Universidad Zaragoza
1542

Master's Thesis

Real-scale 3D reconstruction from monocular endoscope images

Reconstrucción 3D a escala real a partir de imágenes monoculares de endoscopio

Author

Víctor Martínez Batlle

Supervisor

Juan Domingo Tardós Solano

Master Program in Robotics, Graphics and Computer Vision

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2021

Acknowledgements

Special thanks to my supervisor, Prof. Juan Domingo Tardós Solano, for his guidance and advice. I would also like to thank Prof. José M. M. Montiel and Diego Royo for their time and help.

This project has received funding from
EndoMapper: Real-time mapping from endoscopic video.
European Union's Horizon 2020 research and innovation
program under grant agreement No 863146.

Abstract

In today's scientific community, there is a growing research interest in extending augmented reality and autonomous navigation into the human body. The European EndoMapper project aims to solve the problem of real-time 3D mapping of the human colon that will assist colonoscopy, tumour biopsy and other medical procedures.

Due to space limitations inside the human bowel, medical endoscopy can only use monocular vision. However, conventional mapping systems are not able to recover the real scale of the environment from monocular images. This intrinsic limitation leads to unknown scale maps, making it difficult to diagnose some diseases of the human colon. In addition, these maps are often distorted, as they suffer from scale drift, which is a common problem in monocular systems.

This MEng thesis proposes a solution to monocular real-scale 3D reconstruction inside the human colon. Our approach exploits the controlled lighting inside the human body, where the only light source moves jointly with the camera. This allows us to consider a pseudo-stereo pair formed by the endoscope's light and camera, and use it to achieve real-scale perception on a monocular endoscope.

First, we define a model of the illumination and camera of an endoscope, which allows us to understand the imaging process during colonoscopy. Then, this model is adapted and calibrated for a real medical endoscope, using only a sequence of images recorded by the endoscope itself. Finally, we propose a method capable of estimating a dense depth map from a single monocular image, based on the calibration we performed.

To evaluate the accuracy of our depth estimation, we conducted experiments with both synthetic and real images of the human colon. With respect to synthetic data, we obtain results with a 7% error, which is less than 3 mm on average. Lastly, we demonstrate that our method can also work with real images, where we estimate dense depth maps that preserve the structure and discontinuities of the human colon.

Index

Chapter 1

Introduction

Visual SLAM or VSLAM (Visual Simultaneous Localization and Mapping) is a problem that has been intensively researched over the last two decades. It consists of constructing a 3D map of the environment from images taken by a camera, while calculating its trajectory.

This field of study finds its application in different areas, including the health care sector. The European EndoMapper project (EU-H2020 grant 863146) aims to establish the scientific basis for reconstruction of 3D maps of the interior of the human body from medical endoscopy images.

In conventional VSLAM systems, using monocular cameras, the scale of the environment is not observable. This causes inconsistencies in the constructed maps. Stereo [14] and visual-inertial [2] cameras solve this problem, known as scale drift. But, in medical endoscopy and many other applications, size and power limitations restrict visual SLAM to the use of monocular cameras.

However, the interior of the human body is an example of an artificially illuminated environment, like that of a vehicle driving at night, deep-sea submarines or robotic exploration of underground caves. Here, the light source that illuminates this kind of scene is controlled and linked to the camera movement.

Our fundamental hypothesis is that we can take advantage of the illumination used in such environments to reconstruct real-scale 3D scenes using conventional monocular endoscopes, obtaining pseudo-stereo information by considering a light-camera pair.

1.1 Related work

Deep learning based methods. Recent results in single-image depth estimation using deep convolutional networks [6] open the possibility of designing accurate full-scale monocular SLAM systems [22, 24]. Our previous work [11] demonstrates that, using a depth prediction network, it is possible to perform pure real-scale monocular

VSLAM, obtaining almost the same accuracy as with stereo VSLAM, and eliminating scale drift. However, there are not enough stereo images of the inside of the human body to allow these networks to learn how to predict real scale.

Photometric based methods. Other authors focus on the study of photometry to obtain dense, real-scale reconstructions of outdoor and indoor 3D scenes under constant ambient light [4, 15]. Regarding medical applications, controlled illumination seems capable of achieving similar results. [7, 12, 16, 23].

Previous works have three phases in common. First, they propose a lighting model for their working environment. Specifically, they model light emission, interaction with surfaces, and capture by the camera. Next, they calibrate the light and camera of their endoscopy system based on that model. Finally, they combine photometric information to obtain a depth map by global iterative refinement [12] or an upwind scheme from certain to uncertain points [7].

Following these approaches, we rely on photometric methods to model and calibrate the light and camera of an endoscope, and therefore we know how lighting changes when the light source moves along with the camera. Thus, unlike reconstruction methods designed for indoor and outdoor environments, we can estimate a depth map inside the human body without the assumption of constant ambient lighting.

1.2 Objectives and scope

The objective of this MEng thesis is to rely on photometric methods to obtain a system capable of reconstructing at real scale the dense depth of a 3D scene from monocular endoscopy images. Adapted from similar approaches found in the literature, points to be addressed in this work are:

- First, in Chapter 2, we propose a model of illumination system in a medical endoscope whose light source is linked with the camera movement.
- Then, in Chapter 3, we calibrate the illumination model for a real monocular endoscope using the "Hospital Clínico Universitario Lozano Blesa" (HCULB) dataset provided by the EndoMapper project.
- Next, in Chapter 4, we study and implement a method to estimate the depth of a 3D scene from monocular images based on the photometric illumination model.
- Finally, in Chapter 5, we validate the proposed method with synthetic datasets that can provide ground-truth depth information as well as with real colon images.

Chapter 2

Endoscope model

This MEng thesis presents a photometric approach to the problem of 3D reconstruction inside the human body from images taken during a medical endoscopy procedure. Such an approach considers the image formation geometry model and the light-transport photometric model. Both are discussed in this chapter.

2.1 Geometric model

A monocular camera observes the world as projected in a 2D image space. The imaging model determines how 3D geometry is captured by the camera in its image plane. Let's denote $\mathbf{x}_{\mathcal{C}}$ a point in the camera frame of reference $\mathcal{C}$. Then, a camera model $\pi : \mathbb{R}^3 \to \Omega$ relates 3D coordinates $\mathbf{x}_{\mathcal{C}} = [x, y, z]^T$ to 2D coordinates $\mathbf{u} = [u, v]^T$ on the image domain $\Omega \subset \mathbb{R}^2$.

In addition, we consider a global frame of reference $\mathcal{W}$. Point coordinates $\mathbf{x}_{\mathcal{W}}$ in the world frame can be expressed in camera frame by multiplying

$$\mathbf{x}_{\mathcal{C}} = \mathbf{T}_{\mathcal{CW}}\mathbf{x}_{\mathcal{W}}, \tag{2.1}$$

where $\mathbf{T}_{\mathcal{CW}}$ is a matrix composed by rotation $\mathbf{R}_{\mathcal{CW}}$ and translation $\mathbf{t}_{\mathcal{CW}}$, from world to camera frame. These values are called *extrinsic* camera parameters. Its inverse transformation $\mathbf{T}_{\mathcal{WC}} = \mathbf{T}_{\mathcal{CW}}^{-1}$ is referred as the *pose* of the camera.

Conventional perspective cameras are usually correctly modelled as pinhole cameras, once the small distortions caused by the actual lenses have been corrected. However, an endoscope camera is designed to cover a wider view angle. This causes greater distortions when capturing the environment. Consequently, there exist specific models for this type of camera.

2.1.1 Pinhole camera

A pinhole camera model assumes no distortions on the image formation process. The resulting *projection* model is

$$\mathbf{u} = \pi_{\text{ph}}(\mathbf{x}_{\mathcal{C}}), \quad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x \frac{x}{z} + c_x \\ f_y \frac{y}{z} + c_y \end{bmatrix}, \tag{2.2}$$

where f_x, f_y are the camera focal length in pixels and $[c_x, c_y]^T$ is its principal point. These values are called *intrinsic* camera parameters.

Given the above projection model, all possible 3D points that can be imaged in an image point $\mathbf{u}$, are those on a ray whose origin is the optical centre of the camera. The direction of that ray $\mathbf{r}_d$ is defined by the *unprojection* model

$$\tilde{\mathbf{r}}_d = \pi_{\text{ph}}^{-1}(\mathbf{u}), \quad \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{u - c_x}{f_x} \\ \frac{v - c_y}{f_y} \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{r}_d = \frac{\tilde{\mathbf{r}}_d}{\|\tilde{\mathbf{r}}_d\|}. \tag{2.3}$$

We define the Z-depth of a point $\mathbf{x}$ as the z coordinate of that point with respect to the camera frame. Given the Z-depth of an image point $\mathbf{u}$, we can define the *extended unprojection* model as

$$\mathbf{x}_{\mathcal{C}} = \pi_{\text{ph}}^{-1}(\mathbf{u}, z), \quad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \tilde{\mathbf{r}}_d \cdot z. \tag{2.4}$$

Note that this model recovers the original coordinates, solving the ray ambiguity.

2.1.2 Fisheye camera

Fisheye vision can be achieved with different kinds of cameras. In this work, we consider the Kannala & Brandt model [9] with four coefficients as the best alternative to our endoscope distortion.

This model is based on spherical projection. Thus, first we get the projection of the 3D world point $\mathbf{x}_{\mathcal{C}} = [x, y, z]^T$ on the unit sphere as azimuth (φ) and elevation (θ) angles such that

$$\varphi = \arctan(y/x), \qquad \theta = \arctan\left(\frac{\sqrt{x^2 + y^2}}{z}\right). \tag{2.5}$$

Then, the *projection* model results in

$$\mathbf{u} = \pi_{\text{kb}}(\mathbf{x}_{\mathcal{C}}), \quad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x\, d(\theta) \cos\varphi + c_x \\ f_y\, d(\theta) \sin\varphi + c_y \end{bmatrix}, \tag{2.6}$$

where the polynomial $d(\theta) = \theta + k_1\theta^3 + k_2\theta^5 + k_3\theta^7 + k_4\theta^9$ models the actual lens distortion with four coefficients k_{1-4}.

As a result, the endoscope camera model has four intrinsic parameters for the pinhole model (f_x, f_y, c_x, c_y) and four new parameters to explicitly model the large lens distortion (k_{1-4}).

The *unprojection* model of the endoscope camera is equivalent to the one presented in the previous section. First, we undo the pinhole projection such that

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \pi_{\text{ph}}^{-1}(\mathbf{u}) = \begin{bmatrix} d(\theta)\cos\varphi \\ d(\theta)\sin\varphi \\ 1 \end{bmatrix}. \tag{2.7}$$

Then, we can solve for φ and θ, noting that

$$\frac{y'}{x'} = \frac{d(\theta)\sin\varphi}{d(\theta)\cos\varphi} = \tan\varphi, \quad \varphi = \arctan(y'/x'), \tag{2.8}$$

$$r = \sqrt{(x')^2 + (y')^2} = \sqrt{d(\theta)^2(\sin^2\varphi + \cos^2\varphi)} = d(\theta), \tag{2.9}$$

and solving the ninth degree polynomial $\theta = d^{-1}(r)$.

Finally, we convert the ray direction $\mathbf{r}_d$ from polar to Cartesian coordinates

$$\mathbf{r}_d = \begin{bmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{bmatrix} \tag{2.10}$$

The full unprojection model is denoted as $\mathbf{r}_d = \pi_{\text{kb}}^{-1}(\mathbf{u})$. This model can be extended to obtain $\mathbf{x}_{\mathcal{C}} = \pi_{\text{kb}}^{-1}(\mathbf{u}, d) = \mathbf{r}_d \cdot d$, where d is the Euclidean distance to the point.

2.2 Photometric model

The photometric basis of our work requires knowledge of the light transport phenomenon. For this purpose, both the emission of light from the endoscope and the interaction of light with the scene must be modelled. Additionally, our model must consider how endoscopes capture the light that reaches their camera.

2.2.1 Light model

The illumination system mounted on an endoscope usually consists of one or more small lights. These lights are usually modelled as punctual lights [7, 12]. In the Point Light Source model (PLS) emission comes from an infinitesimal point $\mathbf{x}_l$ in space, called *light centre*. Light emitted from the source propagates in a sphere. Consequently, we observe an inverse square fall-off, such that light reaching an arbitrary point can be computed as

$$\sigma_{\text{PLS}}(\mathbf{x}) = \frac{\sigma_o}{\|\mathbf{x} - \mathbf{x}_l\|^2} = \frac{\sigma_o}{d^2}, \tag{2.11}$$

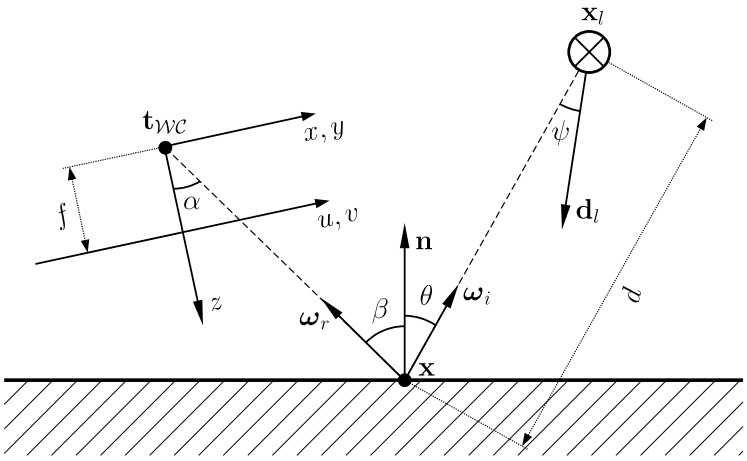


Figure 2.1: Outline of the most important variables of our work.

where σ_o is the radiance observed at unit distance, and $\|\mathbf{x} - \mathbf{x}_l\|$ is the Euclidean distance from the light source to the point, noted as d in Figure 2.1.

As stated by Modrzejewski et al. [12], in most endoscopes light is transmitted to the tip using optical fibre. Therefore, they propose an extension for the PLS model. Based on their approach, we define the General Spot Light Source model (GSLS) by incorporating a spread function $\mu(\mathbf{x})$. This term models the amount of light emitted in each direction of the sphere, so that

$$\sigma_{\mathrm{GSLS}}(\mathbf{x}) = \mu(\mathbf{x})\, \sigma_{\mathrm{PLS}}(\mathbf{x}) \tag{2.12}$$

The spread function μ can be specified by taking a principal direction $\mathbf{d}_l$, that denotes the beam over which maximum radiance is emitted. Assuming a radial attenuation from the principal direction, the spread function can be modelled as a cosine fall-off [7] or as a polynomial decay [12]. We consider both approaches for our model

$$\mu_{\cos}(\mathbf{x}) = \cos^k \psi, \tag{2.13}$$

where k is a parameter of the model, or, alternatively

$$\mu_{\mathrm{poly}}(\mathbf{x}) = 1 + k_1\psi^2 + k_2\psi^4 + k_3\psi^6, \tag{2.14}$$

having three parameters, i.e. coefficients k_{1-3}.

Both models depend on the angle ψ between the sampled outgoing direction and the principal direction, as in Figure 2.1, such that

$$\psi = \langle \mathbf{x} - \mathbf{x}_l, \mathbf{d}_l \rangle, \tag{2.15}$$

that can be computed from cosine similarity of unit vectors

$$\psi = \arccos\left(\frac{\mathbf{x} - \mathbf{x}_l}{\|\mathbf{x} - \mathbf{x}_l\|} \cdot \mathbf{d}_l \right). \tag{2.16}$$

2.2.2 Surface reflectance

When light reaches a surface, most of it will be reflected, going out in different directions depending on the material properties. Reflectance of a surface has been extensively studied in Computer Graphics [17, Chapter 5], giving rise to many models [13]. The term *bidirectional reflectance distribution function* (BRDF) refers to a general function that defines how light is reflected at an opaque surface. Actual definition of this function depends on the surface itself.

On the one hand, perfectly diffuse materials reflect the same amount of light in all the hemisphere domain. The Lambertian model [10] represents this ideal behaviour. His definition is usually formulated as

$$f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r) = \frac{k_d}{\pi}, \tag{2.17}$$

where k_d is a parameter of the model that indicates the amount of incident light that is reflected, versus that absorbed by the surface. It is usually named *surface albedo*. As shown in Figure 2.1, unitary vectors $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_r$ are the directions that connect the surface point with the origin of the incoming light and with the destination of the reflected radiance, respectively.

On the other hand, in perfectly specular materials, for a given $\boldsymbol{\omega}_i$ direction, all light is reflected in a single outgoing direction. Subsequently, this behaviour is usually modelled as a delta function

$$f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r) = \delta_{\boldsymbol{\omega}_s}(\boldsymbol{\omega}_r), \tag{2.18}$$

whose value is one if and only if the outgoing direction $\boldsymbol{\omega}_r$ is the perfect specular direction $\boldsymbol{\omega}_s$ [3, Chapter 10].

However, real materials combine both behaviours. The Phong reflection model [18] is a combination of diffuse reflection on rough surfaces with specular reflection on shiny surfaces. It can be defined as

$$f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r) = \frac{k_d}{\pi} + k_s(\boldsymbol{\omega}_r \cdot \boldsymbol{\omega}_s)^{\alpha}, \tag{2.19}$$

where k_d, k_s model the amount of diffuse and specular behaviour, and α determines material roughness.

In addition, the angle of incidence of light with respect to the surface normal, named θ in Figure 2.1, changes the area of the projected solid angle and thus causes attenuation in the reflected light. We account for this, by adding a *cosine term* to previous BRDF definitions, so that

$$f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r, \mathbf{n}) = f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r)\, \cos\theta = f_{\mathrm{r}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r)\, (\boldsymbol{\omega}_i \cdot \mathbf{n}) \tag{2.20}$$

is the amount of light reflected from $\boldsymbol{\omega}_i$ to $\boldsymbol{\omega}_r$ in a surface point with normal $\mathbf{n}$.

2.2.3 Camera response model

Digital cameras rely on sophisticated light sensors capable of measuring the number of photons arriving through the lens and aperture during an exposure time. All these components influence the measurements that a camera reflects in captured images. In temporal order, light reaching a camera is affected by the following factors.

Lens vignetting. Cameras use complex lens systems to guide light from the scene to their sensor. Those systems introduce an attenuation in image space, known as *mechanical vignetting*, which is in addition to the *natural vignetting* caused by foreshortening [20]. Typically, this second vignetting depends on the cosine of α angle between the camera's forward vector and the direction of incoming light (see Figure 2.1). Natural vignetting tends to approximate to $\cos^4 \alpha$. But mechanical vignetting is not easy to model theoretically. Therefore, vignetting is usually approximated empirically. Engel et al. [5] propose a function $V : \Omega \to [0, 1]$ that models a per-pixel attenuation value from a set of images of diffuse, uniform surfaces. However, this function can be reformulated in a parametric manner, reducing the required number of dense pixel samples. We assume a radial attenuation from the camera's forward vector, such that

$$V(\mathbf{u}) = \cos^k \alpha. \tag{2.21}$$

Auto-gain. Endoscope cameras might automatically adjust some parameters, such as exposure time or analogue signal amplification [20]. In video streams, signal amplification is controlled by an automatic gain control (AGC) logic. We assume this *auto-gain* usually acts as a multiplying factor. Thus, all pixel values on i-th image have the same gain value g_i that depends on the camera's internal logic.

$$I'_i(\mathbf{u}) = g_i \cdot I''_i(\mathbf{u}) \tag{2.22}$$

Image post-processing. When digital cameras convert analogue signal to digital bits (ADC), they perform a variety of *digital signal processing* (DSP) operations. One of these processes is called *gamma correction*. In order to increase perceived dynamic range, cameras map the luminance values through a gamma function such that

$$I_i(\mathbf{u}) = I'_i(\mathbf{u})^{1/\gamma}. \tag{2.23}$$

Although it may vary among manufacturers, $\gamma = 2.2$ is a commonly used value [20]. Moreover, this gamma function parameter does not change overtime.

2.2.4 Complete photometric model

Our complete photometric model considers all concepts introduced above, as a combination of light, surface and camera models

$$\begin{aligned} \mathcal{I}(\mathbf{x}) &= \Big(\sigma_{\text{GSLS}}(\mathbf{x})\ f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r, \mathbf{n})\ V(\mathbf{u})\ g_i\Big)^{1/\gamma} \\ &= \left(\frac{\mu(\mathbf{x})\ \sigma_o}{d^2}\ f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r)\ \cos\theta\ V(\mathbf{u})\ g_i\right)^{1/\gamma}, \end{aligned} \tag{2.24}$$

where

- $\mathbf{x} \in \mathbb{R}^3$ is the 3D point that is being imaged.
- $\mathbf{n} \in \mathbb{R}^3$ is the normal of the surface at that point.
- $\mathbf{u} \in \Omega$ is its corresponding image pixel, such that $\mathbf{u} = \pi(\mathbf{x})$.
- d is the Euclidean distance from the light source to the point.
- σ_o is the outgoing radiance from the light source at unit distance.
- $\mu(\mathbf{x})$ is the light spread function.
- $f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r)$ is the BRDF from $\boldsymbol{\omega}_i$ incoming to $\boldsymbol{\omega}_r$ outgoing direction.
- θ is the incidence angle $\langle \boldsymbol{\omega}_i, \mathbf{n} \rangle$ of the cosine term.
- $V(\mathbf{u})$ is the camera's vignetting.
- g_i and γ are the camera's gain and gamma values.

Chapter 3

Endoscope calibration

The endoscope model presented in the previous chapter is intended to cover a generic device used in medical endoscopy. In our project, we will use the workstation defined within EndoMapper project.

The Olympus EVIS EXERA III endoscopic system was used for taking and recording images of endoscopic procedures at the digestive service of Hospital Clínico Universitario Lozano Blesa (HCULB) in Zaragoza. This platform includes a GIF-H190 endoscope with a CLV-190 light source, connected to a CV-190 video processor.

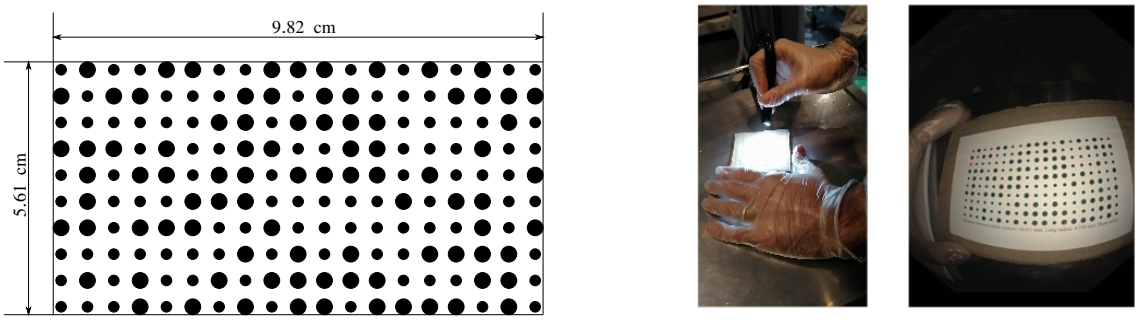
In this chapter, we present an adaptation of the general model to our real endoscope. In Sections 3.1 and 3.2, we propose a calibration methodology based on images taken by the endoscope itself. And Section 3.3 discusses the results obtained from calibrating on pre-existing HCULB dataset from EndoMapper project.

3.1 Geometric calibration

The geometric camera model, discussed in Section 2.1, consists of eight intrinsic parameters: focals, f_x f_y, principal point $[c_x, c_y]^T$ and distortion coefficients k_{1-4}. Using a Vicalib [1] pattern, we calibrate those eight values for the Olympus endoscope. Calibration images were acquired as show in Figure 3.1.

As a result of the geometric calibration, we obtain the eight intrinsic values as well as the camera pose for each calibration image. Reported intrinsic parameters are:

- Focal distance: $f_x = 717.21$ px, $f_y = 717.48$ px.
- Principal point: $c_x = 735.37$ px, $c_y = 552.80$ px.
- Distortion coefficients: $k = [-0.13893, -1.2396\text{E-}03, 9.1258\text{E-}04, -4.0716E\text{-}05]$.



(a) (b)

Figure 3.1: Image acquisition for calibration. (a) Size and distribution of the Vicalib pattern used. (b) Overview of the acquisition of calibration images.

3.2 Photometric calibration

After geometric calibration, we calibrate the endoscope's photometry. First, we adapted the photometric model, discussed in Section 2.2, to our real endoscope. Then, we define a cost function that aims to minimise the photometric error of a sparse set of points.

3.2.1 Model simplifications

Given the specifications of our real endoscope, we introduce a set of simplifications to our initial endoscope model, in order to favour calibration convergence.

Light model. Olympus GIF-H1900 endoscope has a fisheye camera coupled with two fibre optic light beams (Figure 3.2). As both light sources are very close, we will assume that they can be modelled as a single light source (GSLS) with a joint spread function. Furthermore, we will assume that this virtual light source is located at the camera's optical centre, i.e. $\mathbf{x}_l = \mathbf{t}_{\mathcal{WC}}$.

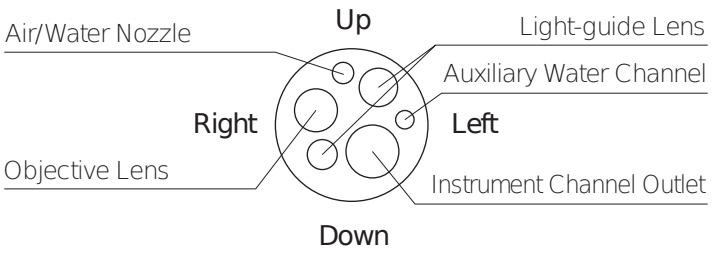


Figure 3.2: Insertion section of an Olympus EVIS EXERA III GIF-H190 endoscope.

Surface reflectance. The general BRDF model $f_r(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r)$ considers directions of incidence and reflection. However, our virtual light is placed over camera's optical centre, so both directions are the same, i.e. $\boldsymbol{\omega}_i = \boldsymbol{\omega}_r$, therefore $\beta = \theta$ (see Figure 2.1).

In addition, most approaches in the literature assume perfectly diffuse surfaces. We compare this assumption against a non-parametric BRDF model that considers only the θ angle between incident/reflected direction and surface normal.

For this purpose, we find a finite set of values F_r, such that

$$f'_r : \Theta \to F_r, \quad \Theta \subset [0, \pi] \tag{3.1}$$

and apply linear interpolation for the rest of the function domain

$$f''_r : [0, \pi] \to [0, 1], \quad f''_r(\theta) = \frac{(\theta - \Theta_1) f'_r(\Theta_2) + (\Theta_2 - \theta) f'_r(\Theta_1)}{\Theta_2 - \Theta_1}, \tag{3.2}$$

where Θ_1 and Θ_2 are the two values in Θ closest to θ, such that $\Theta_1 \leq \theta < \Theta_2$.

We can still denote $f(\theta) = f''(\theta) \cos\theta$, to account for the cosine attenuation.

Camera response model. On the one hand, camera's vignetting $V(\mathbf{u})$ and virtual light spread function $\mu(\mathbf{x})$ are coupled, since we have aligned both centres. Thus, we optimise the effect of both functions jointly, resulting in a single function

$$\mu'(\mathbf{x}) \stackrel{\text{def}}{=} \mu(\mathbf{x})\, V\left(\pi_{\text{kb}}(\mathbf{x})\right). \tag{3.3}$$

On the other hand, auto-gain logic of the endoscope is unknown. Therefore, g_i values of each image are coupled with the absolute radiance σ_o emitted by the light, so that their effects cannot be separated. Consequently, we fix the σ_o parameter to an arbitrary value, and optimise the apparent auto-gain.

Simplified model. Finally, we obtain a simplified photometric model, which is parameterised according to the unknowns we want to estimate for our endoscope:

$$\mathcal{I}\left(\mathbf{x},\, \mathbf{d}_l,\, \mathbf{k},\, \mathbf{f},\, g_i,\, \gamma\right) = \left(\frac{\mu'(\mathbf{x},\, \mathbf{d}_l,\, \mathbf{k})}{d^2}\; f_r(\theta,\, \mathbf{f})\; \cos\theta\; g_i \right)^{1/\gamma}. \tag{3.4}$$

where

- $\mathbf{d}_l \in \mathbb{R}^3$ is the principal direction of the $\mu(\mathbf{x})$ spotlight model.
- $\mathbf{k} = [k_1, k_2, \ldots, k_m] \in \mathbb{R}^m$ are the parameters of $\mu'(\mathbf{x})$ joint vignetting and light spread function. Their number m may vary depending on the model.
- $\mathbf{f} = [f_1, f_2, \ldots, f_q] \in \mathbb{R}^q$ are variable parameters of $f_r(\,)$ BRDF models.
- $g_i \in [0, \infty)$ is the estimated auto-gain value for i-th image.
- $\gamma \in [0, \infty)$ is the gamma correction value applied by the camera.

3.2.2 General calibration method

Our calibration method is defined as an optimisation problem:

$$\{d_l^\theta, d_l^\varphi, \mathbf{k}, \mathbf{f}, g_i, \gamma \mid \forall i\}^* = \underset{\mathbf{d}_l, \mathbf{k}, \mathbf{f}, \mathbf{g}, \gamma}{\operatorname{argmin}} \sum_{i,j} \rho\left(I_{ij} - \mathcal{I}\left(\boldsymbol{x}_{\mathcal{W}j}, \mathbf{d}_l, \mathbf{k}, \mathbf{f}, g_i, \gamma\right)\right) \tag{3.5}$$

- $[d_l^\theta, d_l^\varphi] \in [0, \pi] \times [0, 2\pi]$ is the light principal direction $\mathbf{d}_l$ in spherical coordinates.
- $\boldsymbol{x}_{\mathcal{W}j} = [x_j, y_j, z_j]^T \in \mathcal{X}$ are the coordinates of j-th sampled point, in world frame.
- I_{ij} is the measured intensity value of j-th point on i-th image.
- $\mathcal{I}(\,)$ is our photometric model, as described in Section 3.2.1.
- $\rho(\,)$ is a robust cost function, e.g. Huber loss.

In order to select $\boldsymbol{x}_{\mathcal{W}j}$ points, that are considered by our calibration, we sample a uniform sparse grid on the same Vicalib pattern used for geometric calibration. These points must be in white areas of the pattern, ensuring maximum radiance reflection. In Figure 3.3 we illustrate the selected set of points.

3.2.3 Photometric error

The optimisation described above is based on the error between actual intensity measurements I_{ij} and our photometric model $\mathcal{I}(\,)$. For each sampled point $\boldsymbol{x}_j$ we get as many photometric residuals as the number of images, if the projection of a point falls inside the image plane boundary. Thus, we define

$$I_{ij} = I_i(\pi_{\text{kb}}(\mathbf{R}_{i\mathcal{W}}\boldsymbol{x}_{\mathcal{W}j} + \mathbf{t}_{i\mathcal{W}})), \tag{3.6}$$

where $\mathbf{R}_{i\mathcal{W}}$ and $\mathbf{t}_{i\mathcal{W}}$ are the extrinsic camera parameters, provided by Vicalib's geometric calibration, for the i-th image.

The errors we will consider in this work are:

– The signed photometric error

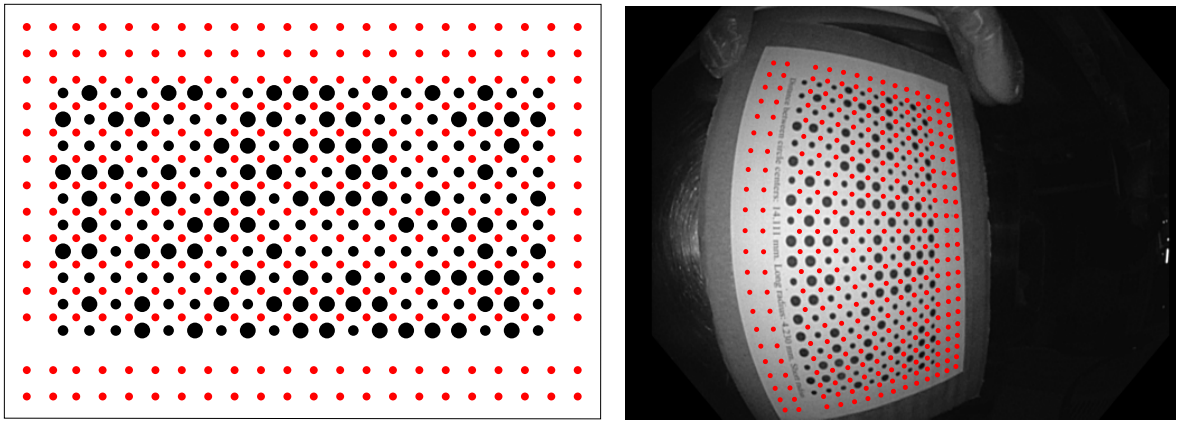
$$E_{ij} = \mathcal{I}(\boldsymbol{x}_{\mathcal{W}j}, \mathbf{d}_l, \mathbf{k}, \mathbf{f}, g_i, \gamma) - I_{ij} \tag{3.7}$$

– The mean absolute error

$$\text{MAE} = \frac{1}{N} \sum_{i,\,j} |E_{ij}| \tag{3.8}$$

– The mean relative error

$$\text{MRE}\ [\%] = \frac{1}{N} \sum_{i,\,j} \frac{|E_{ij}|}{I_{ij}} \cdot 100 \tag{3.9}$$



(a) (b)

Figure 3.3: Sampling the Vicalib pattern. (a) Red marks correspond to each $\boldsymbol{x}_j \in \mathcal{X}$ sampled point. (b) Chosen samples, as projected on one of the optimisation images.

3.3 Calibration results

The endoscope is calibrated using sequence HCULB_00039 from the EndoMapper dataset. We choose 875 images for training the model and reserve 193 images to validate the result. Our final calibration allows us to estimate the intensity of a pixel with an average absolute error of less than 3 grey levels and a relative error of 2.2%. Convergence for each parameter is discussed in the following sections.

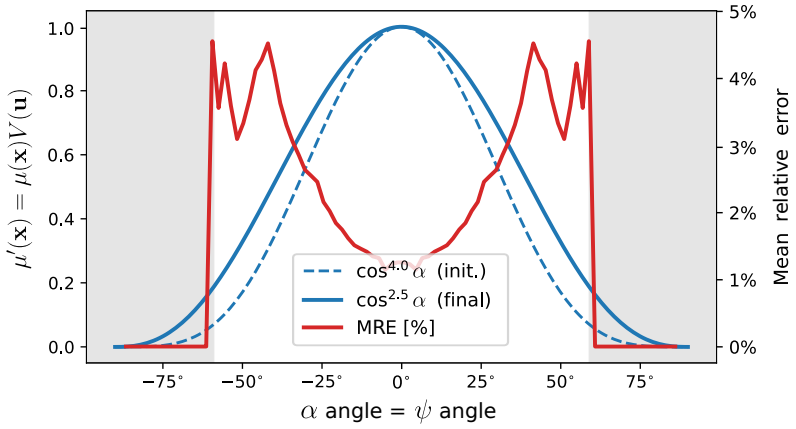
3.3.1 Light principal direction and spread function

In all conducted experiments, the principal direction of the illumination ($\mathbf{d}_l$) was within one degree of the camera forward vector. Given this negligible value, we have simplified the model assuming that the maximum radiance direction is in the camera's z-axis.

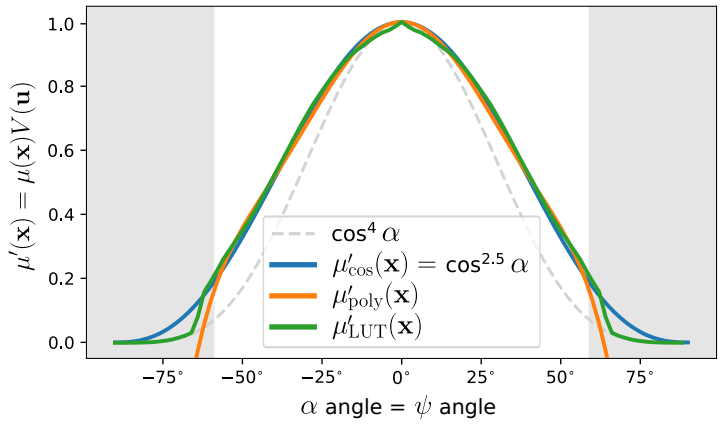
Regarding the dispersion, in Figure 3.4 we can see that the estimated function is equal to $\cos^{2.5}\alpha$. Thus, spread is wider than natural vignetting $\cos^4\alpha$ (see Chapter 2). This is consistent with the illumination system of our endoscope, where two real light beams result in a widening of our virtual beam.

Quantitative evaluation shows a small relative error of 1% in the central area of the image, i.e. 0°. However, error grows towards the edges, reaching 4.5% at 40°. This is because the samples $\boldsymbol{x}_{\mathcal{W}j}$ do not always cover the whole image. Thus, when $\alpha > 60°$ (shaded area) the function remains unsampled in our calibration data.

Finally, Figure 3.4b shows three experiments, comparing $\mu'_{\cos}(\mathbf{x})$ cosine model, $\mu'_{\text{poly}}(\mathbf{x})$ polynomial model and a third $\mu'_{\text{LUT}}(\mathbf{x})$ model, introduced in a similar way to non-parametric BRDF in Section 3.2.1. We conclude that all of them converge to the same solution, so we keep $\mu'_{\cos}(\mathbf{x})$, as it is the simplest one.



(a)



(b)

Figure 3.4: Estimated radiance attenuation caused by light spread function $\mu(\mathbf{x})$ and camera vignetting $V(\mathbf{u})$. (a) A cosine fall-off $\mu'(\mathbf{x}) = \cos^{2.5}\alpha$ yields a 1% to 5% error. (b) Our experiments with different model parametrisations converge to similar results. These models are $\mu'_{\cos} = \cos^{2.5}\alpha$, $\mu'_{\text{poly}} = 1 - 1.5\alpha^2 + 1.36\alpha^4 - 0.65\alpha^6$, and μ'_{LUT} interpolating between 30 optimised values.

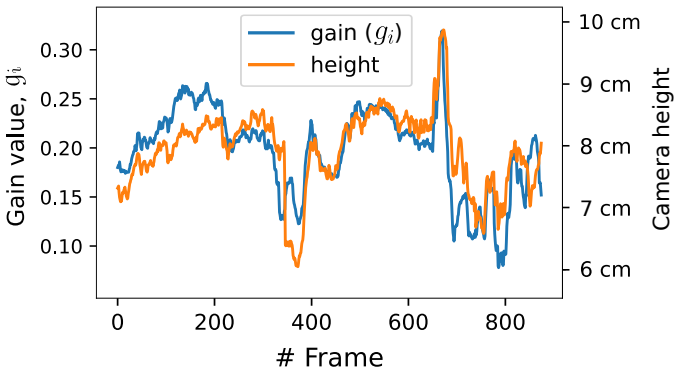


Figure 3.5: Estimated auto-gain factors have a continuous progression over the 875 images used in calibration. In addition, we see a great correlation when we compare it with camera height.

3.3.2 Camera response function and auto-gain

Although we use $\gamma = 1.0$ as initial value, the optimisation converges most of the time to $\gamma = 2.2$ which is the most used value.

Automatic gain cannot be evaluated with test data, because each image has a different gain factor. Instead, in Figure 3.5, we can see that the estimated gain factor follows a continuous progression over time along a calibration sequence.

Moreover, gain value for each frame seems to be closely related to the distance from camera to the illuminated surface. That is, when the camera is closer to the pattern, light is more intense, and the endoscope applies a lower gain value.

In addition, during calibration, we are working at distances between 6 and 10 cm. As we will see later, this depth range is like that found in endoscopy.

3.3.3 Lambertian vs. non-parametric BRDF

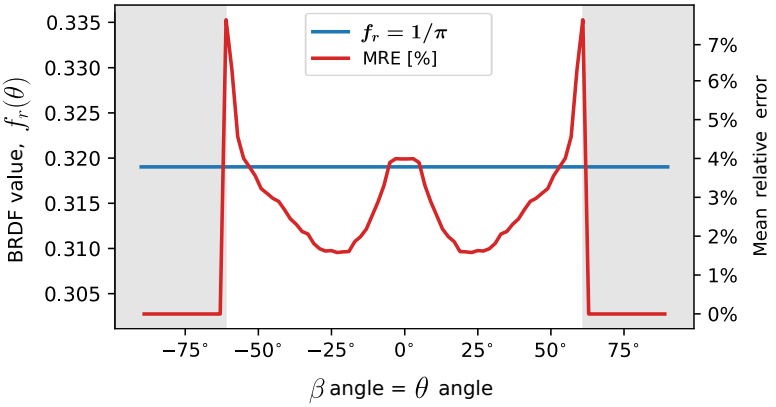
Using a perfectly diffuse BRDF is a common approach in photometric problems. Following this idea, we set the BRDF of our illuminated surface to a Lambertian model k_d/π (see Chapter 2), and we set surface albedo to one ($k_d = 1$) since we sample white areas on the Vicalib pattern.

However, as we can see in Figure 3.6a, this approach causes erroneous predictions near the perfect specular direction, i.e. when $\theta \approx 0$. This shows that the paper used to print the Vicalib pattern does not fit the perfect diffuse Lambertian model. In order to avoid this causing an erroneous calibration, we use a generic non-parametric BRDF model (Figure 3.6b) as discussed in Section 3.2.1.

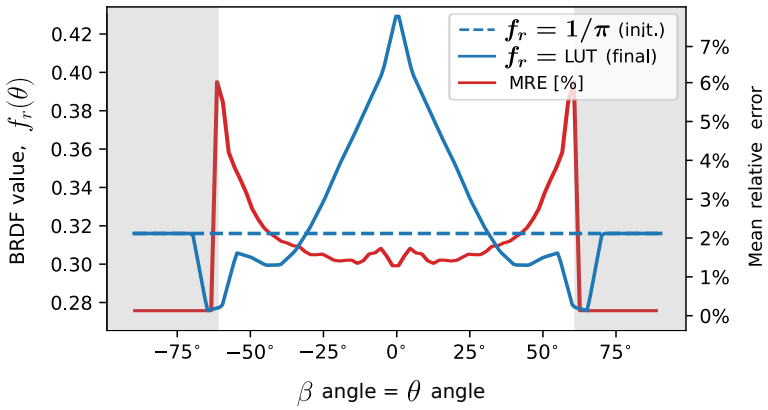
When we assumed a Lambertian BRDF, the distribution of errors was positively skewed (Figure 3.7c), due to unexpected specular reflection. The optimisation increased

the predicted gain for the whole frame, causing a larger number of positive errors (overestimation of intensity, seen in the region from 0 to 5 of Figure 3.7c). Plus, there were a few very negative errors (underestimation, as seen in range -15 to -5) caused by pixels with specular reflection.

With a new BRDF for the paper, our calibration corrects that bias. The result allows us to predict the grey level of a pixel with a standard deviation of 3.2 levels. Moreover, the new estimated BRDF is an isolated component of the model. For example, when we want to apply our calibration in the interior of the human body, we can replace this BRDF by that of the human intestine, and the rest of the calibrated parameters remain valid.

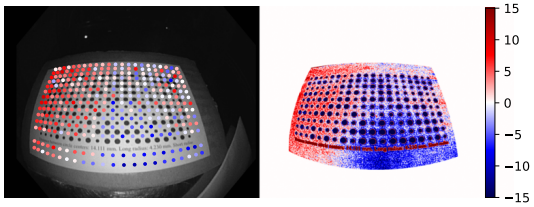


(a) Lambertian BRDF

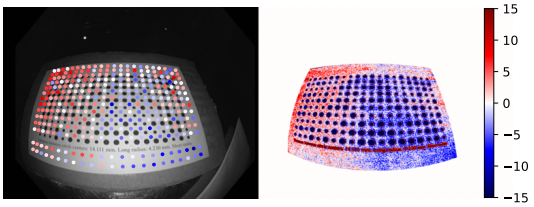


(b) Non-parametric BRDF

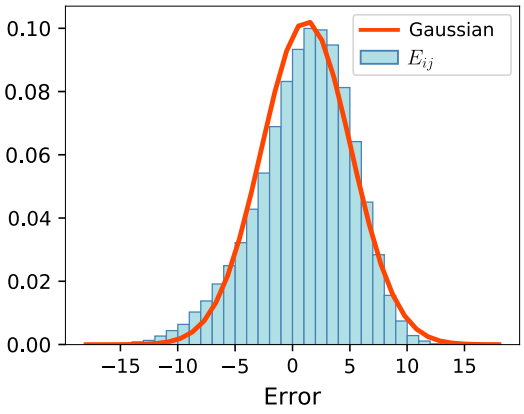
Figure 3.6: (a) Fixed Lambertian BRDF accumulates error when $\theta \approx 0$. (b) Our non-parametric BRDF model converges to a Phong-like distribution that eliminates such an error. For this experiment, we choose fifteen F_r values.



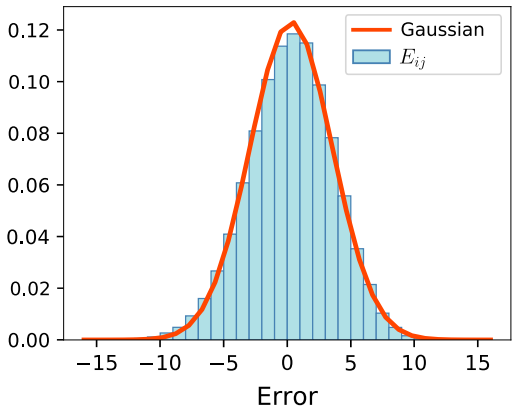
(a) Lambertian BRDF (Frame 128 of 193)



(b) Non-parametric BRDF (Frame 128 of 193)



(c) $\boldsymbol{\mu = 1.2}$, $\sigma = 3.9$ (All frames)



(d) $\boldsymbol{\mu = 0.3}$, $\sigma = 3.2$ (All frames)

Figure 3.7: (a) Photometric error (E_{ij}, in grey levels $0 - 255$) using a perfectly diffuse Lambertian BRDF is higher in the areas where $\theta \approx 0$. This causes (c) bias in E_{ij} errors on test frames. (b) Our non-parametric BRDF reduces overall error, and (d) eliminates calibration bias.

Chapter 4

Depth estimation

Given the endoscope photometric model and a single endoscope image, our goal is to estimate the true-scale depth and surface normal for each imaged point. To achieve this objective, we consider the following assumptions:

Diffuse surface. Human colon tissue can be approximated by a Lambertian material, if specular highlights are masked or treated as spurious. In addition, the surface albedo is considerably constant throughout the colon.

Known auto-gain. We know the capture device's automatic gain control, having information on the applied gain at each frame. Otherwise, real scale would not be observable from a single frame, obtaining an up-to-scale depth map.

Local planarity. Surfaces are assumed to change smoothly, except at occasional discontinuities. This allows us to approximate differential changes of the surface by the corresponding tangent plane.

4.1 Depth map estimation

Based on DTAM method proposed by Newcombe et al. [15], we approach the estimation of a depth map as an optimisation problem, that minimises an energy function:

$$E_{\boldsymbol{\xi}} = \int_{\Omega} \Big\{ R(\mathbf{u}, \boldsymbol{\xi})) + \lambda C(\mathbf{u}, \boldsymbol{\xi}(\mathbf{u})) \Big\} d\mathbf{u} \tag{4.1}$$

where

- $\mathbf{u} \in \Omega$ are coordinates on the image.
- $\boldsymbol{\xi} : \Omega \to \mathbb{R}$ is the depth map (Section 4.1.4).
- $C(\,)$ is a photometric cost function (Section 4.1.1).

- $R(\,)$ is a regularisation cost (Section 4.1.3).
- $\lambda \in \mathbb{R}^+$ is a hyperparameter that adjusts the regularisation weight.

4.1.1 Photometric cost function

DTAM assumes global illumination of the scene. Thus, Newcombe et al. use a cost function based solely on camera geometry and brightness constancy.

Unlike in DTAM, the illumination during endoscopy varies with camera movement. Consequently, we replace the original cost function with a novel cost function based on our photometric endoscope model:

$$C(\mathbf{u}, d) = \rho\left(I(\mathbf{u}) - \mathcal{I}\left(\pi^{-1}\left(\mathbf{u}, d\right)\right)\right) \tag{4.2}$$

where

- $\pi^{-1}(\,)$ is the camera unprojection model, e.g. Kannala & Brandt.
- $\mathcal{I}(\,)$ is our calibrated endoscope photometric model.
- $I : \Omega \to \mathbb{R}^+$ denotes the intensity of the pixels.
- $\rho(\,)$ is a robust cost function.

4.1.2 Normal estimation from a depth map

As seen in Chapter 2, photometry of a scene $\mathcal{I}$ is influenced by both the distance to the points (spotlight spherical attenuation) and the surface normal (cosine term). However, surface normal is directly related to depth variations. Therefore, both parameters should not be optimised separately. Instead, given the local planarity assumption, we can calculate the normal of a point from the estimated depth map [8].

Thanks to this relationship, we keep the depth map as the only unknown variable of the problem. However, it should be noted that this method is influenced by spurious data, especially at surface discontinuities. Therefore, in such areas, we can expect greater difficulties in reaching an optimal solution.

4.1.3 Smoothness regularisation

The defined cost function is trying to find three unknowns per pixel (z, n_θ, n_φ) from only one intensity measurement (I). Newcombe et al. propose a regularisation term

$$R(\mathbf{u}, \boldsymbol{\xi}) = g(\mathbf{u}) \|\nabla \boldsymbol{\xi}(\mathbf{u})\|_\epsilon, \tag{4.3}$$

to solve problem's ill-posedness. With this, they penalise local depth variations, except at points where the luminosity gradient is large, which usually correspond to surface discontinuities. Thus, $\|x\|_\epsilon$ Huber norm with $\epsilon \approx 1.0e^{-4}$ works as total variation (TV) regulariser and $g(\mathbf{u})$ reduces the regularisation strength at high gradient points

$$g(\mathbf{u}) = e^{-\alpha \|\nabla I(\mathbf{u})\|_2^\beta}. \tag{4.4}$$

Both gradients, $\nabla\boldsymbol{\xi}$ and ∇I, can be computed by 2D filtering with classic first difference, Sobel kernels, and others. Thanks to this regulariser, (z, n_θ, n_φ) are now constrained by pixel's neighbourhood.

4.1.4 Depth map representation

Newcombe et al. formulate their depth map function as

$$\boldsymbol{\xi}_{1/z} \stackrel{\text{def}}{=} \frac{1}{z}, \tag{4.5}$$

known as inverse Z-depth. This decision is appropriate for their multi-view based problem, as the pinhole projection model depends directly on this variable.

Instead, we are faced with a single-view problem. In our case, the photometry is quadratically dependent on the inverse of the Euclidean distance

$$\mathcal{I} \propto \frac{1}{d^2}, \quad \text{with} \quad d = \|\mathbf{x}_\mathcal{W} - \mathbf{t}_{\mathcal{WC}}\|. \tag{4.6}$$

Therefore, we will consider two variations of the depth map, such that

$$\boldsymbol{\xi}_d \stackrel{\text{def}}{=} d, \qquad \boldsymbol{\xi}_{1/d} \stackrel{\text{def}}{=} \frac{1}{d}. \tag{4.7}$$

4.2 Initial solution

We make the optimisation method start from an initial solution, where we assume all surface normal vectors pointing towards the camera optical centre.

From the calibrated photometric model, seen in Equation (3.4), we can revert the effects of light spread function, diffuse Lambertian BRDF, as well as known camera gain and gamma correction. Thus, we get

$$I_c(\mathbf{u}) = \frac{I^\gamma}{\mu'(\mathbf{x}) \cdot f_r(\theta) \cdot g_i} = \frac{\cos\theta}{d^2}, \tag{4.8}$$

where I_c is a *canonical intensity value*, which is obtained after compensating all mentioned parameters that influence image formation.

Note that, when a surface normal points towards the camera, the θ angle is zero. Therefore, by solving for d in the above equation, we get an initial solution

$$\boldsymbol{\xi}_o(\mathbf{u}) = I_c(\mathbf{u})^{-1/2}. \tag{4.9}$$

The closer the actual θ is to zero, the closer this initial solution is to the real depth.

Chapter 5

Experimental validation

In this chapter, we validate the real-scale 3D reconstruction algorithm. However, there is no public dataset that provide a useful depth ground truth of real colonoscopy procedures. Therefore, the quantitative validation is carried out with synthetic data.

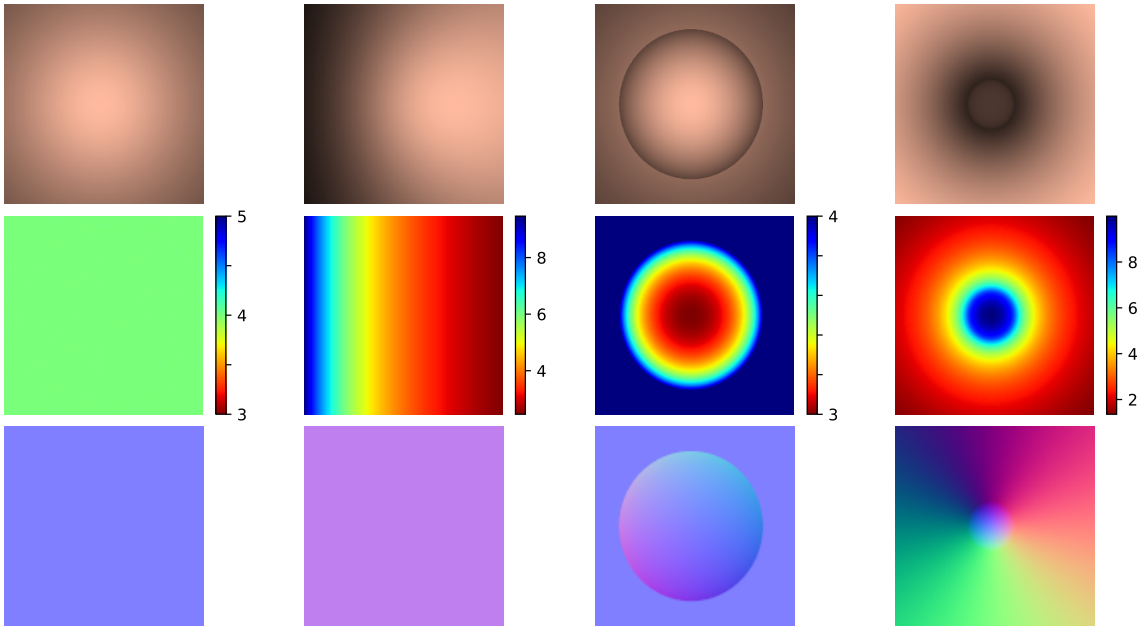
5.1 Simple geometry dataset

In this project, we synthesise a set of scenes, based on simple geometries. The characteristics of the data are presented below, together with a discussion of the results obtained.

Dataset specifications. Our dataset is composed of four scenes, of incremental complexity. Each scene consists of a 475×475 px image, which simulates the camera capture with a *field of view* (FOV) of 92°. We also provide ground truth depth and normal maps, as shown in Figure 5.1. Furthermore, knowing the endoscope's geometry and photometry, we would be able to correct for distortion and attenuation. Thus, the synthesised images constitute *canonical frames*, where distortion and attenuation have already been compensated.

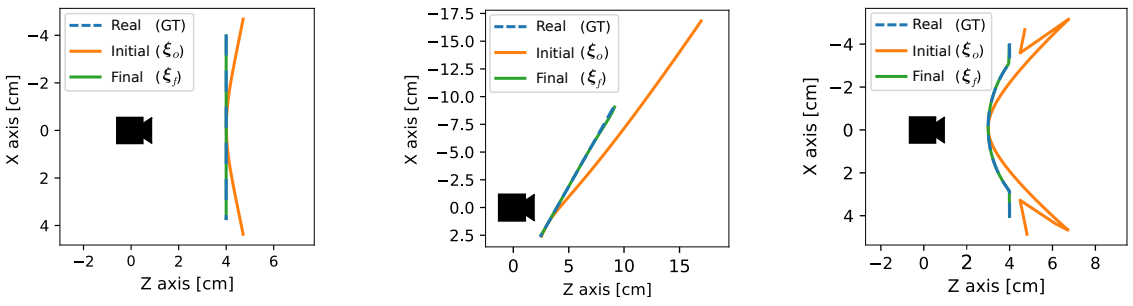
Scene details. In increasing complexity. *Scene 00* consists of an infinite plane parallel to the image plane, located at 4 cm on Z-axis. *Scene 01* is like the previous one, but rotated 18° around Y-axis, giving depths from 2 to 10 cm. *Scene 02* contains a spherical cap, placed on a plane, with a depth range of 3 to 4 cm. Finally, *Scene 03* attempts to simulate the endoluminal cavity by means of a cylinder ending in a hemisphere, with a total length of 10 cm.

Method convergence. It depends on the scene complexity and the initial solution optimality (Section 4.2). In scenes 00 to 02, all conducted experiments achieved a



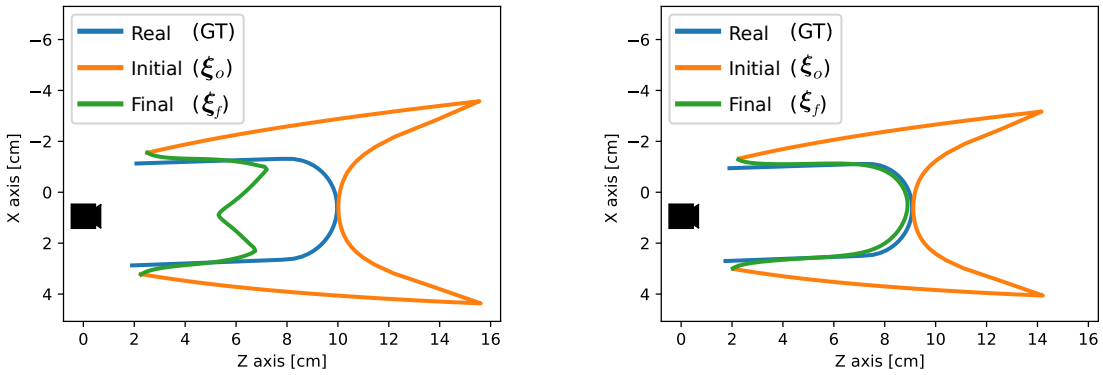
(a) Scene 00 (b) Scene 01 (c) Scene 02 (d) Scene 03

Figure 5.1: Data generated for our simple geometry dataset. **Top:** Canonical image, simulated with our light transport model, coloured brown to improve visibility in this report. **Middle:** Z-depth map (GT) expressed in centimetres. **Bottom:** Normal map (GT), represented in colour space $(R, G, B) = ([n_x, n_y, n_z] + 1)/2$.



(a) Scene 00 (b) Scene 01 (c) Scene 02

Figure 5.2: Slice along Y-axis (middle of the image) of actual surface (GT), initial solution ($\boldsymbol{\xi}_o$) and final optimised estimation ($\boldsymbol{\xi}_f$). Our prediction is very accurately matched to the ground-truth.



(a) Using $\nabla\boldsymbol{\xi}$ regulariser

(b) Using $\nabla^2\boldsymbol{\xi}$ regulariser

Figure 5.3: Difference in convergence of first and second derivative regularisers, after 1 100 iterations of optimisation. (b) converges faster than (a), but both solutions reach similar results as the number of iterations increases.

solution very close to ground truth. Figure 5.2 compares depth ground truth with the initial and final solutions.

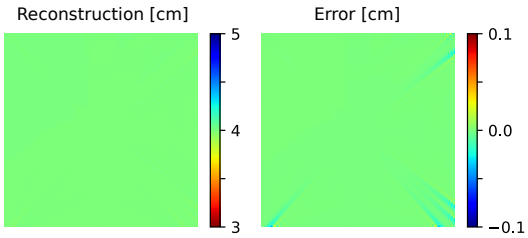
As for Scene 03, regularisation term imposes smooth changes in the surface depth. However, the cylindrical geometry of this scene causes large depth gradients at the centre of the image. Therefore, we have compared $\nabla\boldsymbol{\xi}$ first derivative regulariser, proposed by DTAM, with a $\nabla^2\boldsymbol{\xi}$ second derivative regulariser. The latter imposes constant changes in depth. This results in a faster and better convergence (Figure 5.3).

Reconstruction accuracy. Our method achieves millimetre accuracies, ranging from sub-millimetre errors in the simplest scene to 2 mm errors in the most complex one. Table 5.1 presents quantitative results for each experiment. Overall, regularising with second derivative term provides better results. When we change the variable we are optimising, $\boldsymbol{\xi}$, convergence is faster, while accuracy remains unchanged. In Figure 5.4 we show predicted depth maps for our best variant.

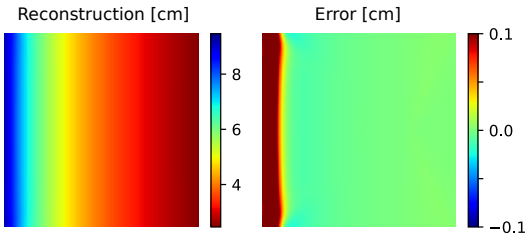
5.2 Simulated colon dataset

Within the EndoMapper project, a more complex dataset [19] is being developed, based on a CT scan of a real human intestine. From this 3D model, they use Unity framework to simulate image acquisition during an endoscopic exploration procedure.

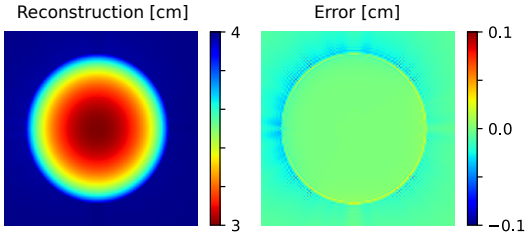
We have tested our model on a scene in the new dataset. Below we present the characteristics of the data provided and the results we obtained.



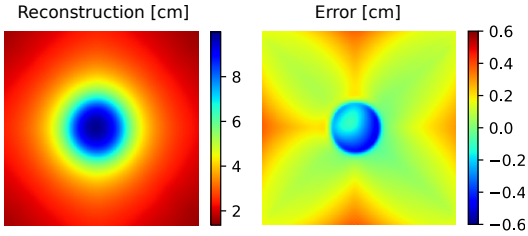
(a) Scene 00



(b) Scene 01



(c) Scene 02

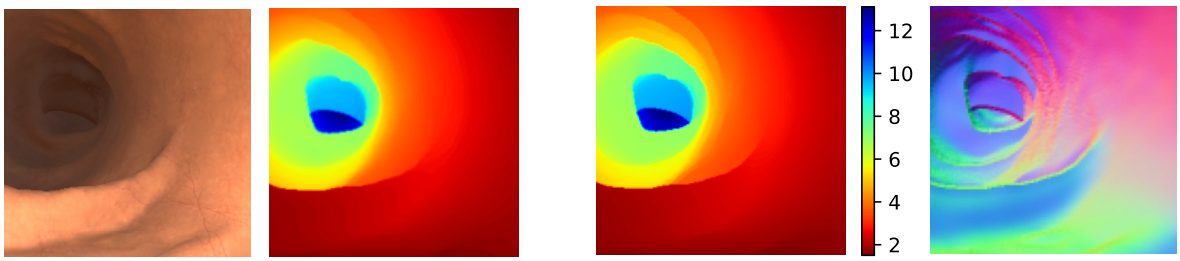


(d) Scene 03

Figure 5.4: Reconstruction results on simple geometry dataset. For each scene, we present the variant located in the last row of Table 5.1.

Scene	ξ	Reg.	# iter.	Depth error [cm]		Depth error [%]		Normals error [deg]	
				Mean	Median	Mean	Median	Mean	Median
00	$1/z$	∇	71	<0.01	<0.01	<0.01	<0.01	0.34	<0.01
01	$1/z$	∇	1 148	0.10	**<0.01**	0.90	**0.04**	1.19	0.20
		∇^2	**73**	**0.03**	**<0.01**	**0.32**	0.09	**0.62**	**0.18**
02	$1/z$	∇	102	0.02	0.02	0.35	0.37	1.00	0.50
		∇^2	**64**	**0.01**	**0.01**	**0.25**	**0.21**	**0.95**	**0.39**
03	$1/z$	∇	>5 550	0.30	0.21	7.30	6.73	12.17	9.44
		∇^2	>1 500	**0.19**	**0.18**	5.83	5.31	**11.07**	**8.00**
	d	∇^2	301	**0.19**	**0.18**	5.85	**4.99**	12.92	9.15
	$1/d$	∇^2	**78**	**0.19**	**0.18**	**5.78**	5.21	11.55	8.30

Table 5.1: Reconstruction accuracy on the simple geometry dataset.



(a) Frame 0000 (b) Depth (GT) (c) Filtered [cm] (d) Normals ($\hat{\text{GT}}$)

Figure 5.5: Simulated colon dataset [19] within EndoMapper project. (a) Image provided by original dataset. (b) Ground truth depth information using 8-bit representation. (c) Filtered depth ground truth, using a bilateral Gaussian filter with $\sigma_s = 7$ and $\sigma_c = 0.1$ cm. (d) Normals estimated from filtered depth map.

Dataset specifications. One image of the dataset has a resolution of 475×475 px, generated from a pinhole projection model. Unity camera's field of view (FOV) has been reduced to 55°. Moreover, the intensity of each pixel is similar to the canonical intensity defined in this work (Section 4.2). These values differ only in albedo and gain factor. However, albedo is known, leaving the gain of each frame as the only unknown. Yet, the latter can be estimated based on depth ground-truth.

However, depth ground-truth is encoded as Z-depth in 8-bit images, obtaining an insufficient resolution for the estimation of normals (Section 4.1.2). We try to solve this problem by applying a Gaussian filter on the depth map. Thus, we obtain an acceptable approximation of the ground-truth normals, except at surface discontinuities, as we show in Figure 5.5.

Scene details. In the presented frame, the camera points in the longitudinal direction of the bowel. Here, we found depths from 2 to 12 cm. In addition, bowel's haustra cause occlusions and discontinuities, larger than those seen in the previous simple dataset.

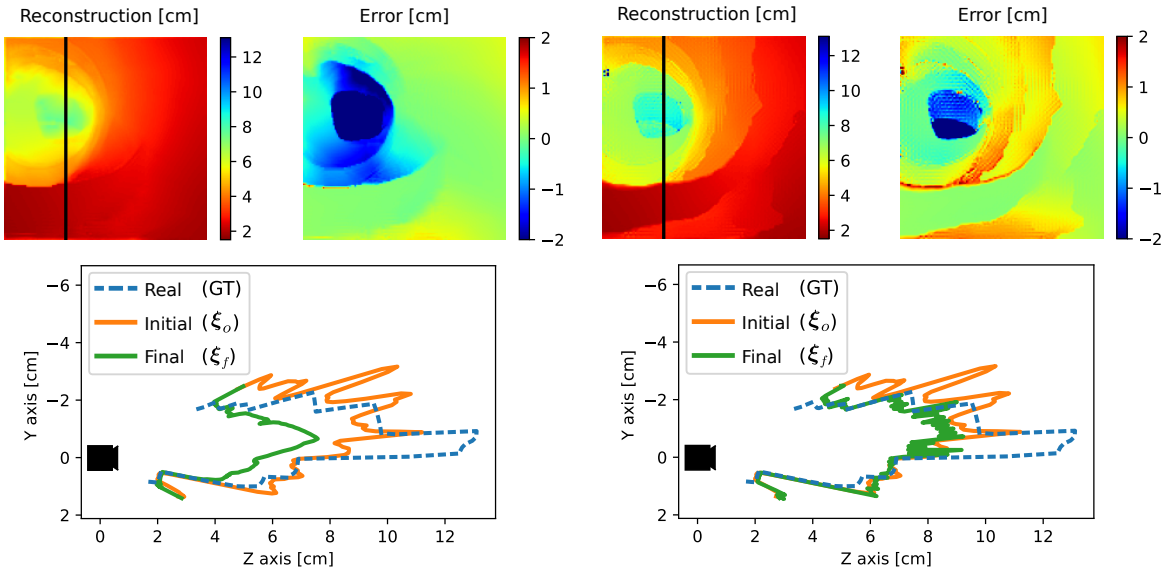
Method convergence. As in the previous section, we compare the method's convergence using first and second derivative regularisation (Figure 5.6). Here, second derivative is more affected by roughness of the bowel. Thus, we obtain better results with the original DTAM regulariser.

On the other hand, in Table 5.2 we study differences in convergence between three $\boldsymbol{\xi}$ variants, paying attention to the reconstruction error after different number of iterations. Our two approaches improve convergence and accuracy over the DTAM method. However, these results are not as conclusive as in the simple dataset. Due to reduced camera FOV, all variants of $\boldsymbol{\xi}$ achieve similar results.

Reconstruction accuracy. Reconstruction obtained in this dataset is noisier and normal estimation is therefore worse (Figure 5.7).

$\boldsymbol{\xi}$	**Reg.**	**# iter.**	**Depth error [cm]**		**Depth error [%]**		**Normals error [deg]**	
			Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
$1/z$	∇	44 500	0.53	0.23	10.60	7.41	30.63	23.77
	∇^2	8 900	0.51	0.40	15.09	11.86	36.06	29.26
d	∇	20 000	0.33	**0.16**	7.90	**4.98**	**26.21**	**18.75**
	∇^2	44 500	0.38	0.20	9.57	6.37	32.00	23.56
$1/d$	∇	44 500	**0.28**	**0.16**	**7.32**	5.01	27.89	19.69
	∇^2	**5 400**	0.51	0.40	15.16	12.05	35.86	28.89

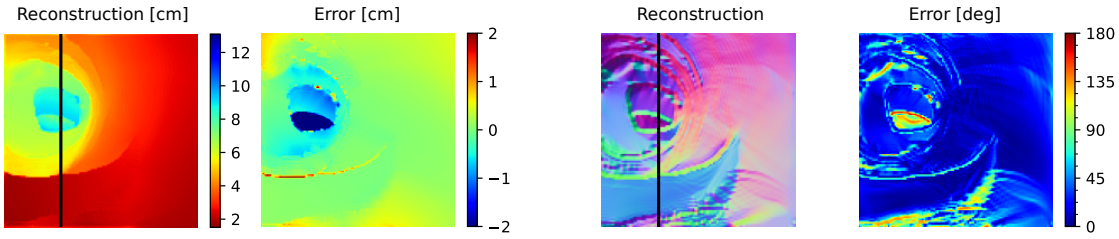
Table 5.2: Reconstruction accuracy on the simulated colon dataset.



(a) Using $\nabla\boldsymbol{\xi}$ regulariser

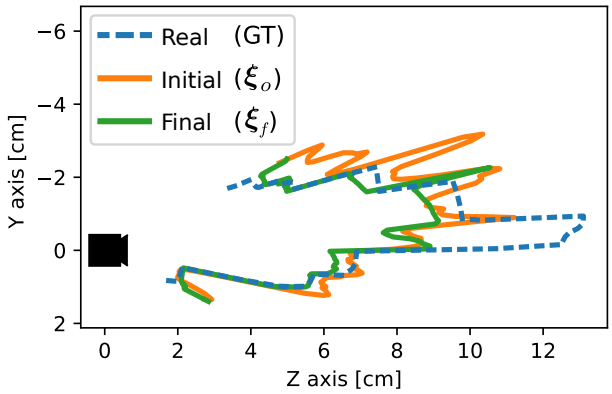
(b) Using $\nabla^2\boldsymbol{\xi}$ regulariser

Figure 5.6: Estimated Z-depth maps and their corresponding errors with respect to Gaussian-filtered ground truth depth. The second derivative regulariser (b) introduces greater error than the original in DTAM (a). This is also visible in cross-section, corresponding to the line marked in black.



(a) Depth accuracy

(b) Normal accuracy



(c) Cross-section along black line

Figure 5.7: Estimated Z-depth, normal map and corresponding errors w.r.t filtered real depth (GT) for $\boldsymbol{\xi}_{1/d}$ and $\nabla\boldsymbol{\xi}$ variant. In (c) we show a cross-section of the reconstruction, corresponding to the line marked on (a) and (b). Here we compare the ground truth with the initial and final solutions.

In addition, Unity's renderer, designed for video games, introduces fog effect in areas far away from the camera. This increases intensity of distant pixels and our initial solution, in Figure 5.7c, presents a big difference w.r.t. ground-truth. As a result, we cannot reconstruct the deepest part of the colon, from 10 to 12 cm.

Moreover, the dataset was generated using Unity's global illumination model. This takes into account second and successive light bounces through the scene. Together with regularisation term, it causes that the reconstruction shown in Figure 5.7c underestimates depth for points farther than 6 cm.

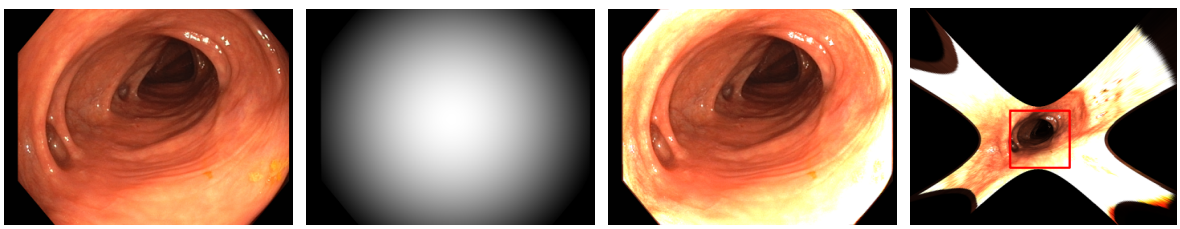
We conclude that depth accuracy of our reconstruction is 0.28 cm on average. However, median error, 0.16 cm, is significantly lower, because it is not influenced by the deeper points, which could not be reconstructed.

5.3 Real colon dataset

Finally, we carried out a qualitative validation on real colonoscopy images provided by EndoMapper project. This allows us to apply both photometric and geometric calibration, as well as to visually check the depth estimation method.

Dataset specification. HCULB dataset consists of real gastroscopy and colonoscopy sequences, recorded at the Hospital Clínico Universitario Lozano Blesa in Zaragoza. During the procedures, Olympus EVIS EXERA III GIF-H190 endoscopes are used, which we already calibrated in Chapter 3. It provides a resolution of 1440 $\times$ 1080 px with a FOV of 140°. However, only images are supplied, with no ground-truth data.

Scene details. For this experiment, we use frame 73880 of HCULB_00039 sequence, as shown in Figure 5.8a. This scene is similar to the generated colon dataset, including occlusions and discontinuities. However, colon's metric scale is unknown.



(a) Frame 073880 (b) Vignetting & $\boldsymbol{\mu}(\mathbf{x})$ (c) Canonical (d) Undistorted

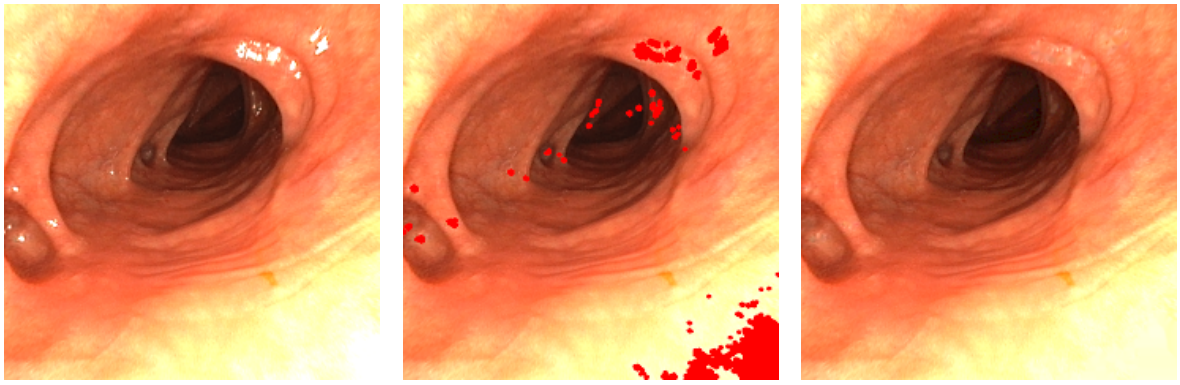
Figure 5.8: HCULB dataset by EndoMapper project. (a) We select one frame of sequence HCULB_00039 and (b) compensate for vignetting and light spread function. (c) Then, we obtain a canonical frame, that can be (d) undistorted and cropped.

Photometric compensation. First, we compensate for the endoscope's vignetting and light spread function (Figure 5.8b) as well as gamma correction. Thus, we obtain a canonical frame (Figure 5.8c) with unknown gain.

Geometric undistortion. Next, we correct distortion caused by endoscope lenses. This way, we get an equivalent pinhole frame (Figure 5.8d), as in previous quantitative evaluation. The obtained image is then cropped to avoid areas where distortion is higher. After cropping, FOV is reduced to 92°.

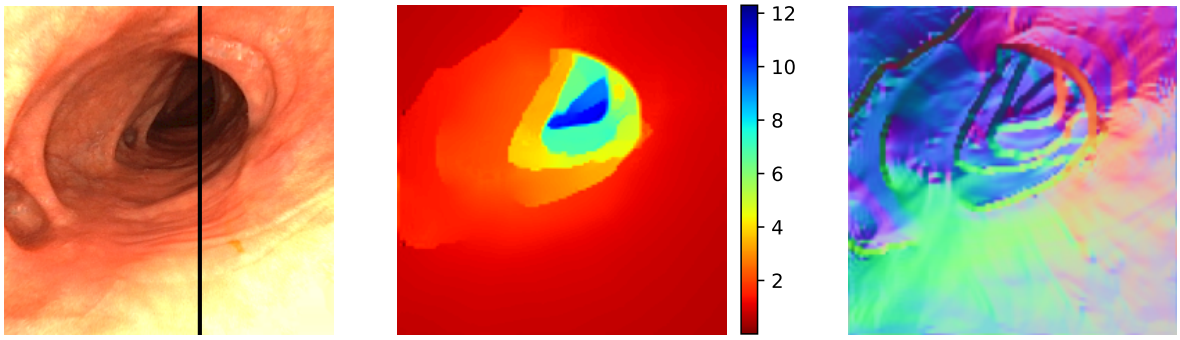
Highlight inpainting. In contrast to synthetic data, specular reflections occur in real images. We apply a brightness threshold to mask these pixels (Figure 5.9b). Our threshold selects pixels that are 75% brighter than their neighbours and those that exceed 225 out of 255 brightness value. The first criterion helps us to detect small and distant reflections. While second criterion filters out large and near reflections. Finally, mask is dilated to improve recall. On these pixels, we apply inpainting (Figure 5.9c) based on the algorithm proposed by Telea [21].

Reconstruction results. Qualitatively, in Figure 5.10, the result obtained maintains relative depth between far and near zones. In addition, main surface discontinuities at the haustra are visible in the depth map, although other minor details are smoothed out. Still, the predicted scale is arbitrary, as we do not have data about the camera's automatic gain.

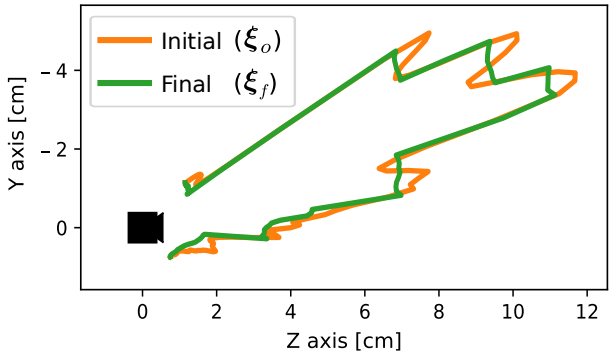


(a) Original frame (b) Highlight mask (c) Inpainted frame

Figure 5.9: Automatic highlight detection and inpainting.



(a) Input frame (b) Predicted depth [cm] (c) Predicted normals



(d) Cross-section along black line

Figure 5.10: Qualitative results on the HCULB_00039 sequence, based on a real image provided by EndoMapper project. (b) With $\boldsymbol{\xi}_d$ and $\nabla\boldsymbol{\xi}$, we obtain an up-to-scale depth map, with good contour details. (c) This depth map allows us to estimate a smooth normal map. (d) Cross-section showing the reconstruction of several haustra.

Chapter 6

Conclusion

6.1 Summary

In this MEng thesis we have proposed a solution for real-scale monocular 3D reconstruction. This allows us to estimate the metric scale in reconstructions of the human colon, where other methods based on stereo systems or self-supervised deep learning are currently not possible.

To do this, we have geometrically and photometrically modelled a real monocular endoscope, considering emission, transport and capture of light. This has allowed us to calibrate endoscope's illumination and camera, being able to model the image formation with an error of 0.3 ± 3.2 grey levels.

With the calibration of endoscope's light-camera pair, we are able to reconstruct a real-scale 3D model of the interior of the human body. Thus, it overcomes the limitations of conventional monocular systems and recovers metric depth of the scene. Such 3D reconstruction can be carried out from a single image, when we know the reflectance function of a surface and the camera gain.

We have validated our model on three different datasets. Two of them are synthetic datasets for which we have ground-truth depth. In the second synthetic dataset, our method obtains dense depth maps at real scale with an error of less than 7%, i.e. around 3 mm, even for complex bowel's geometry. In addition, this error is reduced to less than 1.6 mm in closer areas.

Finally, with a real colon dataset, we have qualitatively shown that our method can be applied on real colonoscopy images, obtaining dense maps of the human colon. These results open the door to SLAM systems capable of reconstructing dense, accurate maps of the interior of the human body. Our method can provide more reliable information about the scale and deformations of the environment, where using stereo cameras is not possible.

6.2 Future work

This Master's thesis is a research project prior to a PhD thesis to be carried out at University of Zaragoza with an FPU grant awarded by the Spanish Ministry of Universities. Future work will aim to improve and complete the presented method by:

Fisheye depth estimation. In the conducted experiments, the depth map was calculated on a pinhole image. In the future, we will also incorporate the fisheye model during depth prediction, preserving the full field of view of the endoscope camera.

Integration with SLAM methods. Our work shows that there is potential in integrating the proposed photometric model with a geometric SLAM system. We propose to investigate a hybrid optimisation scheme capable of simultaneously exploiting both the geometric reprojection error and the photometric error.

Such systems are not based on a single frame, but reconstruct a complete scene map from a sequence of images. Thus, we will analyse the benefits of this new multi-view approach, compared to our single-view model. There is hope that joint optimisation will eliminate the scale drift suffered by monocular SLAM systems, especially in the deformable human colon, where this problem is exacerbated.

Integration with deep learning methods. On the other hand, neural networks have been proven to be effective for depth prediction task. However, so far, real-scale depth prediction requires supervised learning or self-supervised training with stereo data. But such images are not available from the human colon. Therefore, our model could be used together with self-supervised learning techniques, to get neural networks that predict real scale, despite being trained only on monocular images.

New colon datasets. Within and outside the EndoMapper project, all methods proposed to solve 3D reconstruction inside the human colon lack a complete and accurate dataset against which to evaluate and compare their accuracy. Our two lines of research mentioned above, SLAM and deep learning methods, would benefit from a real dataset, with ground-truth data, for quantitative validation.

Bibliography

[1] Autonomous Robotics & Perception Group (ARPG). *VICalib visual-inertial calibration suite*. 2016. URL: `https://github.com/arpg/vicalib`.

[2] Carlos Campos et al. "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM". In: *IEEE Transactions on Robotics* (2021).

[3] Peter Comninos. *Mathematical and computer programming techniques for computer graphics*. Springer Science & Business Media, 2010.

[4] Jakob Engel, Vladlen Koltun, and Daniel Cremers. "Direct Sparse Odometry". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.3 (2018), pp. 611–625.

[5] Jakob Engel, Vladyslav Usenko, and Daniel Cremers. "A photometrically calibrated benchmark for monocular visual odometry". In: *arXiv preprint arXiv:1607.02555* (2016).

[6] Clément Godard et al. "Digging into self-supervised monocular depth estimation". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 3828–3838.

[7] Yang Hao et al. "Photometric stereo-based depth map reconstruction for monocular capsule endoscopy". In: *Sensors* 20.18 (2020), p. 5403.

[8] Stefan Hinterstoisser et al. "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2011, pp. 858–865.

[9] Juho Kannala and Sami S Brandt. "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.8 (2006), pp. 1335–1340.

[10] Johann Heinrich Lambert. *Photometria sive de mensure de gratibus luminis, colorum umbrae*. Eberhard Klett, 1760.

[11] Víctor Martínez Batlle and Juan Domingo Tardós Solano. "Scale estimation in monocular ORB-SLAM2 using deep convolutional networks". BA thesis. University of Zaragoza, 2020.

[12] Richard Modrzejewski et al. "Light modelling and calibration in laparoscopy". In: *International Journal of Computer Assisted Radiology and Surgery* 15.5 (2020), pp. 859–866.

[13] Rosana Montes and Carlos Ureña. "An overview of BRDF models". In: *University of Grenada, Technical Report LSI-2012-001* (2012).

[14] Raul Mur-Artal and Juan D Tardós. "ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras". In: *IEEE Transactions on Robotics* 33.5 (2017), pp. 1255–1262.

[15] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. "DTAM: Dense tracking and mapping in real-time". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2011, pp. 2320–2327.

[16] Takayuki Okatani and Koichiro Deguchi. "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center". In: *Computer vision and image understanding* 66.2 (1997), pp. 119–131.

[17] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016.

[18] Bui Tuong Phong. "Illumination for computer generated pictures". In: *Communications of the ACM* 18.6 (1975), pp. 311–317.

[19] Anita Rau et al. "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy". In: *International journal of computer assisted radiology and surgery* 14.7 (2019), pp. 1167–1176.

[20] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[21] Alexandru Telea. "An Image Inpainting Technique Based on the Fast Marching Method". In: *Journal of Graphics Tools* 9.1 (2004), pp. 23–34.

[22] Lokender Tiwari et al. "Pseudo rgb-d for self-improving monocular slam and depth prediction". In: *European Conference on Computer Vision*. Springer. 2020, pp. 437–455.

[23] Chenyu Wu, Srinivasa G Narasimhan, and Branislav Jaramaz. "A multi-image shape-from-shading framework for near-lighting perspective endoscopes". In: *International Journal of Computer Vision* 86.2 (2010), pp. 211–228.

[24] Nan Yang et al. "D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 1278–1289.

Appendix A

Project management

This project was carried out with a grant from the European EndoMapper Project. The dedication to the project has been 8 hours per day (40 per week) for a total of 5 months. Table A.1 shows the total number of hours spent on each project's tasks. Figure A.1 shows the time schedule of the project.

Task	*Time*
Workspace configuration	10 h
Research and literature review	40 h
Endoscope model definition	55 h
Geometric camera calibration	100 h
Photometric camera calibration	190 h
Depth estimation algorithm	280 h
Project report	95 h
Co-ordination and meetings	30 h
Total	800 h

Table A.1: Time spent on each project task.

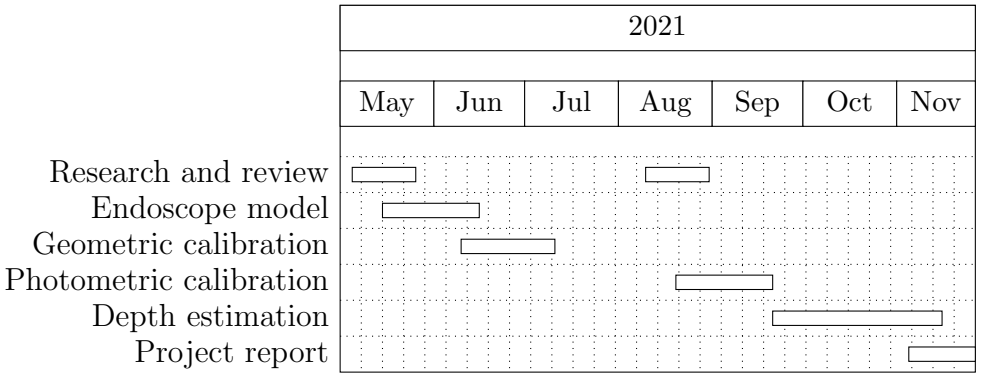


Figure A.1: Gantt chart of project execution.