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 work-arounds 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

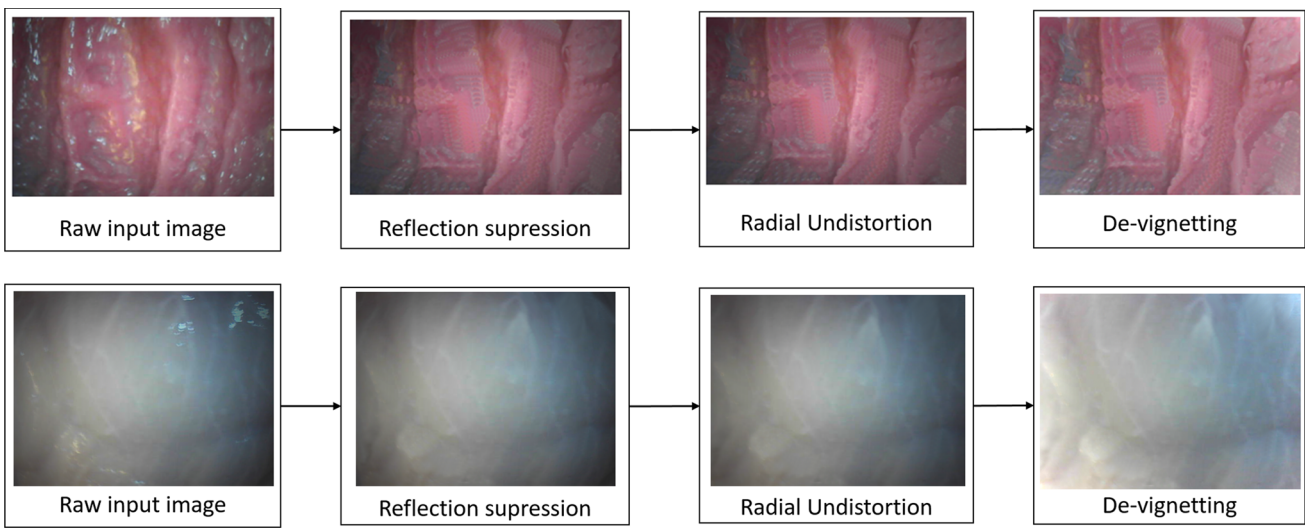


Fig. 1 Preprocessing pipeline: reflection removal, radial undistortion, de-vignetting

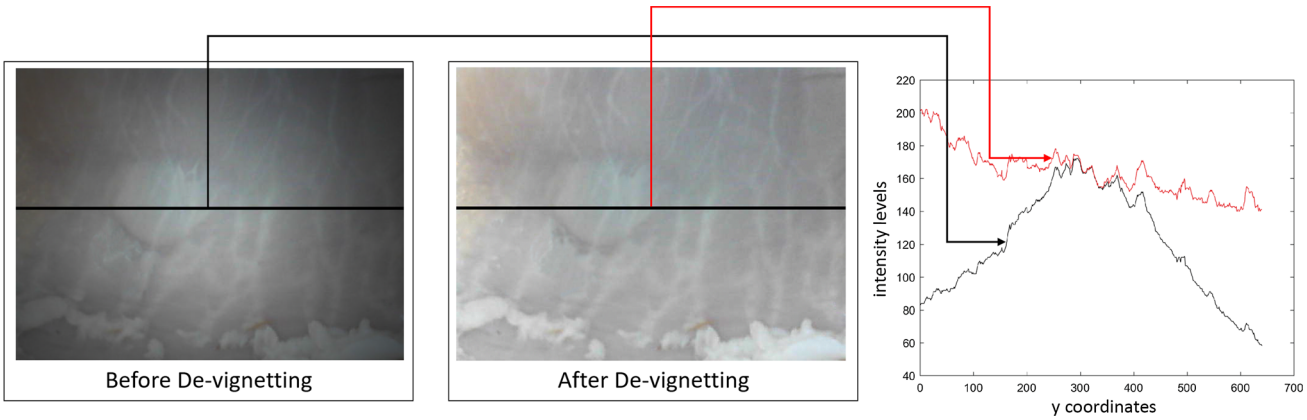


Fig. 2 Demonstration of the de-vignetting process

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

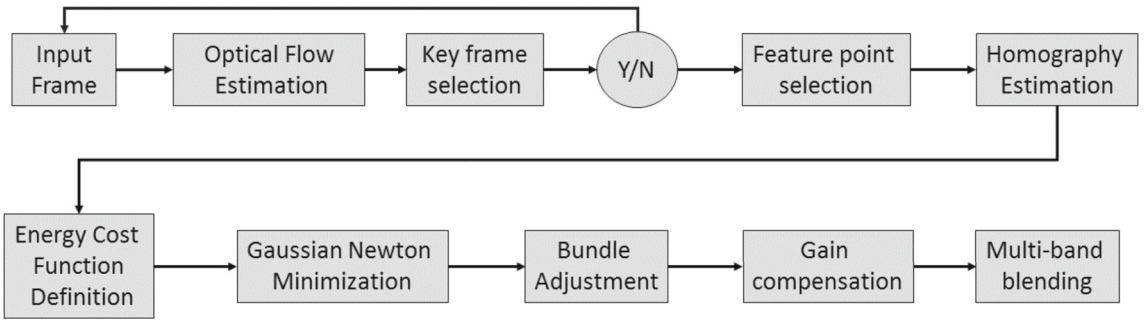


Fig. 3 Image stitching flowchart

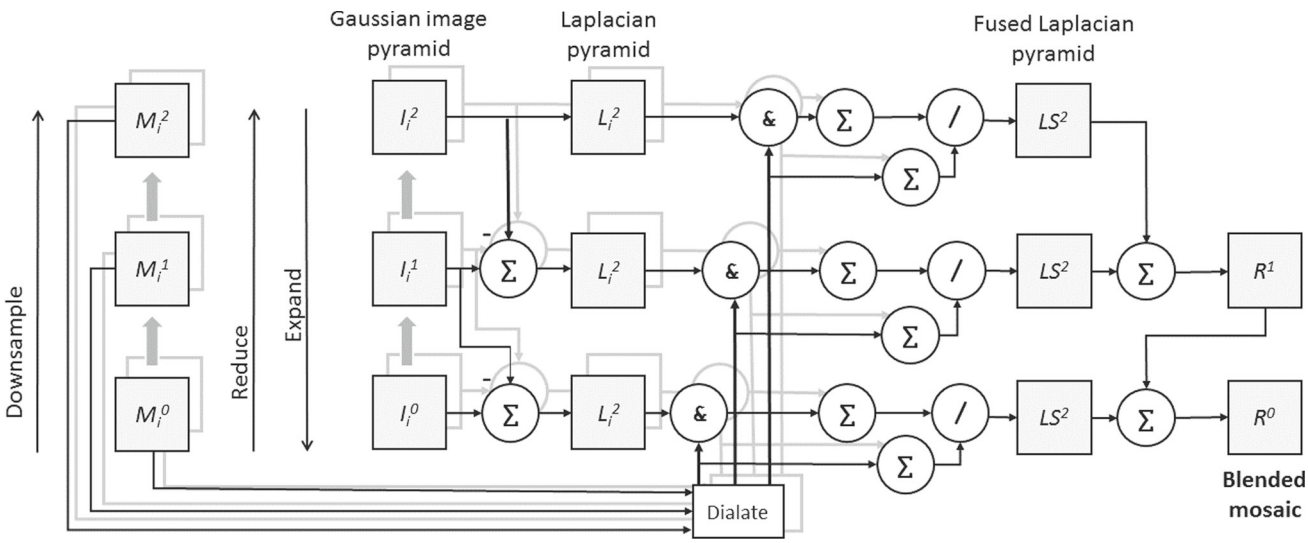


Fig. 4 Multi-band blending flowchart

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

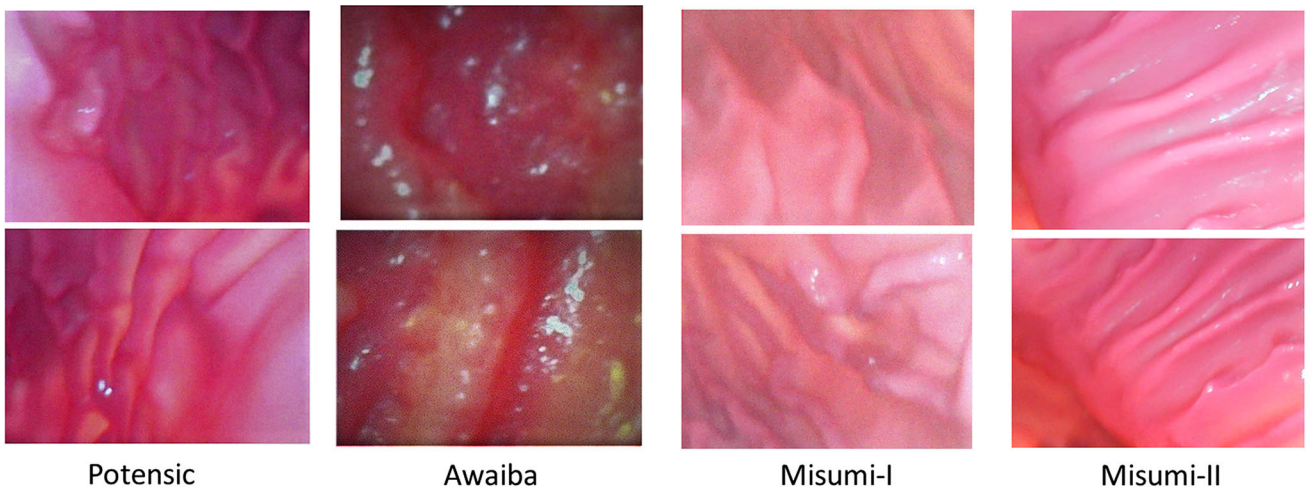


Fig. 6 Non-rigid esophagus gastroduodenoscopy simulator dataset overview for different endoscopic cameras

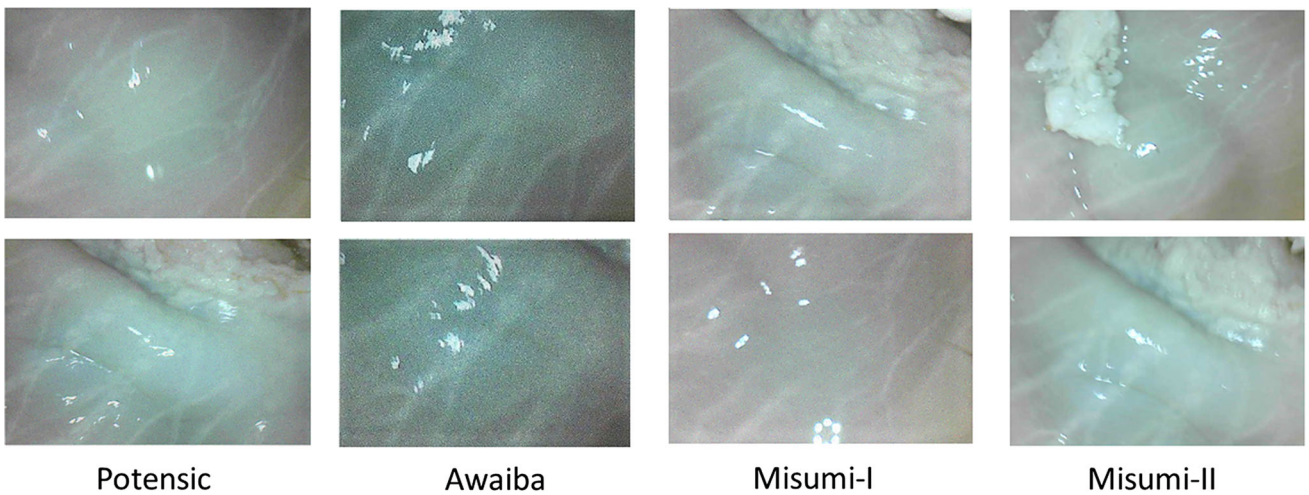


Fig. 7 Real pig stomach dataset overview for different endoscopic cameras

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

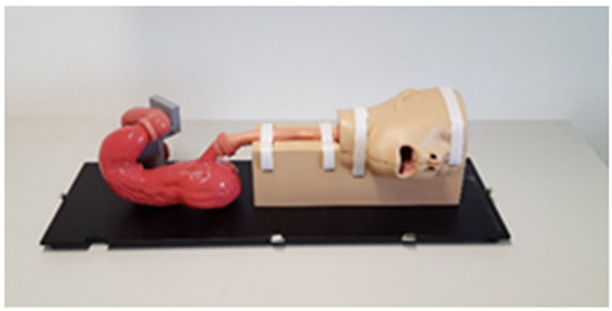
Resolution	1280 × 720 pixel
Diameter	8.8 mm
Pixel size	$10.0 \times 10.0\,\mu m^2$
Pixel depth	10 bit
Frame rate	30 fps

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).

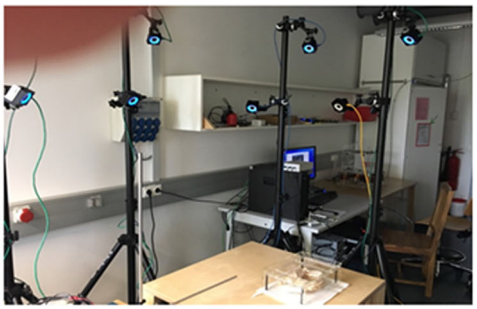
4.2 Trajectory estimation

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

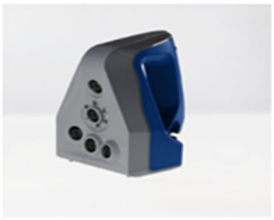


EGD surgical simulator LM-103

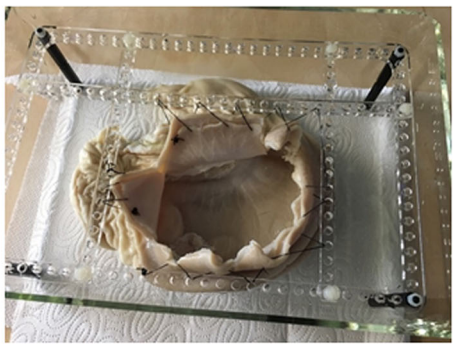


Optitrack Prime 13 Tracking System

Magnetically actuated soft capsule endoscopes



Artec 3D Space Spider



Real pig stomach

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

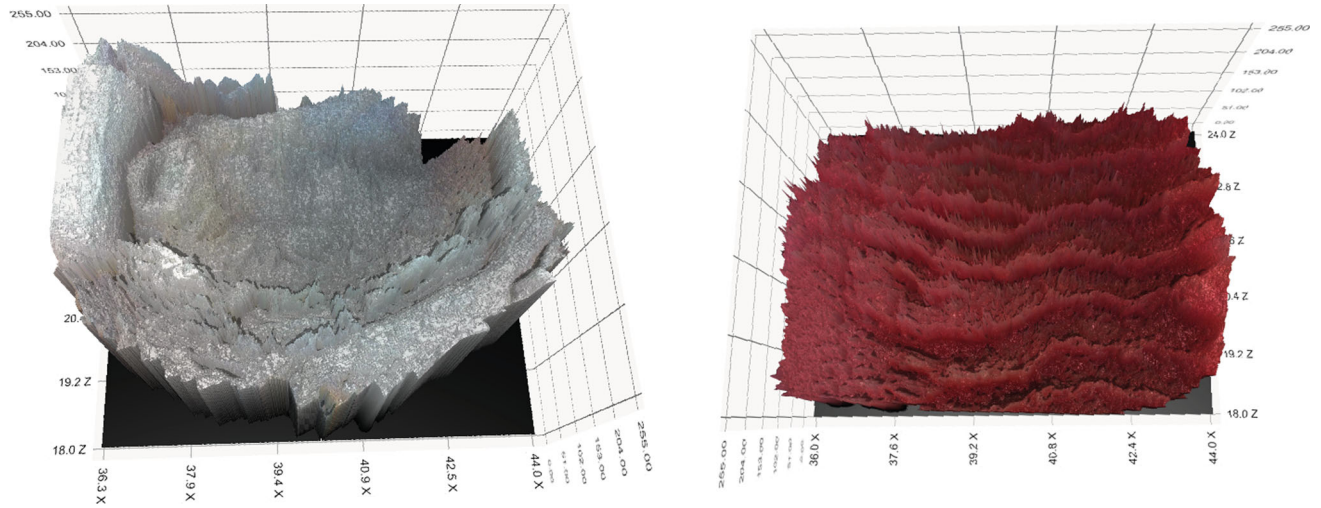


Fig. 10 Qualitative 3D reconstructed map results for different cameras [(real pig stomach (left), synthetic human stomach (right)]

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.