications and perspectives," *Curr. Opin. Pharmacol.*, vol. 18C, pp. 69–75, Sep. 2014.

[165] S. Mura, J. Nicolas, and P. Couvreur, "Stimuli-responsive nanocarriers for drug delivery," *Nat. Mater.*, vol. 12, pp. 991–1003, Nov. 2013.

[166] E. Diller, J. Giltinan, G. Z. Lum, Z. Ye, and M. Sitti, "Six-degrees-of-freedom remote actuation of magnetic micro-robots," *Proc. Robot.: Sci. Syst.*, 2014.

[167] E. M. Purcell, "Life at low Reynolds number," *Amer. J. Phys.*, vol. 45, p. 3, 1977.

[168] G. Loget and A. Kuhn, "Electric field-induced chemical locomotion of conducting objects," *Nat. Commun.*, vol. 2, p. 535, Nov. 2011.

[169] W. Wang, L. A. Castro, M. Hoyos, and T. E. Mallouk, "Autonomous motion of metallic microrods propelled by ultrasound," *ACS Nano*, vol. 6, pp. 6122–6132, Jul. 2012.

[170] A. Sen, M. Ibele, Y. Hong, and D. Velegol, "Chemo and phototactic nano/microbots," *Faraday Discuss*, vol. 143, pp. 15–27, 2009, discussion 81–93.

[171] H. R. Jiang, N. Yoshinaga, and M. Sano, "Active motion of a Janus particle by self-thermophoresis in a defocused laser beam," *Phys. Rev. Lett.*, vol. 105, p. 268302, Dec. 2010.

[172] W. F. Paxton *et al.*, "Catalytic nanomotors: Autonomous movement of striped nanorods," *J. Am. Chem. Soc.*, vol. 126, pp. 13 424–13 431, Oct. 2004.

[173] D. Kagan, S. Balasubramanian, and J. Wang, "Chemically triggered swarming of gold microparticles," *Angew. Chem. Int. Ed. Engl.*, vol. 50, pp. 503–506, Jan. 2011.

[174] Y. Yoshizumi, Y. Date, K. Ohkubo, M. Yokokawa, and H. Suzuki, "Bimetallic micromotor autonomously movable in biofuels," in *Proc. 2013 IEEE 26th Int. Conf. Micro Electro Mechanical Syst. (MEMS)*, Taipei, Taiwan, 2013, pp. 540–543.

[175] M. E. Ibele, Y. Wang, T. R. Kline, T. E. Mallouk, and A. Sen, "Hydrazine fuels for bimetallic catalytic microfluidic pumping," *J. Amer. Chem. Soc.*, vol. 129, pp. 7762–7763, Jun. 2007.

[176] S. Sengupta *et al.*, "Self-powered enzyme micropumps," *Nat. Chem.*, vol. 6, pp. 415–422, May 2014.

[177] H. Choi *et al.*, "Electromagnetic actuation system for locomotive intravascular therapeutic micro-robot," in *Proc. 2014 5th IEEE RAS EMBS Int. Conf. Robot. Biomechatron.*, Sao Paulo, Brazil, 2014, pp. 831–834.

[178] F. Chen *et al.*, "In vivo tumor vasculature targeted PET/NIRF imaging with TRC105(Fab)-conjugated, dual-labeled mesoporous silica nanoparticles," *Mol. Pharm.*, vol. 11, pp. 4007–4014, Nov. 3, 2014.

[179] H. F. Wehrl *et al.*, "Preclinical and translational PET/MR imaging," *J. Nucl. Med.*, vol. 55, pp. 11S–18S, May 15, 2014.

[180] M. S. Judenhofer *et al.*, "Simultaneous PET-MRI: A new approach for functional and morphological imaging," *Nat. Med.*, vol. 14, pp. 459–465, Apr. 2008.

[181] C. H. Siu, T. J. C. Harris, J. Wang, and E. Wong, "Regulation of cell-cell adhesion during Dictyostelium development," *Seminars in Cell Develop. Biol.*, vol. 15, pp. 633–641, Dec. 2004.

[182] W. F. Loomis, "Cell signaling during development of Dictyostelium," *Dev. Biol.*, vol. 391, pp. 1–16, Jul. 1, 2014.

[183] G. Shaulsky and R. H. Kessin, "The cold war of the social amoebae," *Current Biology*, vol. 17, pp. R684–R692, Aug. 2007.

[184] S. T. Rutherford and B. L. Bassler, "Bacterial quorum sensing: Its role in virulence and possibilities for its control," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, Nov. 1, 2012.

[185] D. A. Robinton and G. Q. Daley, "The promise of induced pluripotent stem cells in research and therapy," *Nature*, vol. 481, pp. 295–305, Jan. 2012.

[186] P. A. Patah *et al.*, "Microbial contamination of hematopoietic progenitor cell products: Clinical outcome," *Bone Marrow Transplant*, vol. 40, pp. 365–368, Jun. 2007.

[187] M. Stormer *et al.*, "Bacterial safety of cell-based therapeutic preparations, focusing on haematopoietic progenitor cells," *Vox Sang*, vol. 106, pp. 285–296, May 2014.

[188] S. Patyar *et al.*, "Bacteria in cancer therapy: A novel experimental strategy," *J. Biomed. Sci.*, vol. 17, p. 21, 2010.

[189] M. Loeffler, G. Le'Negrate, M. Krajewska, and J. Reed, "Attenuated Salmonella engineered to produce human cytokine LIGHT inhibit tumor growth," *Proc. Nat. Acad. Sci.*, vol. 104, pp. 12 879–12 883, 2007.

[190] J. F. Schenck, "Safety of strong, static magnetic fields," *J. Magn. Reson. Imaging.*, vol. 12, pp. 2–19, Jul. 2000.

[191] I. C. Atkinson, L. Renteria, H. Burd, N. H. Pliskin, and K. R. Thulborn, "Safety of human MRI at static fields above the FDA 8 T guideline: Sodium imaging at 9.4 T does not affect vital signs or cognitive ability," *J. Magnet. Resonance Imaging*, vol. 26, pp. 1222–1227, 2007.

[192] M. V. Berry and A. K. Geim, "Of flying frogs and levitrons," *European Journal of Physics*, vol. 18, p. 307, 1997.

[193] Y. Kinouchi, H. Yamaguchi, and T. S. Tenforde, "Theoretical analysis of magnetic field interactions with aortic blood flow," *Bioelectromagnetics*, vol. 17, pp. 21–32, 1996.

[194] F. G. Shellock, E. Kanal, and T. B. Gilk, "Regarding the value reported for the term 'spatial gradient magnetic field' and how this information is applied to labeling of medical implants and devices," *Amer. J. Roentgenol.*, vol. 196, pp. 142–145, 2011.

[195] B. Huang, M. W. M. Law, and P. L. Khong, "Whole-body PET/CT scanning: Estimation of radiation dose and cancer risk 1," *Radiology*, vol. 251, pp. 166–174, 2009.

[196] *Guidance for Industry and FDA Staff-Information for Manufacturers Seeking Marketing Clearance of Diagnostic Ultrasound Systems and Transducers*, U.S. Food Drug Administration, 2012.

[197] A. K. Capulli *et al.*, "Approaching the *in vitro* clinical trial: Engineering organs on chips," *Lab Chip*, vol. 14, pp. 3181–3186, Sep. 7, 2014.

[198] E. Diller and M. Sitti, "Micro-scale mobile robotics," *Foundations and Trends in Robotics*, vol. 2, no. 3, pp. 143–259, 2013.

ABOUT THE AUTHORS

Metin Sitti (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1992 and 1994, respectively, and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999.

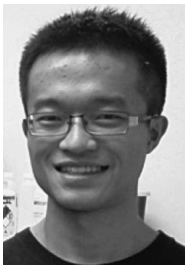
He was a research scientist at the University of California at Berkeley, Berkeley, CA, USA, during 1999–2002. Since 2002, he has been a professor in Department of Mechanical Engineering and Robotics Institute at Carnegie Mellon University, Pittsburgh, PA, USA. In 2014, he became a director in Max-Planck Institute for Intelligent Systems in Stuttgart, Germany. His research interests include physical intelligence, mobile micro-robots, bioinspired materials and miniature robots, medical milli/micro-robotics, and micro/nanomanipulation.

Dr. Sitti received the SPIE Nanoengineering Pioneer Award in 2011 and the National Science Foundation CAREER Award in 2005. He received the IEEE/ASME Best Mechatronics Paper Award in 2014, the Best Poster Award in the Adhesion Conference in 2014, the Best Paper Award in the IEEE/RSJ International Conference on Intelligent Robots and Systems in 2009 and 1998, respectively, the first prize in the World RoboCup micro-robotics Competition in 2012 and 2013, the Best Biomimetics Paper Award in the IEEE Robotics and Biomimetics Conference in 2004, and the Best Video Award in the IEEE Robotics and Automation Conference in 2002. He is the editor-in-chief of *Journal of Micro-Bio Robotics*.

Hakan Ceylan received the B.S. degree in molecular biology from Bilkent University, Ankara, Turkey, in 2010 and the Ph.D. degree in materials science and nanotechnology from National Nanotechnology Research Center affiliated to Bilkent University in August 2014.

Since September 2014, he has been a postdoctoral researcher in the Max Planck Institute of Intelligent Systems, Stuttgart, Germany. His research interests include self-organization, programmable matter, reconfigurable processes, and bioinspired adhesive interfaces. He is the author of one book chapter and seven research articles.

Dr. Ceylan's awards and fellowships include Lindau Nobel Laureate Meeting Fellowship (2011), undergraduate and graduate scholarships from the Scientific and Technological Council of Turkey (TUBITAK) in the periods of 2005–2010 and 2010–2014, respectively, and summer research fellowship from Marie Curie Research Institute, Oxted, UK, in 2009. He was a recipient of Ultratech/Cambridge NanoTech 2013 Best Paper Award.



Wenqi Hu (Student Member, IEEE) received the B.S. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2005, and the Ph.D. degree in electrical engineering from the University of Hawai'i at Manoā, Honolulu, HI, USA, in 2014.

He is currently working as a postdoctoral researcher at Max Planck Institute for Intelligent Systems, Stuttgart, Germany. His research interest is biomedical applications of micro-robots.

Dr. Hu was the recipient of the University of Hawai'i College of Engineering's 2014 Outstanding Ph.D. Research Award.

Joshua Giltinan (Student Member, IEEE) is currently pursuing the Ph.D. degree in Department of Mechanical Engineering at Carnegie Mellon University, Pittsburgh, USA.

In 2014, he joined the Physical Intelligence department at the Max-Planck Institute for Intelligent Systems, Stuttgart, Germany.

Mehmet Turan received the Diploma degree in electronics and telecommunication engineering from RWTH Aachen University, Germany in 2011. Since June 2014, he has been a Ph.D. student in the Max Planck Institute for Intelligent Systems, Stuttgart, Germany.

His research interests include computer vision in medical robotics, SLAM, and robot control.

Sehyuk Yim (Student Member, IEEE) received the B.Eng. degree in mechanical engineering, the B.Eng. degree in electronic engineering in 2007, and the M.S. degree in mechanical engineering in 2009, all from Sogang University, Seoul, Korea, and the Ph.D. degree in mechanical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in December 2012.

He is currently working as a postdoctoral researcher at Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. His research interests include the design, modeling, and control of bioinspired robots, soft robots for medical applications, modular robots, and programmable matter.

Eric Diller (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from the Case Western Reserve University, Cleveland, OH, USA, in 2009. He received the Ph.D. degree in mechanical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2013, where he continued as a postdoctoral researcher. Since 2014, he has been an Assistant Professor in the Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada.

Dr. Diller's research interests are in microscale robotics and bioinspired novel locomotion systems, fabrication and control relating to remote actuation of microscale devices using magnetic fields, microscale robotic manipulation, smart materials, and swimming at low Reynolds number.

CONCLUSION

This dissertation introduces many novel ideas for endoscopic capsule robot localization and mapping task, making use of deep learning, sensor fusion and direct SLAM techniques. To summarize, following frameworks are proposed:

- A dense and direct medical SLAM approach which makes use of GPU accelerated non-rigid frame-to-model fusion, joint volumetric-photometric pose estimation and dense model-to-model loop closure techniques;
- A fully dense, non-rigidly deformable, strictly real-time, intraoperative map fusion approach for actively controlled endoscopic capsule robot applications, which combines benefits of magnetic and vision-based localization, with non-rigid deformations based frame-to-model map fusion;
- A supervised deep monocular visual odometry (VO) method for endoscopic capsule robots based on recurrent convolutional neural networks (RCNNs);
- An unsupervised deep localization and depth estimation approach for endoscopic capsule robots consisting of two simultaneously trained sub-networks, the first one assigned for depth estimation via encoder-decoder strategy, and the second assigned to regress the camera pose in 6-DoF;
- A sequence-to-sequence deep sensor fusion approach which does not need any spatial or temporal synchronization between sensors;
- A novel multi-sensor fusion algorithm on switching state space models with particle filtering using robot dynamics modelled by recurrent neural networks (RNNs), which can handle sensor faults and non-linear motion models;
- A comprehensive medical 3D reconstruction method, which is built in a modular fashion including preprocessing, keyframe selection, sparse-then-dense alignment-based pose estimation, bundle fusion, and shading-based 3D reconstruction;
- A comprehensive review of the current advances in biomedical untethered mobile milli- and microrobots with an emphasis on the potential impacts of such devices in the near future and existing and emerging challenges associated with medical operations performed via such minituarized robotic technologies.

The presented methods are showing high accuracy camera pose and 3D mapping performances on the qualitative and quantitative analyses performed on deformable porcine stomachs and realistic surgical EsophagoGastroDuodenoscopy simulator. In future, in-vivo testings on real human patients are required to validate the accuracy and robustness of the methods in real GI tract and under real medical operation

conditions. We also intend to extend our work into stereo capsule endoscopy applications to achieve even more accurate localization and mapping results. Moreover, we intend to incorporate a robot operating system platform (ROS), disease/lesions detection, segmentation algorithms and nearest frontier based exploration capabilities into the platform to transform our actively controllable endoscopic capsule robot into an intelligent medical robot system which autonomously explores and detects lesions inside the GI tract and performs drug delivery and biopsy-like interventions with submillimeter precision.

Despite the success of the proposed ideas in this dissertation, there are still several open challenges for deep learning based endoscopic capsule robot applications such as:

- Large datasets of labeled medical datasets are not generally available due to privacy issues and underrepresentation of rare conditions; e.g. diseases and abnormalities. Synthetically-generated medical data with an accurate forward model for the capsule imaging system and anatomically-realistic model of the organ could be used for the training purposes.
- A spatial and temporal control of the capsule illumination source would lead to better robot localization, mapping, lesion detection and recognition. The projection of a known and spatially and temporally varying texture (and color), would facilitate the detection of lesions (which respond differently to different wavelengths) as would facilitate the extraction of image data given the projection of texture. Varying textures for different areas of the visual field may also facilitate detection and estimation of shape (photometric shape and motion).
- Low resolution and low camera framerates (3-5 fps) of state-of-the art capsule endoscopy videos are still limiting factors for computer vision related tasks such as disease detection, topography estimation and visual odometry. Thus, to enhance the resolution of the capsule endoscopy videos, a super-resolution generative adversarial network might be trained. Moreover, low frame rate endoscopic videos could be interpolated via inter-frame generative adversarial networks to upscale the framerate.
- Detection of lesions in the small bowel is specially difficult given its length, homogeneous texture, and absence of visual references. In addition, the medical alternatives to performing small bowel exams are very difficult and complex. The use of capsule endoscopy is therefore specially important. Capsule-based detection and localization of lesions is both more difficult and important in the case of the small bowel. New robust methods and techniques are required, which can take advantage of actively controlled capsules.
- Most capsule endoscopy manufacturers have multi-camera versions of capsules. Motion control of such multi-cam capsules can advantageously use the information provided by the several cameras. Such a multi-camera based visual feedback would increase the control accuracy of the endoscopic capsule robot.