3D Reconstruction of the Human Colon from Capsule Endoscope Video*

Pål Anders Floor*, Ivar Farup and Marius Pedersen

Department of Computer Science, Norwegian University of Science and Technology (NTNU), Teknologivegen 22, 2815 Gjøvik, Norway

Abstract

In this paper we investigate the possibility of constructing 3D models of longer sections of the human colon using image sequences obtained from wireless capsule endoscope (WCE) video to provide enhanced viewing for gastroenterologists. As images from WCE contain severe distortions and artifacts non-ideal for 3D reconstruction algorithms, the problem is difficult to attack. However, recent developments of virtual graphics-based models of human gastrointestinal system, where most of the distortions and artifacts can be enabled or disabled, makes it possible to determine how each factor disturbs such algorithms individually. In this paper we disable distortions and artifacts in order to determine if longer sections of the human intestinal environment is at all feasible to reconstruct. Though simulation we show that this is possible using structure from motion and simultaneous localization and mapping (SLAM).

Keywords

3D reconstruction, capsule endoscopy, structure from motion, SLAM.

1. Introduction

Severe diseases in the gastrointestinal (GI) system like Chron's disease, inflammatory bowel disease, and cancer, are reducing many peoples quality of life. One way to detect such diseases at an early stage, making them more likely to combat, is to make screening of the GI system a common procedure beyond a certain age. However, fear of pain and difficulties caused by endoscopy is a major factor limiting the number of people screening themselves without clear symptoms [1].

The wireless capsule endoscope (WCE) [2], which is a pill-sized capsule that the patient swallows, is a good alternative for preventive screening, as it avoids the above mentioned problem and is capable of reaching all parts of the GI system. The WCE carries one- or several cameras on board, recording video while travelling through the GI system. However, current standard WCE's have significantly lower resolution and frame rate than typical endoscopes, and the images contain more severe noise and distortions. Further, the video is usually over eight hours long, making it challenging for gastroenterologists to detect pathologies in the intestinal wall. With increasing demand on intestinal screening, tools that make the gastroenterologists workload less demanding, and thereby reduces time-use per patient, are needed.

The 11th Colour and Visual Computing Symposium, September 8–9, 2022, Gjøvik, Norway

* Funding was provided by the Research Council of Norway under the project CAPSULE no. 300031.

*Corresponding author.

paal.anders.floor@ntnu.no (P. A. Floor)

One method that can help gastroenterologists is a 3D model enhancing pathologies in the intestinal wall, making them easier to detect. A 3D model may also prove useful for planning of treatment. The inspiration for this approach comes from the positive feedback of using 3D reconstruction in gastrointestinal endoscopy [3]. Currently, 3D models are obtained through CT-scanning, which is expensive and may expose the patient to unnecessary radiation. Therefore, we will investigate the construction of 3D models based solely on WCE images.

There are at least two methods that may be applied in order to reconstruct the 3D structure of a scene based on WCE images: 1) *Direct methods*, like *shape from shading* (SFS), which recovers 3D structure based on geometric reasoning on how light is reflected of relevant surfaces [4], here the GI wall. 2) *Feature-based methods* like *structure from motion* (SfM), which recovers 3D shapes from features captured in multiple views of the same scene [5]. SfS can reconstruct 3D shapes from only one image, while SfM needs at least two images.

With many images of the same (rigid) scene available, SfM can provide accurate 3D reconstruction. However, this is not necessarily easy to obtain for WCE from WCE images for the following reasons: i) Sometimes only one image is available due to rapid movement of the WCE, or debris in the intestine. ii) SfM assumes rigid motion, which is sometimes violated due to muscle contractions and *peristalsis*. iii) Sometimes the WCE position does not change enough from frame to frame to avoid degeneracies. In case i) single image techniques, like SfS, have to be applied [6]. In case ii) one can apply *non-rigid* SfM (NR-SfM) [7, 8] taking non-rigid scene movement into account. In case iii) a *simultaneous localization and mapping* (SLAM) [9, 10] approach may be applied.

SLAM uses the fact that both camera position and 3D structure are obtained from SfM, and applies SfM locally among so-called *key frames*, which are frames with significantly different poses. Therefore, SLAM is potentially able to detect and ignore frames that may lead to degeneracies.

Another problem is that WCE images are are highly corrupted. Examples are debris in the intestinal fluids, specular reflections, motion blur, heavy lens distortion, chromatic aberrations, compression artefact etc. All of these factors makes it hard, if at all possible, to design algorithms for accurate 3D reconstruction as it is difficult to single out how each of these corruptions affect the reconstruction individually. However, with recent developments of virtual graphics-based models of human GI system, where most of the distortion and artifacts can be enabled and disabled, it is possible to dissect the problem and determine how each factor disturbs the reconstruction individually. One such model is VR-CAPS [11], which is a realistic looking artificial GI system built from CT scans of humans, where also most corruptions in the WCE imaging process is modelled. Further, their GI-model can easily be exported and dissected in 3D modelling applications like *blender*[1], thereby providing a *ground truth* for evaluation of 3D reconstruction algorithms, something which is hard to obtain for real WCE.

In this paper we conduct a feasibility study using SfM and SLAM for 3D reconstruction of longer sections of human colon in an ideal situation where most of the distortions mentioned above are turned off. We will use typical WCE image resolution and frame-rate. This enables us to conclude if 3D reconstruction using a feature based approach is at all possible for the unusual and repetitive geometry of a typical colon. If the conclusion is negative, there is no point in

[1] https://www.blender.org/

pursuing this problem further. We will first investigate SfM to gain basic knowledge, then use this knowledge to investigate a SLAM approach, named *ORB-SLAM* [10], which is a fast and accurate approach for monocular cameras.

In Section 2 the problem formulation is given and the existing methods we apply for our experiments are described. In Section 3, which contains the novel results of this paper, we experiment with SfM and ORB-SLAM to obtain 3D models of human colon segments generated in VR-CAPS. Summary, conclusion and future research ideas are given in Section 4.

2. Problem Formulation and Methods

Direct methods usually rely on accurate radiometric information, whereas feature based methods rely on image features. As the WCE is continuously adapting its camera response, accurate radiometric information is hard to obtain over image sequences. Therefore we consider feature based approaches here.

We will assume that lens distortion and other non-linearities have been compensated for so that we have a *pinhole model*. We will also assume that typical distortions seen in pillcam images, like specs on lens, motion blur etc., are taken care of through pre-processing.

Generally, we assume that the image capturing process is some mapping between 3D projective space $\mathbb{P}^3$ and 2D projective (image) plane $\mathbb{P}^2$. Points in space are described in homogenous *world coordinates* as $\mathbf{X} = [X, Y, Z, W]^T$ and image point are in homogenous *image coordinates* $\mathbf{x} = [x, y, w]^T$ ($W, w \in \mathbb{R}^+$ are some unspecified scaling factors) [5, p. 7]. For 3D points in a *point cloud*, we denote the i'th point as $\mathbf{X}_i$, $i = 1, \cdots, N$, and its image $\mathbf{x}_i$. With a pinhole camera model the relation between a point in world coordinates and image coordinates is a mapping $P : \mathbb{P}^3 \to \mathbb{P}^2$. Then, for M views (images) of a given point, $\mathbf{X}_i$, in the point cloud, the imaging process of the j'th view is given by [5, p. 154]

$$\mathbf{x}_i^j = P^j \mathbf{X}_i,\ j = 1, \cdots, M,\ i = 1, \cdots, N, \tag{1}$$

where P^j is the 3×4 *camera matrix* for the j'th view given by [5, p. 156]

$$P^j = K[R^j|\mathbf{t}^j]. \tag{2}$$

R^j is a 3×3 rotation matrix and $\mathbf{t}^j$ is a 3×1 translation vector, both in $\mathbb{P}^3$. With WCE, the same camera captures all views, and so the *calibration-* or *intrinsic matrix*, K, is the same for all views, given by [5, p. 156]

$$K = \begin{bmatrix} fm_x & s & p_x m_x \\ 0 & fm_y & p_y m_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where f is the *focal length*, p_x, p_y is the *principal point*, s is the *skew* and m_i is the number of pixels per unit distance. The m_i-factors in K makes Eq. (2) a transformation from world coordinates to *pixel coordinates*. Skew is normally zero for pillcams, therefore we set $s = 0$ in what follows. The other parameters can be found through a *calibration* procedure [5, p.226].

We assume that the first view of any image sequence is taken by the camera when located at the world origin, i.e., $P^1 = K[I_{3\times3}|\mathbf{0}]$, with $I_{3\times3}$ the 3×3 identity matrix. That is, camera coordinates of the first view in the sequence are equivalent to world coordinates.

2.1. Structure from Motion (SfM)

SfM recovers both 3D structure and individual camera poses. For two, three, or four views one can solve the SfM problem using *tensors* named *fundamental matrix* (FM), *trifocal tensor* and *quadrifocal tensor*, respectively [5], which generate multi-linear relationships among the coordinates of image measurements, providing closed form mathematical relations in terms of camera matrices. For $M > 4$ views one has to deal with the problem through *bundle adjustment* (BA).

2.1.1. Two views and fundamental matrix (FM)

Take Eq. (1) with $M = 2$. Then any two points $\mathbf{x}_i^1, \mathbf{x}_i^2$, being the images of $\mathbf{X}_i$ in the two views, have to be related by the *epipolar constraint* [5, p.245]

$$(\mathbf{x}_i^2)^T F \mathbf{x}_i^1 = 0, \ \forall i, \tag{4}$$

with F, the FM, a 3×3 rank 2 matrix given by $F = [\mathbf{e}^2]_\times P^2 (P^1)^+$ [5, p.244]. Here $\mathbf{e}^2$ is the *epipole*, i.e., the image of the camera center of view 1, and $(P^1)^+$ is the Moore-Penrose pseudoinverse of P^1. $A = [\mathbf{e}^2]_\times$ is a skew-symmetric matrix where $a_{21} = e_3^2$, $a_{31} = -e_2^2$, and $a_{32} = e_1^2$.

F can be numerically estimated from common features in two images. Typically, SIFT, SURF, Eigen- or ORB features are generated in the two images and matches between them are searched. With $n \geq 8$ such matches the *normalized 8-point algorithm* can estimate F [5, p. 282]. With significant noise in the image, *outliers* can be problematic. These can be dealt with by the RANSAC [12] algorithm. With F estimated, assuming that the camera center of the first view is at the world origin, the two camera matrices are found by $P^1 = K[I|\mathbf{0}]$ and $P^2 = [[\mathbf{e}^2]_\times F|\mathbf{e}^2]$ [5, p.256].

With P^1, P^2 determined, one can find 3D point $\mathbf{X}_i$ for the correspondence $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$, satisfying the constraint (4), by a *triangulation method* [5, p. 311]

$$\mathbf{X}_i = \tau(\mathbf{x}_i^1, \mathbf{x}_i^2, P^1, P^2). \tag{5}$$

A common method is to use the fact that $\mathbf{x}_i^j \times \mathbf{x}_i^j = \mathbf{x}_i^j \times P^j \mathbf{X}_i = 0$. For two corresponding points $j = 1, 2$, this generates four linearly independent equations contained in matrix B. To find $\mathbf{X}_i$ one solves $B\mathbf{X}_i = 0$ numerically (see [5, pp. 312-313]). Typically, one minimizes the *reprojection error* [5, p. 314], $C(\mathbf{x}_i^1, \mathbf{x}_i^2) = d(\mathbf{x}_i^1, \hat{\mathbf{x}}_i^1)^2 + d(\mathbf{x}_i^2, \hat{\mathbf{x}}_i^2)^2$, subject to (4), with $d(\cdot, \cdot)$ some distance measure.

With K and F known it is shown in [5, p. 272-273] that the 3D scene can be reconstructed up to a *similarity transform*, i.e, a Euclidean reconstruction with an unknown scaling factor. The exception is the *degenerate case* which can occur when the camera centers and $\mathbf{X}_i$ are co-linear, or in a practical noisy case, close to co-linear. Also, under pure rotation about the camera center, the degenerate case $F = 0$ occur. For WCE, degeneracies can occur due to both of these cases. With K known, one can estimate the *Essential Matrix* (EM) $E = K^T F K$ [5, p. 257], instead of the FM, which is simpler to compute.

To obtain metric reconstruction one will need additional information about the known length of some object in the scene, which is hard to obtain in the GI-system. One effort dealing with this issue is [13].

2.1.2. Multiple Views and Bundle Adjustment (BA)

For three and four views the trifocal- and quadrifocal tensors provide relations in a similar way as the FM did for two. However, the relations are more general. One example is the possibility of *transfer*. That is, with a point correspondence between two views, the point in the third (or forth) will be determined. For $M > 4$ views, the problem has to be dealt with numerically through bundle adjustment (BA), which is a minimization problem on the form [5, p. 434]

$$\min_{\hat{P}^j, \hat{\mathbf{X}}_i} \sum_{i,j} d\big(\hat{P}^j \hat{\mathbf{X}}_i, \mathbf{x}_i^j\big),\ j = 1, \cdots, M\ i = 1, \cdots, N, \tag{6}$$

with $d(\cdot, \cdot)$, some distance measure, typically Euclidean norm. That is, BA is the reprojection error over all views and 3D points. BA needs a good initial estimate, $\hat{P}^j, \hat{\mathbf{X}}_i$, of camera poses and 3D points, which is typically obtained by computing the FM (or trifocal tensor) sequentially over pairs (or triplets) of neighboring images until all views in the sequence are covered [5, p.453].

The main problem with BA is that it is very costly to compute for large M [5, p. 435]. This problem is addressed by some SLAM algorithms (like ORB-SLAM) by using BA locally over sub-sets of key frames.

2.2. ORB-SLAM

Performing 3D reconstruction on hundreds or even thousands of images, can be necessary for WCE video streams. Then a pure SfM approach is inconvenient both due to computational complexity and the difficulty of keeping track of which features are visible in a given view. For this a SLAM approach is more convenient. As WCE video is a sequences of monoculus images, we consider an approach known to be efficient for that case, namely *ORB-SLAM* [10].

ORB-SLAM performs SfM (using FM and BA) locally among key frames, which can be seen as structures (or objects) connected in a *co-visibility graph*. That is, a weighted graph where each node is a key frame with all relevant information included (like number of features, their *strength*, and all necessary adjacency information). There are edges among key-frames with common features, where the weight corresponds to the number of features they share. The local computation of camera poses and 3D geometry greatly reduces computational cost. A global optimization is also performed to optimize the position of camera poses. ORB-features, are used throughout as they are significantly faster to compute than SIFT or SURF features.

ORB-SLAM is done in three steps in addition to an initialization procedure. We provide a brief summary here and refer to [10] for details.

0) Initialization: One out of two methods are chosen based on the scene in question: i) A homography if the scene is plane, or if the parallax is low. ii) A FM if the scene is not plane and with sufficient parallax. With K known, the EM, $E = K^T F K$, is estimated. Solutions are chosen based on [5, p. 257-260]. The choice between the two cases are done automatically using a *heuristic* [10, p.1151]. A detection of low parallax case is also included and will refuse the initialization as this leads to a bad reconstruction.

1) Tracking: Localizes the pose of each frame w.r.t. the first view, which is assumed to be at the world origin ($P^1 = K[I|0]$), by matching ORB-features. It also decides if a given frame

should be inserted as a key frame in the co-visibility graph. The poses are then optimized using BA (6) over the P^i's only. If tracking is lost a *place recognition module* is used in a *global re-localization* procedure.

2) Local Mapping: Processes new key frames and performs local BA to obtain a sparse 3D reconstruction in the surroundings of the relevant pose/frame. New correspondences for unmatched ORB-features are searched in key frames directly connected in co-visibility graph to triangulate new 3D points. If a key frame is found to be redundant, i.e., if it does not change enough compared to other key frames, or if it lacks high quality point matches, it is discarded.

3) Loop Closure: With every new key frame the algorithm searches for loops (i.e., when the camera re-visits a previous part of the scene). When a loop is detected it is possible to estimate *drifts* in the data, like drift in scale and position. This is the essential step to remove or minimize such errors.

2.3. VR-CAPS

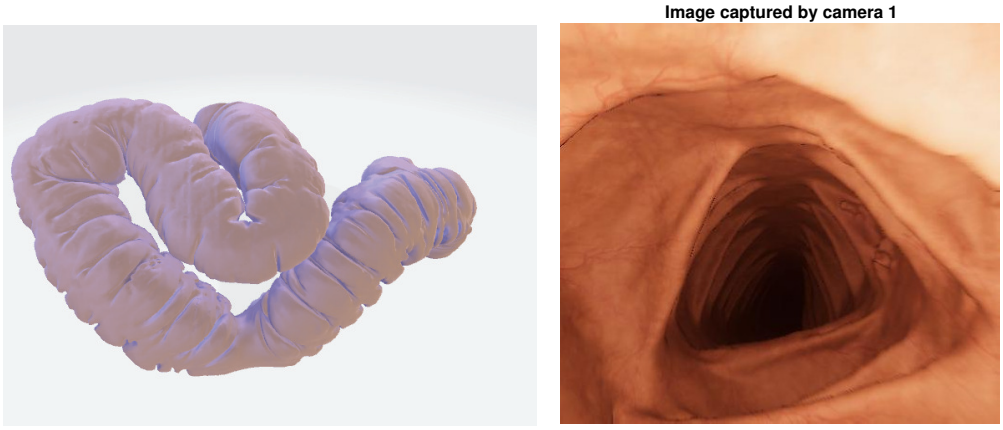
VR-CAPS is a virtual environment for WCE [11] which is publicly available in github[2]. The environment is based on *Unity*, which is a game-platform developed by *Unity Technologies*[3]. The environment simulates a range of organ types, capsule endoscopy designs, normal and abnormal tissue conditions as well as many other features detailed in [11]. It is also possible to emulate non-rigid movements like peristalsis. Therefore, VR-CAPS enables testing of medical imaging algorithms both for current and future WCE designs.

The standard setup in VR-CAPS is a virtual colon model which is built from CT scans of a real human colon and covered with realistic looking textures. A section of this colon is depicted in Fig. 1(a) and an example image captured by the WCE inside this segment is depicted in Fig. 1(b). Many pillcam models can be built, but the default is a standard-sized pill with one camera and a spot-light with conical beam, emulating several point-lights surrounding the lens often seen in standard WCE's. We will use the standard setup for our experiments.

3. Simulation setup and Experiments

We run the VR-CAPS simulator through several subsets of the colon-segment shown in Fig. 1. These subsets can be seen in Figs. 2(a), 4(a), 5(a) and serve as ground truth for the example 3D reconstructions. We set image size to 512×512, framerate to 20 fps, focal length $fm_i = 163$ and principle point $p_x m_x = p_y m_y = 163$ (in pixel units). Further, all distortion effects are disabled as default. However, we will enable some distortions in turn to evaluate the impact on the reconstruction. The WCE is controlled in VR-CAPS by key buttons and mouse. A steady movement is difficult to obtain, therefore the resulting WCE trajectory becomes irregular and ragged, especially through sharp bends. However, this movement appears quite similar to that of a real WCE, and will therefore test the ORB-SLAM's ability to cope with quite a realistic movement.

[2]https://github.com/CapsuleEndoscope/VirtualCapsuleEndoscopy (31/10-21)
[3]https://unity.com/



(a) (b)

Figure 1: (a) Section of colon from VR-CAPS applied for 3D reconstruction in this paper. (b) Example image taken by WCE inside the colon section in (a).

We first consider SfM on a short sequence of images to gain insight into what distortions the algorithm is sensitive to. This knowledge will be useful in deciding suitable pre processing for ORB-SLAM.

We evaluate the reconstruction mainly trough geometric inspection, visually comparing the resulting 3D point cloud models to the ground truth in Figs. 2(a), 4(a) and 5(a). For more exact evaluation, a numerical comparison to ground truth is needed, which can be obtained by computing the average distance between all reconstructed points and ground truth. However, this is not straight forward to obtain as one has to exclude all parts of the colon from the ground truth not captured by the camera over long image sequences, as well as triangulate the reconstructed point cloud in an optimal way. This is currently work in progress.

3.1. Structure from motion

We consider two cases: 2-view and 6-view SfM. Since K is known, we estimate the EM, $E = K^T F K$. SURF features are used to estimate E (and thereby P^i) and 3D points, whereas Eigen features are used to compute dense point clouds once E is known. For $M = 6$ views, an initial reconstruction is made by sequentially computing the EM for pairs of consecutive frames (as in Algorithm 18.3 in [5, p. 453]) followed by BA. All relevant computation and estimation methods for our purposes are found in Matlab's *computer vision toolbox*[4].

Based on experimentation on images obtained from VR-CAPS we concluded that the following pre-processing is needed: One has to remove specs on lens and specular reflections as they tend to confuse the feature detection algorithm. Motion blur causes similar problems, particularly in conjunction with rapid rotations and panning caused by rapid movements of the WCE from image to image. Lens distortion makes the assumption of pinhole camera fail, and therefore leads

[4]https://se.mathworks.com/products/computer-vision.html (10/11-21)

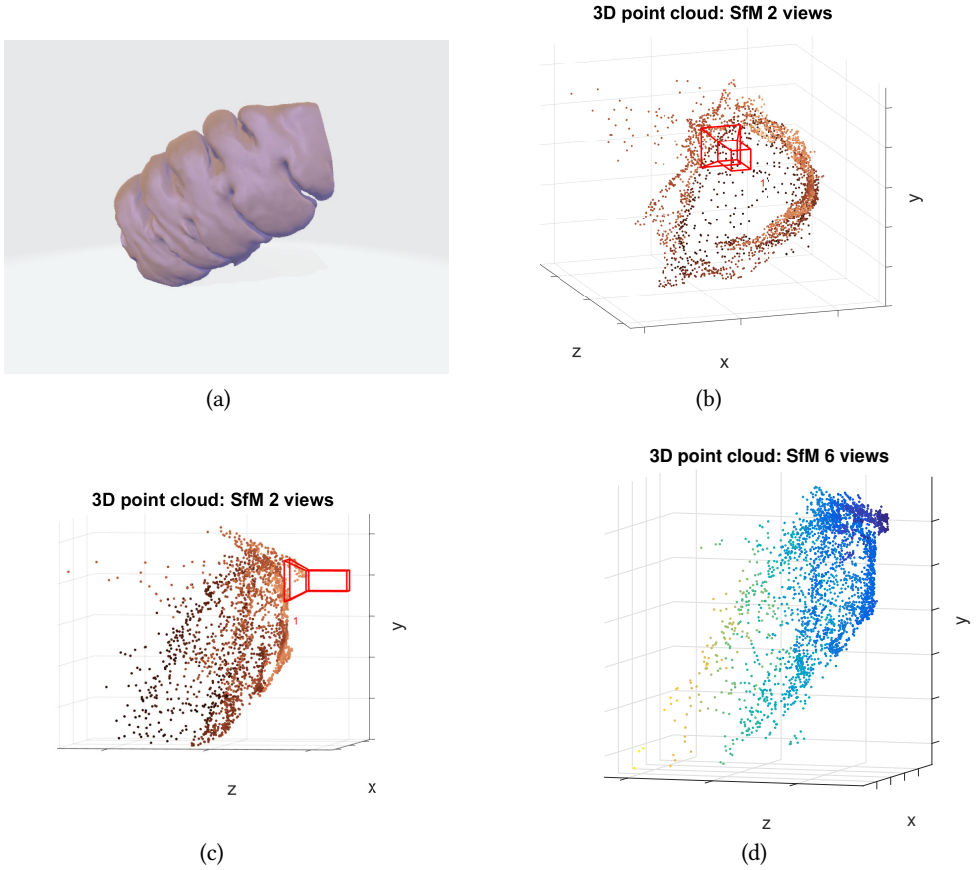


Figure 2: Structure from Motion (SfM) 3D reconstruction (a) Relevant colon segment (ground truth) (b) Reconstruction from 2 views seen from the front. (c) Reconstruction from 2 views seen from the side. (d) Reconstruction from 6 views seen from the side.

to very sparse and inaccurate 3D point clouds. As the WCE has a spotlight source, lighting will vary hugely across the image. Due to dim lighting, particularly in fields imaging deeper parts, contrast enhancement is essential to detect stable features. Due to the variation in brightness we applied *adaptive histogram equalization*.

The colon segment we aim to reconstruct is shown in Fig. 2(a). The image seen by view 1 is depicted in Fig. 1(b). This view is the reference for the 3D computation. The 2-view reconstruction is depicted in real color from the front and side in Figs. 2(b) and 2(c) respectively, with the position of camera 1 included. The 6-view reconstruction is shown in Fig. 2(d) from the side. The reconstruction is quite convincing, even with two images, but clearly more noisy than the 6-view case. The 6-view case is more restrictive and accurate as it rules out some outliers. Therefore, the point cloud may be somewhat less dense. Note in particular that the overall "cylindrical" geometry of the colon in Fig. 2(a) is reconstructed quite well, indicating that the algorithm eliminates perspective distortion.

3.2. ORB-SLAM

For longer segments than the one in Fig. 2(a) a large number of views may be necessary. Then ORB-SLAM is needed.

Algorithm for Densification: As ORB-SLAM is optimized for fast computation and and accurate localization, it produces a sparse point cloud only containing those 3D points of high accuracy needed to optimize camera localization. To make a denser point cloud for the purpose of GI inspection, we use the camera poses and co-visibility graph obtained by ORB-SLAM, then traverse the graph computing dense SfM over sub-sets of key-frames as detailed in Algorithm 1 below.

Algorithm 1. Densification of ORB-SLAM point cloud
Input: *i) Tracking data from ORB-SLAM, $P^i = K[R_i|\mathbf{t}_i]$, for all key-frames $i = 1, \cdots, N_{KF}$. ii) Co-visibility graph of key-frame objects.*
Initialization: *i) Point cloud array ii) Max number of views, M_V, used in dense 3D reconstruction*
Algorithm:
for $i = 1$ to N_{KF}
i) Determine number of key-frames, N_{CV}, with strong co-visible features shared with key-frame i for frames $j > i$
ii) if $N_{CV} < M_V$, set M_V to N_{CV}
iii) if $M_V = 0$, set point cloud to zero and jump to i) for next key frame
iv) else Perform M_V-view SfM (as in Section 2) with dense features, given $P^j, j = i, \cdots, i + M_V - 1$, with key frame i as reference view
v) Rule out degeneracies: If unproportionately large values exist in point cloud, set it to zero and jump to i) for next key frame
vi) Denoise point cloud and store in point cloud array
end
vii) Concatenate all point clouds using available position data, P^i

Simulation Setup: In initialization step 0) (see section 2.2) we force the algorithm to choose a FM model as plane scenes never occur in the GI-system. Due to the WCE movement, initialization may fail due to low parallax. If the initialization is rejected, we skip to the next frame and re-start the algorithm until the initialization succeeds. Loop closure (Step 3) should be disabled as loops never occur when the WCE travels through the GI system. The repetitive geometrical structure of the colon also tend to confuse the ORB-SLAM algorithm, mis-interpreting these for being potential loop closure candidates. Without loop closure one has to expect inaccuracies as both scale and position will drift, particularly over longer segments. In sharp bends of the colon this will be most noticeable due to large rotations or panning of the camera. Here we try to avoid huge scaling errors by running ORB-SLAM several times over different colon segments. In a real scenario, one would likely make 3D models only in segments of the colon surrounding pathologies. However, if longer segments are needed, one may concatenate several reconstructed segments after estimating some scale factors. One may then use the fact that the colon can be approximately described as a tube with radius being contained within certain boundaries.

ORB-features are detected under SLAM. However, SURF and eigen features are applied to compute dense reconstruction in Algorithm 1 as they appear to produce more reliable features for typical colon geometry and texture. We assume the same pre-processing as for SfM. The textures of the colon walls as well as its geometry are both crucial to obtain enough features to obtain a decent reconstruction. This leads to feature detection over a range of scales. To cover all relevant scales, 8 pyramid levels in the feature detection is needed. Further, all frames with significant motion blur are removed manually as they make the algorithm fail due to lack of ORB-feature matches.

Experiments: We consider three scenarios of colon segments that a 3D reconstruction algorithm should be able to handle: 1. Nearly straight short segments. 2. Longer sections bending slowly. 3. Shorter segments with sharp bends.

A version of ORB-SLAM has been implemented by the MatLab community[5]. We build on and extend this example for our purposes here.

Scenario 1: The colon segment under consideration is given in Fig. 3(a) and is the same as in the SfM case. 68 images were generated of this segment in VR-CAPS, and 49 key-frames was chosen by the ORB-SLAM algorithm for reconstruction. Notice that the reconstruction is less noisy and denser than what was the case for pure SfM in Fig. 2. It also appears to be a better reconstruction which fits quite well with the colon model as shown in Fig. 3(d). This indicates that ORB-SLAM copes with colon geometry, providing better accuracy than pure SfM.

Scenario 2: The colon segment under consideration is given in Fig. 4(a). 996 images were generated of this segment in VR-CAPS, and 445 key-frames was chosen by the ORB-SLAM algorithm for reconstruction. The estimated camera poses, i.e., the movement of the camera through the relevant segment, as well as the corresponding sparse point cloud is shown in Fig. 4(b). "Optimized trajectory" refers to a global optimization over all key-frame camera poses after ORB-SLAM. The "ragged" trajectories fits with the simulated movement obtained through VR-CAPS. The sparse cloud seems to capture the overall shape of the colon segment. The dense reconstruction in Figs. 4(c) and 4(d) shows a clearer outline of the reconstruction, and appears to have quite similar shape to the relevant segment. Note in particular that the narrowing of the colon is captured quite well. However, there is quite some noise in the cloud, particularly around the sharpest bend as well as close to the end of the segment, which is expected due to lack of loop closure. Overall, the result is quite promising.

Scenario 3: The colon segment under consideration is given in Fig. 4(a). 1173 images were generated of this segment in VR-CAPS, and 332 key-frames was chosen by the ORB-SLAM algorithm for reconstruction. The movement of the camera through the relevant segment, as well as the corresponding sparse point cloud is shown in Fig. 5(b). The sparse cloud again seems to capture the rough outline of the colon segment, and the "ragged" trajectory is inline with the simulated movement. The dense reconstruction in Figs. 5(c) and 5(d) shows a clearer outline of the reconstruction, and appears to have quite similar shape to the relevant segment. However, there is even more noise than in Scenario 2, especially after the sharp bend, which is expected due to scale- and position drift. Also, we can see that the cloud is denser on the outer side of the bend, whereas it is very sparse, or lacking completely, on the inner side. The reason is that

[5]https://se.mathworks.com/help/vision/ug/monocular-visual-simultaneous-localization-and-mapping.html (20/11-21)

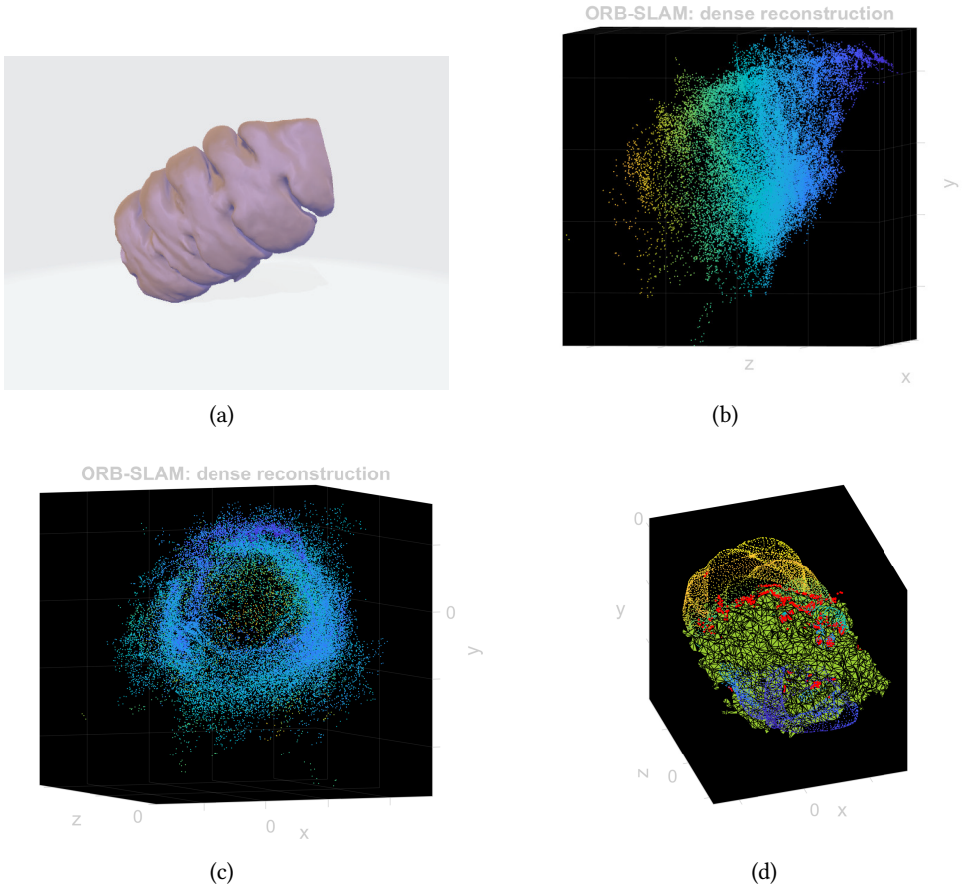


Figure 3: ORB-SLAM 3D reconstruction, scenario 1: (a) Relevant colon segment (ground truth) (b) Dense point cloud from the side. (c) Dense point cloud in front. (d) 3D Point cloud (red) with α-shape triangulation overlayed ground truth colon segment

the camera is mainly facing outwards while it moves throughout the bend. A real WCE has a fisheye lens with a much larger viewing angle, and so one may expect this effect to be less severe (but still present). Anyway, the results seem quite promising, especially given the fact that the results are not yet optimized and fully (post-) processed.

As ORB-SLAM copes quite well with all three scenarios, even sharp bends, it seems that 3D reconstruction of the human colon is indeed possible. As mentioned earlier, for further evaluation and post-processing we first need a numerical comparison to ground truth.

4. Summary and Conclusions

In this paper we have investigated the possibility for 3D reconstruction of the human colon from WCE images using structure from motion and ORB-SLAM. To generate data sets, we used

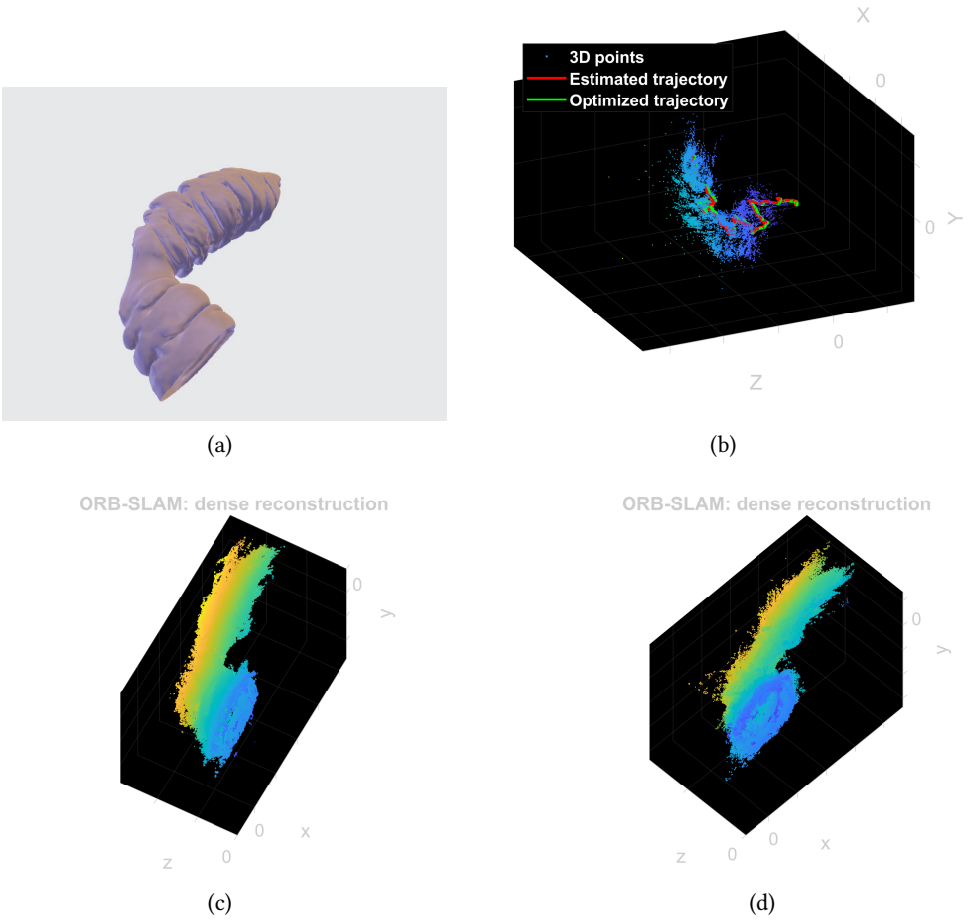


Figure 4: ORB-SLAM 3D reconstruction, scenario 2: (a) Relevant colon segment (ground truth) (b) Estimated camera poses with sparse point cloud. (c) Dense point cloud from the side. (d) Dense point cloud at slightly different vantage point.

a virtual graphics-based environment emulating both the human colon as well as the WCE's movement through it. Experimental results in this paper indicate that 3D reconstruction of human colon is possible.

Future research should aim at optimizing the 3D reconstruction process and find suitable post-processing methods to improve the resulting point cloud. Extensions to more realistic scenarios include non-rigid motion as well as enabling of distortions and artefact seen in real WCE images, all of which can be emulated in VR-CAPS. Then, with suitable pre-processing algorithms based on what we have learned through experiments in place, a study on real WCE videos will be possible. One should also combine or merge methods studied here with single image techniques, like the effort in [6] using shape from shading, to cope with a broader scenario. Lastly, one may eliminate drifts due to lack of loop closure in ORB-SLAM through additional

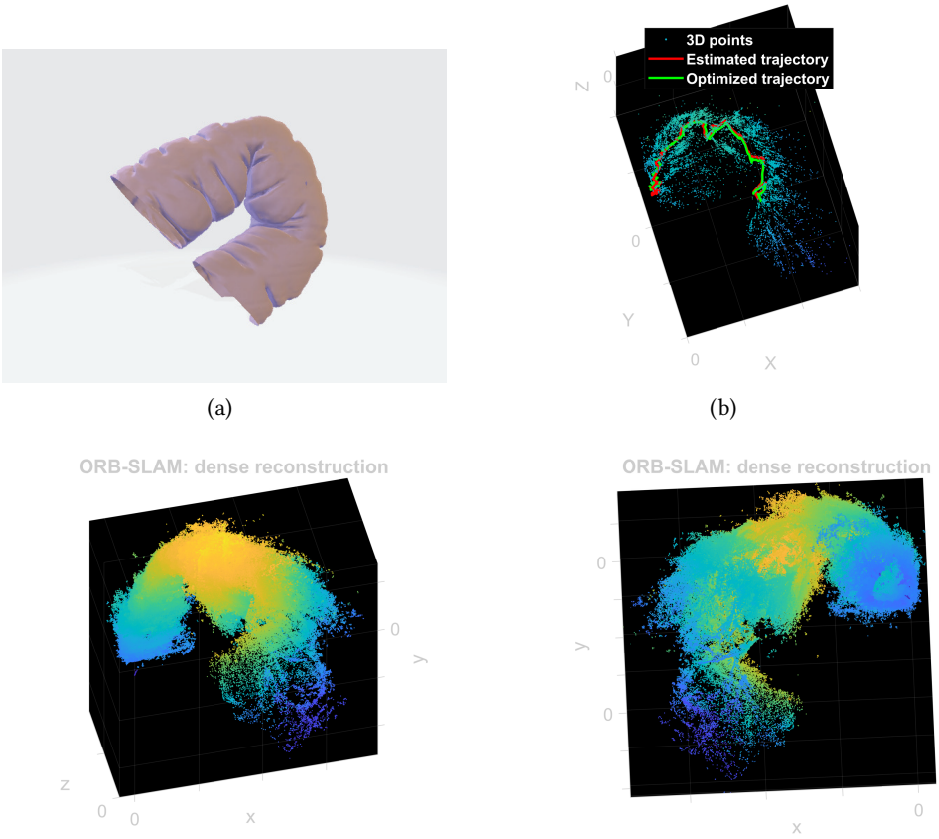


Figure 5: ORB-SLAM 3D reconstruction, scenario 3: (a) Relevant colon segment (ground truth) (b) Estimated camera poses with sparse point cloud. (c) Dense point cloud from the side. (d) Dense point cloud at opposite side.

information available. The WCE emits electromagnetic radiation received by several on-body sensors that can be used to track its position quite accurately [14], or compute the path length traveled [15]. This can help to correct for drift in position.

5. Acknowledgments

We would like to give our appreciation to Anuja Vats for bringing our attention to the VR-CAPS environment.

References

[1] Editorial, Improving uptake of colorectal cancer screening, The Lancet Gastroenterology & Hepatology 2 (2017).

[2] A. T. C. et al., Wireless capsule endoscopy, Gastrointestinal Endoscopy 78 (2013) 805–815.

[3] K. N. et. al., Three-dimensional upper gastrointestinal endoscopy: A clinical study of safety and an ex vivo study of utility in endoscopic submucosal dissection, Gastrointestinal Endoscopy 87 (2018).

[4] B. Horn, M. Brooks, The variational approach to shape from shading, Computer Vision, Graphics, and Image Processing 33 (1986) 174–208.

[5] R. I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, second ed., Cambridge University Press, ISBN: 0521540518, 2004.

[6] B. Ahmad, P. A. Floor, I. Farup, 3d reconstruction of gastrointestinal regions from single images, in: Colour and Visual Computing Symposium (CVCS), Gjøvik, Norway, 2022.

[7] S. Jensen et al., A benchmark and evaluation of non-rigid structure from motion, International Journal on Computer Vision 129 (2021) 882–899.

[8] V. Sidhu, E. Tretschk, V. Golyanik, A. Agudo, C. Theobalt, Neural dense non-rigid structure from motion with latent space constraints, in: European Conference on Computer Vision (ECCV), 2020.

[9] G. Klein, D. Murray, Parallel tracking and mapping for small ar workspaces, in: 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007, pp. 225–234.

[10] R. Mur-Artal, J. M. M. Montiel, J. D. Tardós, Orb-slam: A versatile and accurate monocular slam system, IEEE Transactions on Robotics 31 (2015) 1147–1163.

[11] K. Incetan et al., Vr-caps: A virtual environment for capsule endoscopy, Medical Image Analysis 70 (2021).

[12] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. of the ACM 24 (1981) 381–395.

[13] G. Dimas, F. Bianchi, D. K. Iakovidis, A. Karargyris, G. Ciuti, A. Koulaouzidis, Endoscopic single-image size measurements, Measurement Science and Technology 31 (2020) 9–15.

[14] B. Moussakhani, J. T. Flåm, S. Støa, I. Balasingham, T. Ramstad, On localisation accuracy inside the human abdomen region, IET Wireless Sensor Systems 2 (2012) 9–15.

[15] A. Bjørnevik, P. A. Floor, I. Balasingham, On path length estimation for wireless capsule endoscopy, in: 12th International Symposium on Medical Information and Communication Technology (ISMICT), 2018, pp. 1–5.