IMAGE-BASED MOTION ESTIMATION FOR TELEOPERATED FLEXIBLE ENDOSCOPES

By

Charreau S. Bell

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE
in
MECHANICAL ENGINEERING
December, 2014
Nashville, Tennessee

Approved:

Pietro Valdastri, Ph.D.

Nilanjan Sarkar, Ph.D.

D. Mitchell Wilkes, Ph.D.

To my parents, without whom this work would not
have been remotely possible,

and

To my brothers, who encouraged me to continue,

and

To Chickpea, my ever-present, most loyal friend.

ACKNOWLEDGEMENTS

I would first like to acknowledge and thank my advisor, Dr. Pietro Valdastri, and thank him for all his guidance and support in this endeavor. Without him, I would not have had this opportunity, and am extremely grateful to have worked on this project and acquired so much knowledge and expertise from this experience. I would also like to thank my committee members, Dr. Mitch Wilkes and Dr. Nilanjan Sarkar for their efforts in helping me to complete this work.

Additionally, I'd like to acknowledge my collaborators on this work from the University of Texas-Arlington, Dr. Gian-Luca Mariottini and Gustavo Puerto. Their input, suggestions, and contributions were essential for creating the best work possible.

I would also like to thank my labmates and friends Trevor, Robert, Marco, and Arnetta for their help and support in this project, from just providing suggestions and moral support to helping out with the electronics for testing the algorithm.

Lastly, I'd like to thank the Center for Compact and Efficient Fluid Power under the National Science Foundation (NSF) (Grant #0540834), Vanderbilt Initiative in Surgery and Engineering (VISE), and the Broad Medical Research Program of the Broad Foundation for their financial support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Colonoscopy and Colorectal Cancer

Colorectal cancer is the fourth leading cause of cancer worldwide, claiming more than 600,000 deaths annually [1]. Despite accessibility and awareness of the importance of the procedure, the western world accounts for more than 63% of these deaths [2]. Colorectal cancer progression follows a trajectory similar to that of other cancers, generally reaching malignancy in 5 to 10 years. However, colorectal cancer possesses the unique quality that if the tumor is detected at an early stage, the prognosis for survival is 90%; however, if detected too late, the chances of survival decrease to 5% [3]. This emphasizes the importance of compliance to recommended screening guidelines for populations at risk (i.e., people with a family history of colorectal cancer or older than 50 years of age), regardless of whether symptoms are observed.

The traditional method for diagnostic and therapeutic assessment of colorectal cancer is through colonoscopy, an endoscopic procedure in which a 1.5m-long flexible tube is inserted through the anus. The endoscopist is able to visualize the colon lumen using the endoscopic camera and on-board illumination provided by the endoscope. In this manner, polyps are able to be detected and removed. White light illumination (WLI) is the illumination modality used during standard colonoscopy; however, WLI can fail to reveal important characteristics of the colon wall [4]. Even experienced endoscopists can miss up to 30% of all potential cancer lesions when using standard WLI [5].

In the last decade, narrow band imaging (NBI) has been introduced to augment diagnostic capabilities during endoscopy. NBI employs filters to narrow projected white light to blue (415 nm) and green (540 nm) wavelengths to generate a colored image. Blue-green light is able to enhance superficial mucosal capillaries and mucosal surface patterns; greater absorption of illuminating bands by hemoglobin yields darker-looking blood vessels. Although recent studies demonstrate that NBI does not increase the polyp detection rate when compared to WLI [6], this imaging modality is today increasingly common in commercial colonoscopes (e.g., H180AL/I, Olympus, Japan).

Colonoscopy is an outpatient surgery performed under sedation, and usually takes less than 30 minutes; however, patient compliance with recommended screening is low (i.e., 1 in 3 adults are not being screened [7]). This is attributed to the unpleasant preparation required, fear of pain during the procedure, and perceived embarrassment. Technological improvements in the field of endoscopy aim to help patients to overcome these hindrances.

1.2 Teleoperable Flexible Endoscopes

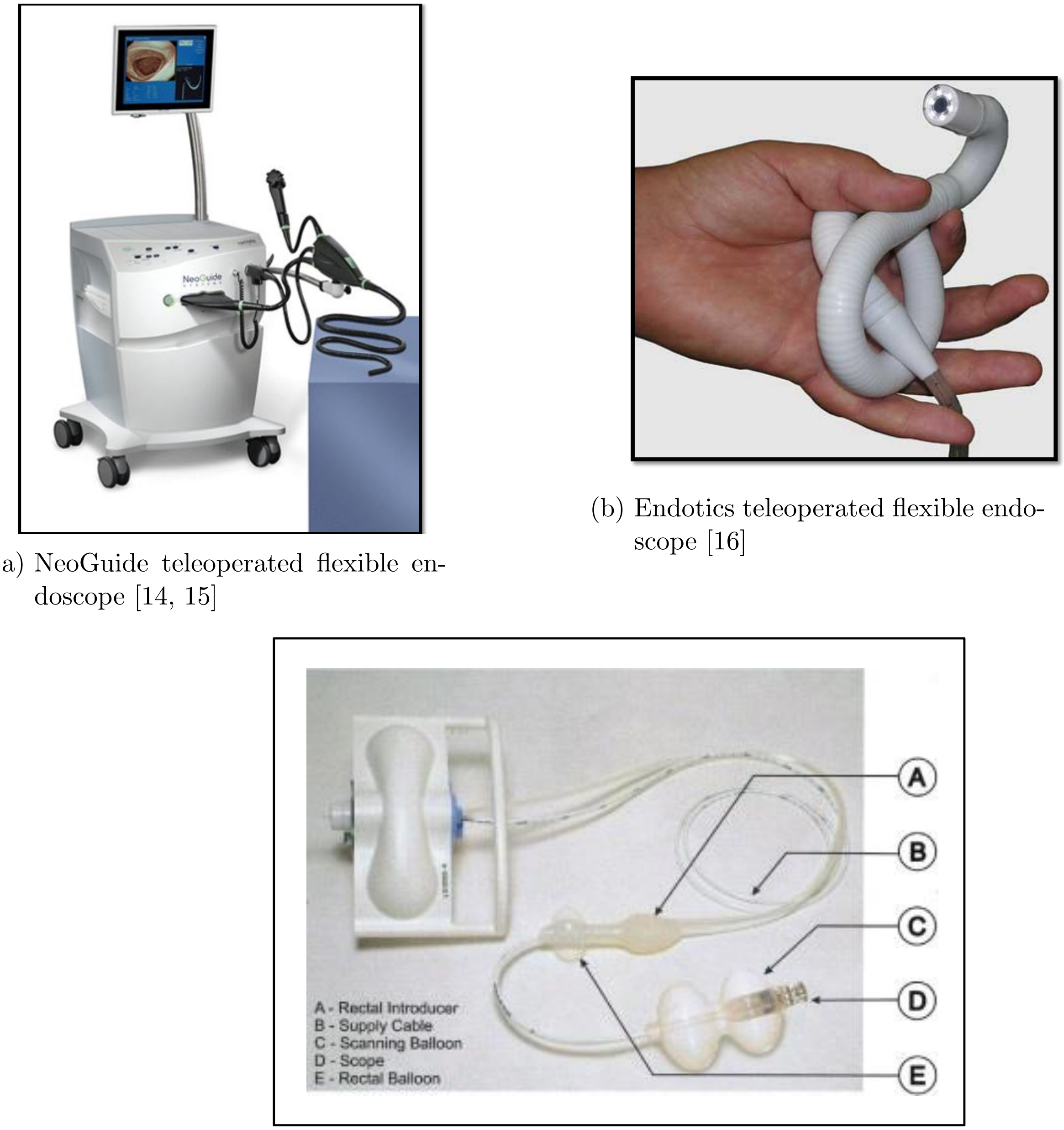
There are several different approaches to encouraging patient participation in colorectal cancer screening, including the development of increasingly flexible endoscopes, wireless capsule endoscopy (WCE), and virtual colonoscopy [3]. Alternately, computer-assisted technologies are also emerging to aid the doctor in detecting malignancies and increasing control over the intended trajectory of the endoscope. Robotics is playing an increasingly important role in this field with the development of fully- or semi-automated endoscopic systems [8, 9, 10, 3, 11, 12].

There are several clinically available computer-assisted endoscopes for increasing the comfort of colonoscopy and reducing the rigor and training necessary to perform the procedure.

The NeoGuide® system, shown in Figure 1.1a, specifically aims to avoid pain during the procedure by eliminating the incidence of looping. Looping accounts for 90% of pain episodes experienced by patients during colonoscopic procedures; it results from the continuing introduction of the endoscope into the tube without a corresponding advance of the endoscope head [13]. This results in stretching of the colon and displacement of the colon mesentery (attaches organs to the wall of the abdomen). The NeoGuide® addresses this problem by combining endoscope insertion with an electromechanically actuated insertion tube. Upon insertion of the colonoscope by the endoscopist, the system measures the endoscope's position and angle; thus, at each depth measured, the insertion tube then takes on the shape of the patient's colon [14, 15].

The Endotics® System is a self-propelled highly flexible robotic colonoscope which seeks to reduce the incidence of pain during colonoscopy. The endoscopic platform is shown in Figure 1.1b. The principal of operation of the device is much like that of an inchworm, stretching and shortening by creating anchor points within the colon. This motion propels the device through the colon. The physician is able to control the system through a hand-held console which allows for steering, advancing of the device through the colon, and common endoscopic controls such as insufflation, lens cleaning, and suction. A major asset of this device is that it maintains all the capabilities of a conventional endoscope, including a therapeutic channel for removal of polyps and tissue biopsy. Additionally, the Endotics systems boasts a very short learning curve of a few weeks to properly and effectively operate the device [16].

The Aero-O-Scope® is a self-propelling, self-navigating endoscope which functions via CO_2 insufflation. The device has 5 main parts as shown in Figure 1.1c: the rectal introducer, the supply cable, the scanning balloon, the scope, and the rectal balloon. The scope is inserted via the rectal introducer, whose balloon then expands to effectively plug the anus. The scope advances via the scanning balloon, which forms another gas barrier at the tip



(a) NeoGuide teleoperated flexible endoscope [14, 15]

(b) Endotics teleoperated flexible endoscope [16]

(c) Aeroscope teleoperable endoscopic platform [17]

Figure 1.1: Clinically available teleoperable flexible endoscopes.

of the scope; and the release of carbon dioxide gas behind the scanning balloon therefore propels the device forward through the colon [17]. A considerable advantage of this platform is its simplicity from the perspective of the gastroenterologist; it requires only inserting the rectal introducer and then pressing navigational buttons in order to control the direction of the endoscope. This platform effectively removes the physical demands and rigorous training associated with navigating an endoscope through the colon, and allows the physician to focus on inspection and detection of polyps.

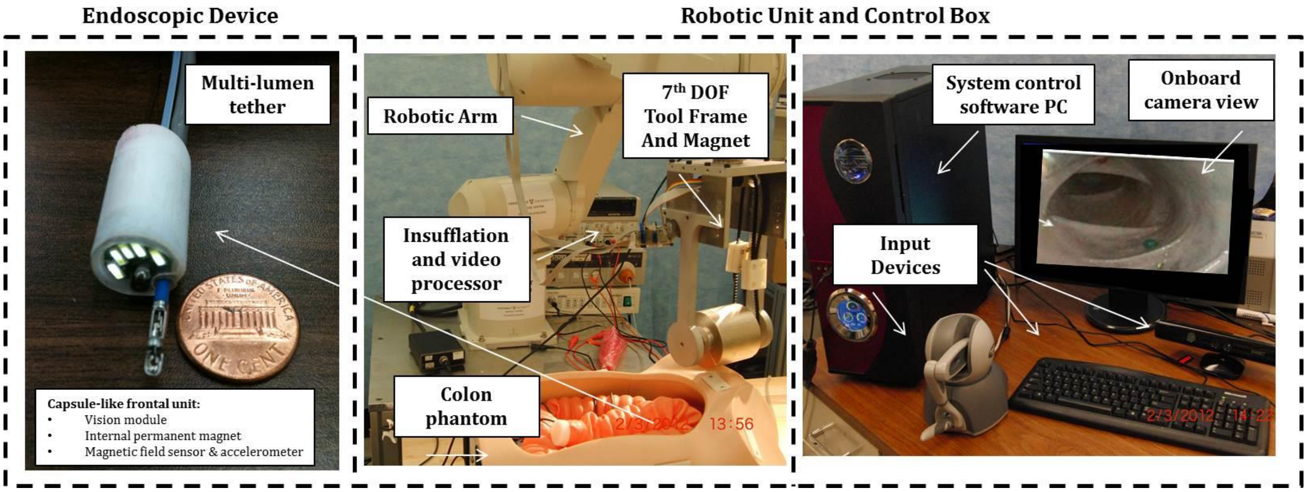

Endoscopic Device
Robotic Unit and Control Box
Multi-lumen tether
Capsule-like frontal unit:
• Vision module
• Internal permanent magnet
• Magnetic field sensor & accelerometer
Robotic Arm
7th DOF Tool Frame And Magnet
Insufflation and video processor
Colon phantom
System control software PC
Onboard camera view
Input Devices

Figure 1.2: Magnetically actuated teleoperated robotic endoscopic platform components.

A number of research platforms are also in development. The Magnetic Air Capsule (MAC) System, the target platform for this work, is a magnetically actuated flexible endoscope which aims to reduce pain episodes among patients and significantly reduce the learning curve for performing colonoscopy. The platform is shown in Figure 1.2. As shown, the system consists of a fully equipped flexible endoscope capsule, which contains LEDs, an endoscopic camera, a 5mmx12mm N42 diametric neodymium magnetic, a water/air channel, a suction channel, and a tool channel. The flexible endoscope is actuated by a N42 10cmx10cm diametric cylindrical neodymium magnetic, which is affixed to a 6 degree of freedom (DOF) industrial robot. The robot is controlled by computer software, which allows several different interfaces, including a Phantom Omni 6 DOF haptic joystick, a touchscreen monitor, and standard mouse/keyboard. The system reduces the incidence of looping and pain by introducing a “front wheel drive”; that is, instead of being pushed from behind, the device is pulled from the front [8]. For this platform, closed-loop control is a highly desirable capability; since the link between the endoscope and the actuation mechanism is magnetic and therefore not rigid, the actuation by the external magnet may not lead to the desired configuration instructed by the operator.

The introduction of teleoperable robotic endoscopes has the potential to widen the implementation of colorectal cancer screening and surveillance programs to rural areas, to mobile camps, or to in-field military bases, and the physical presence of an expert endoscopist may no longer be required.

1.3 Motion Estimation for Teleoperable Endoscopes

Real-time pose (i.e., position and orientation) estimation of the tip of a flexible endoscope is desirable for achieving reliable and effective teleoperation. Medical procedures require high precision and accuracy; an implementation of real-time pose detection confers calculated, controllable movements which are able to enhance system stability [18]. Additionally, the environment of the colon is highly variable among patients and by definition compliant; this aspect may be difficult to accurately model, thus inhibiting the effectiveness of model-based open-loop control. On the other hand, closed-loop control is effective for disturbance rejection and error minimization, which results in a system that is able to reduce the difference between the intended pose (i.e., user-commanded desired pose) and the actual pose (i.e., measured location of the endoscope tip). Real-time pose estimation facilitates closed-loop control by providing a feedback signal of the estimated pose of the endoscope head after actuation; Figure 1.3 illustrates a possible closed-loop control strategy based on image analysis.

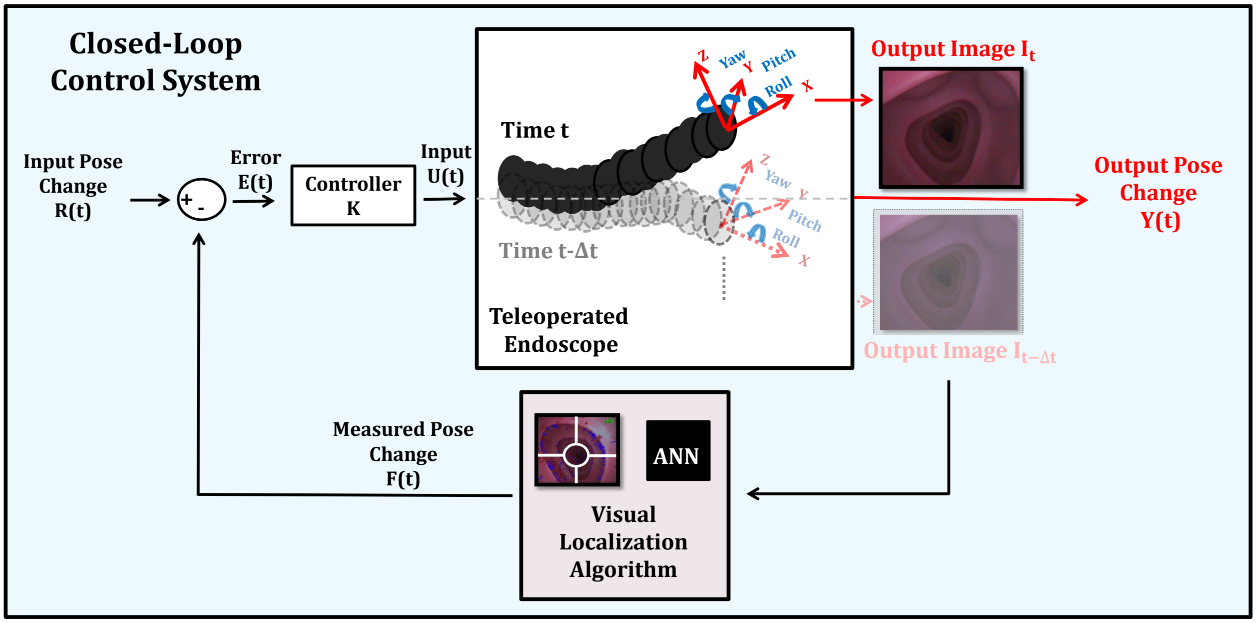


Figure 1.3: Closed-loop control system taking advantage of the proposed pose detection approach to guide a teleoperated endoscope.

Several methods have been introduced in order to achieve real-time pose detection. Magnetic tracking has emerged as a reliable method, and there are several commercial manufacturers of 5 and 6 degree of freedom (DOF) electromagnetic tracking systems [19, 20]. Practically, they can be placed down the tool channel of the endoscope due to their minute size (i.e., <1.8mm [19]), and used to track the pose of the endoscope head in real-time. In an alternate implementation, such as in the commercially available ScopeGuide® (Olympus, Japan), magnetic trackers have been placed along the entire length of the colonoscope. This provides the endoscopist visual feedback of the entire instrument pose given with respect to a global coordinate frame [21]. Real-time pose detection has additionally been achieved in bronchoscopy through the combination of the endoscopic camera stream with image registration and fluoroscopy [22, 23, 24, 25].

However, the use of a magnetic tracker requires additional space in the endoscope; this can result in an increase in the size of the device, possible reduction in the flexibility of the endoscope, and reduced accessibility to the therapeutic tool channel of the endoscope during colonoscopy. For endoscopes with extremely limited operational space such as encephaloscopes, rhinoscopes, and bronchoscopes, minimization of the size of the endoscope is fundamental. Additionally, a magnetic tracker continually occupying the tool channel of the endoscope can possibly compromise the standard of care during colonoscopy. Furthermore, both commercial entities and research labs worldwide are proposing platforms based on magnetic manipulation of endoscopic devices [26, 27, 28, 29, 30, 31, 32]. However, the magnetic manipulation of the endoscope interferes with the electromagnetic fields generated in the magnetic tracker system; this results in degraded, inaccurate, or missing localization estimates from the tracker. Soft body cavities in particular are regions in which accurate tracking of the endoscope head is essential; localization in conjunction with image registration is not effective due to the drastic deformations that occur upon repositioning of the patient [33].

1.4 Related Work

Real-time motion estimation and steering of flexible endoscopes presents a number of challenges [34, 35]. The endoscopic image stream has been used extensively to identify structural features of the colon including: isolating the lumen of the colon via evaluating the darkest region of the image [36, 37], identifying the ring-like haustral folds of the colon lumen [38, 39, 40], and using specular highlights resulting from illumination [41]. Several works have used colon structure identification as a basis for endoscopic steering (i.e., correcting the current heading of the camera towards to lumen center) [9, 42], enabling the possibility for automation. However, these works do not utilize metric motion estimation in their control strategies, and do not implement closed-loop control (i.e., although the motors are actuated towards the center of the lumen, there is no feedback as to whether they reached their intended destination).

Image analysis for tracking and motion estimation has been quite successful in other fields, including mobile robotics, unmanned vehicle navigation [43, 44], and autonomous egomotion estimation [45]. Popular techniques include using different types of features including: Shi-Tomasi features [46], FAST features [47], and Scale-Invariant Feature Transform (SIFT) [48] and Speeded Up Robust Features (SURF) [49] descriptors; different optical flow techniques, including: Lucas-Kanade [50], Hierarchical Multi-Affine (HMA) [51], and dense HMA [52]; and other popular techniques including visual simultaneous localization and mapping [53], and structure from motion (SFM) [54].

These approaches have also been applied within the field of image processing in gastroenterology. A 3-dimensional reconstruction of the colon using sequential images from a monocular camera was achieved using SFM reconstruction [55]. This implementation assumes zero rotation within the image; using this assumption, the algorithm is able to produce estimates of the 3 DOF translation between two images. The SFM algorithm is able to calculate 6 DOF motion up to scale; as a consequence, the metric translation cannot be estimated unless the scale factor is known. The spherical camera model has been used as a more accurate model due to the endoscopic lens; however, this requires simplifying assumptions about the rotation of the camera [9]. Focus of expansion has additionally been employed to avoid numerical instabilities related to optical flow vectors within SFM calculations, and has been successfully employed in virtual colonoscopy and other real image sets[56, 11]. However, algorithm performance on computer-generated datasets can vary significantly from datasets gathered from a silicon human colon simulator or a real human colon [9], and the focus of expansion may not always be in the image.

Application of artificial intelligence and machine learning techniques within the field have been mostly limited to signal filtering and amelioration of computer-aided diagnosis (e.g., object recognition, segmentation, etc.) [57, 58]. Moving picture expert group (MPEG)-7 features (popular for video and audio compression) combined with fuzzy logic were used for localization of a wireless endoscopic capsule within general anatomical regions of the gastrointestinal tract. Rule-based production systems using fuzzy logic have also been used to identify the lumen within an image [38]. However, the effectiveness of these algorithms for teleoperable flexible endoscopes has not been evaluated, and they do not provide quantitative feedback concerning the metric motion of the endoscope tip.

1.5 Thesis Overview

The remainder of this thesis will describe an algorithm suitable for motion estimation for a teleoperable flexible endoscope platform and demonstrate its feasibility. The proposed motion estimation system will be independent of the technological platform on which it is implemented, provide accurate motion estimates suitable for real-time feedback, and neither create unwanted interference in the endoscopic system nor increase the size of the endoscope.

The methodology presented in this paper builds upon previous work by using established optical flow methods to measure the relative motion between two sequential image frames, and then relate this description of motion within the image to the actual pose displacement using machine learning methods. The machine learning technique chosen was artificial neural networks, which are known for their noise rejection and nonlinear estimation capabilities. The applicability of the proposed method in both a controlled setting and in clinical use was tested in several robotically actuated experiments as well as using commercial endoscope operated by an expert endoscopist (>2,000 lifetime procedures). This work will then explore the impact of the elements of the algorithm upon its performance: the optical flow estimation method chosen, the illumination modality, and partitioning of the image.

Chapter 2 will provide an overview of main components of the algorithm: optical flow estimation and neural networks. Chapter 3 will describe the general steps of the algorithm. Chapter 4 will detail the experiments used to test the algorithm, and also describe elements of the algorithm which were compared to produce a more accurate algorithm. Chapter 5 will present and evaluate the results of the experiments, and Chapter 6 describes the final findings of the work.

CHAPTER 2

BACKGROUND

2.1 Coordinate Systems

Essential to the motion estimation and tracking of the endoscope is the calculation of the translational and rotational pose displacement of the endoscope tip. In several of the following experiments, the true location of the endoscope tip is given by a magnetic tracker placed in the tool channel of the endoscope. The magnetic tracker readings are given with respect to a global coordinate frame established by the electromagnetic fields produced by the magnetic tracker.

In order to calculate the relative pose displacement, a local coordinate frame is assigned to the endoscopic tip, and assumed to be coincident with that of the magnetic sensor. It is essential to calculate the pose displacement of the endoscope relative to its previous coordinate frame (i.e., not the absolute displacement in the global coordinate frame). To do this, the initial transformation (rotation and position) is subtracted from the rest of the data; thus, all the data is given relative to the initial position.

Given a local orientation specified as roll (rotation about x-axis), pitch (rotation about y-axis), and yaw (rotation about z-axis) parameters, a ZYX Euler rotation matrix $\mathbf{R}$ can be formed as:

$$R = \begin{bmatrix} c\psi c\theta & -s\psi c\phi + c\psi s\theta s\phi & s\psi s\phi + c\psi c\phi s\theta \\ s\psi c\theta & c\psi c\phi + s\phi s\theta s\psi & -c\psi s\phi + s\theta s\psi c\phi \\ -s\theta & c\theta s\phi & c\theta c\phi \end{bmatrix} \tag{2.1.1}$$

Given the initial pose $\mathbf{T_0} = [\mathbf{R_0}, \mathbf{t_0}]$, where $\mathbf{P} = [0, 0, 0]$ describes the global position of the sensor in 3d space, and the initial orientation is given by:

$$\mathbf{R_0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.1.2}$$

Given pose following actuation, $\mathbf{T_1}$, the change in position of the sensor $\Delta\mathbf{t_1}$ is calculated as:

$$\Delta\mathbf{t_1} = \mathbf{R_0^T}(\mathbf{P_1} - \mathbf{P_0}), \tag{2.1.3}$$

where $\mathbf{P_0}$ and $\mathbf{P_1}$ are given in the global coordinate frame. The change in rotation of the sensor $\Delta\mathbf{R}$ is then given as:

$$\Delta\mathbf{R} = \mathbf{R_0}^T\mathbf{R_1} \tag{2.1.4}$$

These two quantities form the pose displacement $\mathbf{\Delta T_1} = [\Delta\mathbf{R}, \Delta\mathbf{t}]$. This process is repeated for each pose following actuation. This calculation can be easily generalized to the entire sequential data set, where

$$\Delta\mathbf{t_i} = \mathbf{R}_{i-1}^T(\mathbf{P_i} - \mathbf{P_{i-1}}), \tag{2.1.5}$$

$$\Delta\mathbf{R_i} = \mathbf{R_{i-1}}^T\mathbf{R_i} \tag{2.1.6}$$

for the i^{th} pose in the data set.

The ZYX Euler representation is a common orientation representation in robotics, but particularly in aviation and nautical applications, in which it is frequently referred to as the Roll-Pitch-Yaw (RPY) convention [59]. The ZYX Euler convention is contrived as a set of 3 rotations about extrinsic angles (i.e., relative to the reference frame) as follows:

1. A rotation about the reference x-axis by an angle ϕ
2. A rotation about the reference y-axis by an angle θ
3. A rotation about the reference z-axis by an angle ψ.

This work adopts the Euler naming conventions of [59]; however, in this work, *roll* is considered to occur about the X-axis, and *yaw* is regarded to occur about the Z-axis. As in [59], *pitch* is considered to occur about the Y-axis.

2.2 Artificial Neural Networks (ANNs)

ANNs are computational networks which are useful in function approximation and pattern recognition, and were created based on biological neurons. The most basic unit of a neural network is the perceptron, a single "neuron" which performs the mapping

$$o = f(\mathbf{w}^T \mathbf{x} + b) \tag{2.2.1}$$

where o is the scalar output of the unit, $\boldsymbol{w}$ is a set of trainable weights which scale the input, $\boldsymbol{x}$ is the feature vector input to the unit, and b is a scalar bias. The function f is the activation function, which is usually implemented by hard limiters (i.e., on or off), linear, or sigmoid/logistic functions. A diagram of a single neuron is shown in Figure 2.1.

Multilayer feed-forward ANNs are interconnected networks of these single units of neurons, depicted in Figure 2.2. As shown, the neurons are arranged in layers, commonly referred to as *input layers*, layers which receive an input vector $\mathbf{x}_i$ from the training set $\boldsymbol{X}$

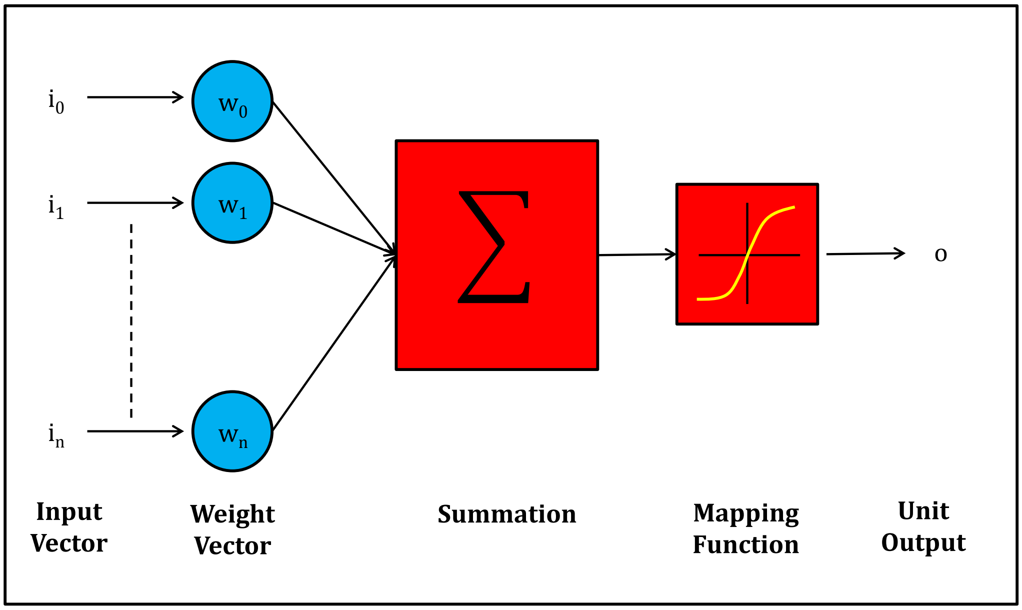


Figure 2.1: Flow diagram for the proposed method for calculating the change in the position and orientation of the endoscopic module, including the investigated variations in illumination (WLI or NBI) modality and spatial partitions (grid-based or lumen-centered).

and perform an identity mapping on its inputs, any number of *hidden layers*, layers of neurons whose inputs are the outputs of the previous layer, and the *output layer*, which performs a similar mapping as the hidden layer, but produces the final output $\boldsymbol{o}$ of the ANN. The graph structure of these networks requires no self-connections (i.e., the output of a unit is its own input) or look-ahead connections (i.e., output of of the unit connected to the input of a non-adjacent layer).

Multilayer feed-forward ANNs are desirable function approximators due to their rejection of noise and flaws in the training set, ability to accomodate high dimensional problem spaces with complex interactions, and speed of computation during operation [60, 61]. Multilayer feed-forward ANNs with certain characteristics (single hidden layer with a finite number of hidden neurons and arbitrary activation function) are universal approximators. This means that these ANNs are able to approximation continuous functions of n real variables with support in the unit hypercube, although convergence depends on the number of neurons in the hidden layer and properties of the function being approximated. Thus, ANNs are powerful for learning complex mappings, given the correct number and size of hidden layers, and certain characteristics of the function to be mapped from a set of exemplars and their target outputs [62].

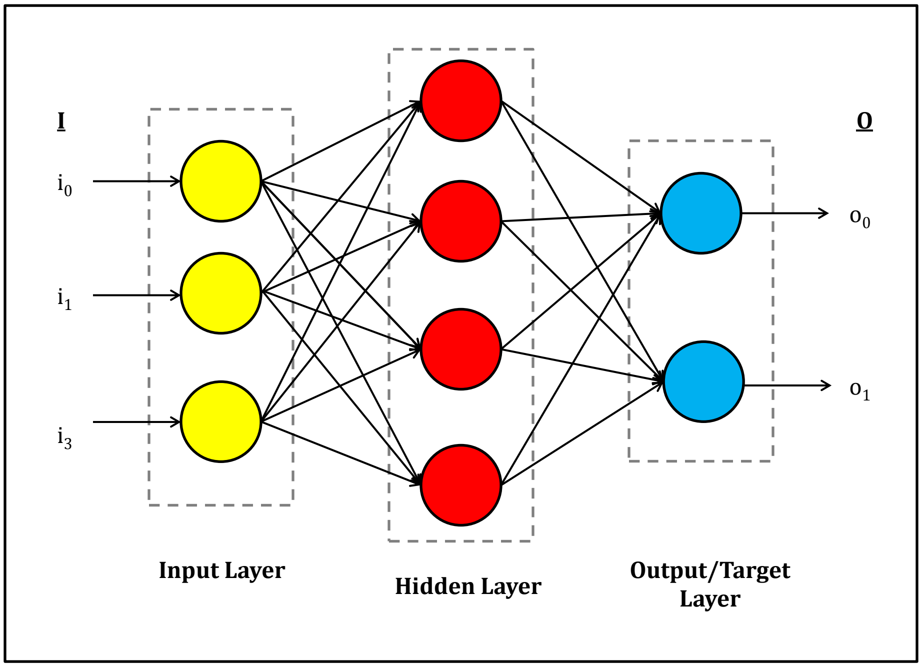


Figure 2.2: 3x4x2 multilayer feedforward ANN topological structure showing input, hidden, and output layers. Each node in any given layer performs the perceptron mapping.

Multilayer feed-forward ANNs employed in this paper are trained via a supervised learning approach. This means that the method requires a training set representative of the function to be approximated. A training set consists of n training samples composed of an input vector and a target output vector pair. Training proceeds as outlined in Algorithm 2.3. As described, training requires iterating through each input/output vector pair in the training set. For each pair, the input vector is forward propagated through the network to produce an output estimate. After this estimate is compared to the target, the error between these two values is backpropagated via the desired training algorithm; that is, the ANN weights adjusted such that the error between the estimations provided by the neural network and the known outputs is minimized. This proceeds until some criteria for termination is satisfied.

```
while error minimization cost function has not been satisfied do
    repeat
        Forward propagate input feature vector $\mathbf{x_i}$ through ANN to calculate output $\mathbf{o_i}$
        Compare estimated outputs to target outputs in training set
        Calculate and backpropagate error using desired training algorithm
    until end of training set has not been reached (i.e., i=n)
end
```

Figure 2.3: Algorithm for training of neural networks for function approximation

There are many popular algorithms used for neural network training, including gradient descent and variants, Levenberg-Marquardt backpropagation [63], and Gauss-Newton backpropagation; all require an error measure to be minimized. The most common error measures are sum-of-squared error, root mean squared error (RMSE), and mean squared error.

2.3 Optical Flow Computation

Optical flow is a pixel-based description of the relative motion between two images. There are many optical flow calculation algorithms; however, in this work we concentrate on three methods due their ubiquity in practice and applicability to the described problem. These optical flow methods can be divided into two groups: sparse and dense optical flow. Sparse optical flow techniques generally begin by identifying small template regions in the image

$I_{t-\Delta t}$ that present unique qualities and are likely to persist and have similar characteristics in a proceeding image I_t. These regions are then identified in image I_t. Given the assumption that the scene is static, all motion in the image can be assumed to be caused by motion of the camera.

2.3.1 Lucas-Kanade (LK) Optical Flow

LK optical flow [50] is a well-established and widely used algorithm for estimating the relative motion in an image sequence based on certain features. The features commonly used are Shi-Tomasi (ST) features [46], which are identified based on the strength of their eigenvalues within a local pixel neighborhood. In the LK formulation, given a pixel region at location *x,y* at time *t* with intensity I, a small camera movement Δx, Δy over time Δt is given by the equation:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2.3.1}$$

Since the movement is small, a Taylor series approximation can be made such that:

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\delta I}{\delta x}\Delta x + \frac{\delta I}{\delta y}\Delta y + \frac{\delta I}{\delta t}\Delta t + HOT \tag{2.3.2}$$

Since

$$\frac{\delta I}{\delta x}\Delta x + \frac{\delta I}{\delta y}\Delta y + \frac{\delta I}{\delta t}\Delta t = 0 \tag{2.3.3}$$

then dividing by Δt results in the optical flow equation:

$$\frac{\delta I}{\delta x}V_x + \frac{\delta I}{\delta y}V_y = -I_t \tag{2.3.4}$$

or

$$I_x V_x + I_y V_y = -I_t \tag{2.3.5}$$

where V_x and V_y are the optical flow in the x and y directions of the image. LK solves this equation by applying it to all pixels within a local neighborhood of a certain pixel p, and solves a least squares problem to estimate V_x and V_y. As is evident in this formulation, LK requires three assumptions - brightness constancy (i.e., the brightness of a tracked pixel stays constant among frames), small movements, and neighboring points belong to the same surface and have similar motion [50, 64].

2.3.2 Scale-invariant Optical Flow

Scale-invariant feature matching algorithms are particularly well-suited for identification of robust matching of features between images. Scale-invariant algorithms introduce an important augmentation of features such as corners by conferring image scale and rotation invariance; corner descriptors fail in cases in which a feature appears rotated or slightly larger/smaller than in the reference image. One popular scale-invariant feature matching algorithm is the Scale-Invariant Feature Transform (SIFT) [48], which proceeds in 4 steps. The first step is scale-space extrema detection, in which points of interest are identified. This is done by convolving an image with the Gaussian kernel n times, and taking the difference between two successive applications of the Gaussian function. This yields a set of difference-of-Gaussians (DOGs) for a given scale, and is referred to as an *octave.* The original image is then downsampled by a factor of 2, and the process is repeated for a number of scales. This step is essentially for providing scale invariance, and since it is a very close approximation of the Laplacian, which confers rotation invariance, both scale and rotation invariance can be achieved.

The second step of the algorithm is detecting the local extrema. This is done by observing a particular image at each sample point, and comparing it to each of its eight neighbors in the current image and nine neighbors in the scales above and below it. The point is only considered a maxima or minima if it is larger than all of these neighbors (maxima) or smaller than all of these neighbors (minima). This step serves to eliminate low-contrast and poorly localized edge extrema.

The third step assigns a consistent orientation to each keypoint to achieve rotation invariance. For computational efficiency, the gradient magnitude and orientation are pre-calculated for each image point in each image. For each keypoint, an orientation histogram is formed, which has 36 bins corresponding to the 360 degree range of orientations. Since the peaks of the histograms represent the strongest directionality of local gradients, this orientation is strongly representative of the keypoint. A parabola is fit to the 3 histogram values closest to each peak to accurately estimate the true orientation peak. Additionally, if any bin of the histogram falls within 80% of the peak, and additional keypoint is created at this same location with this additional orientation information.

The final step composes the keypoint descriptor. For each point of interest, a Gaussian weighting function is applied to each pixel within the neighborhood of the pixel of interest. The descriptor is created by observing 4 4x4 pixel sample regions around the pixel of interest. An orientation histogram is made for each 4x4 region with 8 orientation bins each, which results in an 128 element feature vector for each keypoint. Illumination invariance is then achieved by normalizing the feature vector to unit length; if a brightness change occurs which uniformly increases the intensity of each of the pixels, this will be cancelled out by normalization. Non-linear illumination changes are addressed by thresholding the normalize feature vector at 0.2, and forcing any individual value in the feature vector to be a maximum of 0.2. The feature vector is then renormalized to unit length. This allows matching of the distribution of orientations to play more of an important role than matching of strong gradients.

SURF [49] is a similar algorithm for scale-invariant feature descriptor generator, although differences include utilization of a different mechanism for generating differences in the image pyramid and calculation of the feature descriptor. Although SIFT and SURF have comparable performance, SURF is generally a faster algorithm. An alternative method for using scale invariant features is to eliminate the feature detection parts of the algorithm, and instead use other algorithms for feature detection. One popular algorithm is the Features from Accelerated Segment Test (FAST) corner detector [47] due to its computational efficiency.

2.3.3 Hierarchical Multi-Affine (HMA) Optical Flow

HMA is a feature matching refinement algorithm whose purpose is to find regions of similarity between two distinct images of the same scene. This differs slightly from optical flow, in which there are assumptions placed on the movement of the pixels (i.e., HMA does not necessarily two consecutive frames of video). This algorithm is known for its speed, accuracy, and robustness to image clutter, and has been used specifically in minimally-invasive laparoscopic surgical scenarios. After using a feature identification mechanism (usually SIFT [48]) to identify and describe keypoints, matching is initially performed using the nearest neighbor distance ratio (NNR) appearance-based criteria [65, 51]. A similarity transformation is then used to relate the two features of the match, and consists of a rotation, translation, and scale as $s_i = [\delta x_i, \delta y_i, \delta \sigma_i, \delta \theta_i]$.

HMA then proceeds in 3 iterative phases. The clustering phase first partitions the matches into k subsets based on their keypoint descriptors and their positions in the image. For all the matches of each cluster, similarity-space clustering, based on the similarity space parameters s_i occurs. A scaling vector is applied to weight the importance of each of the parameters, and is clustered via the k-means clustering algorithm. The second phase is the affine-estimation phase, in which the keypoints in each cluster are used to estimate the image transformation matrix $\mathbf{A}$, which represents the affine transformation between the two images. RANSAC [66] and a non-parameteric outlier-detection technique [67] are used for robust rejection of outliers. The final phase of the algorithm is the correction phase, which reassigns the outliers to more appropriate clusters, and the transformation matrix $\mathbf{A}$

is recomputed if necessary. Additionally linear discriminant analysis [61] is used to separate the clusters such that each cluster belongs to only one class.

The algorithm terminates when all nodes are classified as termination nodes. This is done by an analysis of the number and ratio of inliers; if the number of inliers is below an arbitrary threshold τ_C, or the ratio of inliers is greater than an threshold τ_ρ.

CHAPTER 3

METHODOLOGY

3.1 Overview

The proposed algorithm calculates the rotational and translational displacement (i.e., 6 DOF transformation matrix, a common representation of pose in robotics [59] utilizing 3 DOF for position and 3 DOF for orientation) between sequential image frames using the endoscopic camera stream. The accuracy of the metric 6 DOF motion estimation of the endoscope relies heavily on the extraction of stable features from the image stream. However, the gastrointestinal lumen is well-known for its textureless appearance and lack of brightness constancy due to the changes in illumination due to the motion of the endoscope [9, 68, 56]. In this algorithm, we rely on well-established feature detectors and descriptors (Shi-Tomasi features, FAST, SIFT, and SURF) to address this challenge, while also reducing possible noisy data from these disturbances by using a neural network and summary statistics over image regions.

Figure 3.1 depicts the steps of our proposed algorithm. It begins by first finding strong feature correspondences in two sequential images, providing a description of the optical flow (i.e., a measurement of the motion of the endoscope in pixels). A spatial grid is then applied to form a feature vector which corresponds to the motion of the endoscope, as described in pixels. At the same time, ground truth data is acquired via an industrial robotic manipulator or magnetic tracker to provide the actual metric motion of the endoscope. These two components form a single pair within the training set, gathered over an entire trajectory of motion of the endoscope. The performance of the ANN is then tested using another separate diverse training set.

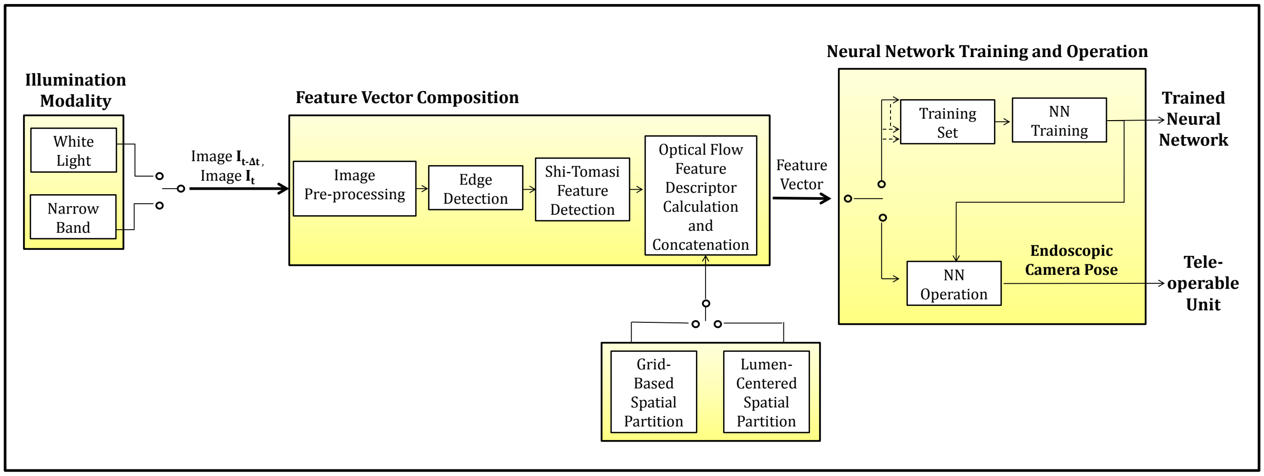

Illumination Modality
White Light
Narrow Band
Image $\mathbf{I}_{t-\Delta t}$,
Image $\mathbf{I}_t$
Feature Vector Composition
Image Pre-processing
Edge Detection
Shi-Tomasi Feature Detection
Optical Flow Feature Descriptor Calculation and Concatenation
Grid-Based Spatial Partition
Lumen-Centered Spatial Partition
Feature Vector
Neural Network Training and Operation
Training Set
NN Training
Trained Neural Network
NN Operation
Endoscopic Camera Pose
Tele-operable Unit

Figure 3.1: Flow diagram of methodology used for calculating change in 6 DOF displacement in orientation and translation from the endoscopic camera stream.

As previously stated in Section 2.3, the essential assumption made is that the scene is static; as a consequence, movement perceived in the image can be assumed to be due to the motion of the endoscope only. How valid is this assumption? There are four major contributors to possible movement of the colonic scene: respiration, deformation of the colon wall due to the flexible nature of the endoscope, haustral contractions, and insufflation. The effects of the respiration of the patient are assumed to be negligible since the colon is insufflated during colonoscopy. Furthermore, on average, a displacement of approximately 7.85mm occurs in the anterior/posterior plane during deep respiration [69]. Due to the high frame rate of the camera, the contribution of this movement is assumed to be minimal.

The motion of the colon due to the pressure of the endoscope upon the colon walls can be significant; excessive pressure and stretching of the colon due to looping of the colonoscope can lead to perforation [13]. Although this motion is clearly significant, this movement occurs *behind* the endoscopic camera; thus, it does not contribute significantly to a change in the endoscopic scene captured by the camera.

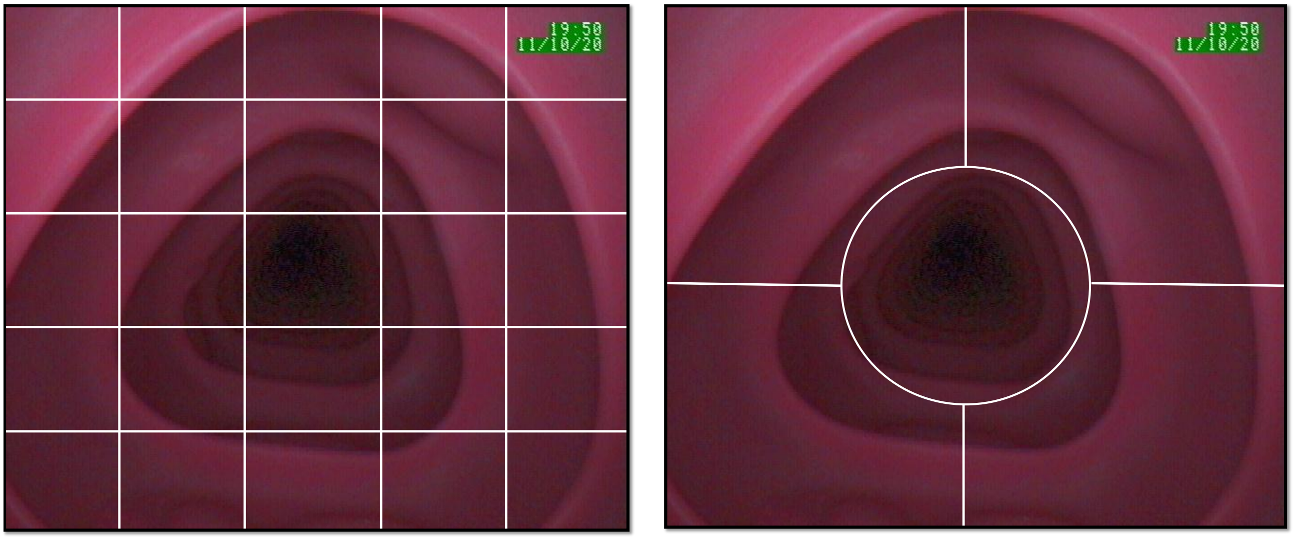
Haustral contractions also cause significant movement within the scene; these events occur as a result of peristalsis to move the content of the colon forward. These movements significantly affect the inertness of the scene viewed by the camera. However, these only occur every 25–30 minutes [70], and a specialized control loop within the teleoperation software can be used to handle this exception. This control loop would also be useful in halting motion estimation during insufflation, which also changes the scene viewed by the camera. However, since this is a function of the endoscope, it can be carefully monitored for changes commanded by the user.

3.2 Input Descriptor Composition

Figure 3.1 presents a flow diagram of the method used to acquire and process the endoscopic camera images. After the illumination modality is chosen, frames are sequentially captured from the video processor at times t-Δt and t. The images are first processed in order to preserve only their effective pixels; an endoscopic video feed usually displays information superfluous to the algorithm, such as a black region containing the patient's name and other personal information and endoscopic system information. Inclusion of this extraneous information into the algorithm creates easily-avoided noise and data inconsistencies.

After the effective image is converted to grayscale, the optical flow algorithm is applied to the image. For example, in the case of Lucas-Kanade optical flow, the Shi-Tomasi (ST) features [46] are first found within image $I_{t-\Delta t}$. Using the Lucas-Kanade optical flow algorithm, the corresponding features are found in image I_t. Thus, a broad description of the optical flow (i.e., relative image motion) is encoded.

The next step of the algorithm is to form an optical flow descriptor of the optical flow between the two images. This is an essential step; the formulation of these vector descriptors



(a) Grid-based spatial partition.

(b) Lumen-centered spatial partition.

Figure 3.2: Spatial partitioning rules for feature vector composition.

defines the unknown function to be approximated by the ANN, and provide the input to the ANN for training and testing. In this work, the optical flow descriptors were created to summarize the nature of the correspondences in specific pre-determined regions (i.e., partitions) of the image. The partitioning methods chosen are shown in Figure 3.2. Grid-based spatial partitioning, shown in Figure 3.2a is a common partitioning method used in many computer vision applications [71, 72, 73, 56]. It consists of dividing the image in 25 regions (i.e., a 5 × 5 rectangular grid) of equal area. For each grid location $g \in G$, two feature descriptors are calculated as

$$\overline{dx}_g = \frac{\sum_{i=1}^{n_g} dx_i}{n_g}$$

and

$$\overline{dy}_g = \frac{\sum_{i=1}^{n_g} dy_i}{n_g}$$

where n_g is the number of feature correspondences present in image I_t at grid location g, and dx and dy are the change in coordinates in the X and Y directions between corresponding features in image $I_{t-\Delta t}$ and I_t. These region descriptors are then concatenated into an overall optical flow descriptor of size 50 (25 grid regions with 2 feature descriptors each) to be used as input to the ANN.

Lumen-centered spatial partitioning, shown in Figure 3.2b, is based on the anatomical structural appearance of the colon lumen within an endoscopic image; the colon appears as a tubular structure with a dark region usually corresponding to the center of the lumen, and the rest of the image corresponds to the colon wall. This partitioning method is based on a consistent alignment of the center of the lumen with the center of the dark region in the image. A lumen segmentation approach similar to a method presented in [9] was taken by first histogram equalizing the region to increase contrast, and then applying an arbitrary threshold. This threshold was determined empirically. The resultant image consists of the lumen, which appears white, while the rest of the image appears black.

To begin to define the regions of the lumen-centered partition, two methods were compared. Both methods require the computation of the centroid of the lumen, (x_c, y_c), to define the location of the center of the lumen. The first method utilized the circumference of the lumen. This was achieved by summing the edge pixels of the lumen region in the thresholded image. The radius r is calculated by dividing the circumference by 2π. The second method resulted in a slightly more stable radius size, and was achieved by calculating the second moment of the lumen region (i.e., the area A), and then calculating the radius as $r = \sqrt{\frac{A}{\pi}}$. This second method was therefore employed in the algorithm. Together, the centroid of the lumen and the area within its radius define the first region of the lumen.

The other four quadrants are defined by dividing the image horizontally at y_c, and vertically at x_c, excluding including the area defined as the lumen center. For each of these 5 regions, the two feature descriptors are then calculated as

$$\overline{dr}_g = \frac{\sum_{i=1}^{n_g} \sqrt{dx_i^2 + dy_i^2}}{n_g}$$

and

$$\overline{d\theta}_g = \frac{\sum_{i=1}^{n_g} tan^{-1}(\frac{dy_i}{dx_i})}{n_g}$$

where $\overline{dr}_g$ is the average distance of optical flow between corresponding features in region g, and $\overline{d\theta}_g$ is the average inclination of the flow of the features in region g. These features are then concatenated into a feature vector of size 10 (5 regions described by 2 feature descriptors each) for use as inputs to the ANN.

3.3 ANN Training and Usage

The optical flow descriptors created by the procedure outlined in Section 3.2 then define the inputs to the ANNs. A description of how the output targets in the training set are acquired is left to Chapter 4, since several different methods were employed. It can be assumed that after the data are acquired and processed, ground truth target data corresponding to particular input vectors are available, and together, these form the training set.

The set of feature descriptor/target pose pairs generated via the grid-based or lumen-centered partitioning methods are then used to train multilayer feed-forward ANNs, previously described in Section 2.2. In order to train the ANNs to accurately estimate the metric endoscope motion, the full training set is further divided into a slightly smaller training set, a validation set, and a test set. As outlined in Section 2.2, each input descriptor is presented to the neural network and forward propagated to produce an output. This output is compared against ground truth data, and the error is backpropagated through the neural network using a selected training algorithm. The algorithm utilized in this work was Levenberg-Marquardt (LMA) error backpropagation, which is a robust algorithm used in ANN training since it is able to find a solution even if the initial weights of the network are very incorrect. However, similarly to other training algorithms, given a complex error surface, LMA can converge to a *local* minima rather than a *global* minima, resulting in sub-optimal network estimation.

Training ends via to *early stopping*, a termination criterion met when the error in the validation set begins increasing over a specific number of epochs. Early stopping reduces the incidence of overtraining (i.e., memorization of the training set, including the noise within it), and usually confers an enhanced generalization capabilities for the ANN. The testing set is used as a test of performance by measuring the error before and after training. At this point, the ANN is considered trained, and is ready for use in practice.

CHAPTER 4

EXPERIMENTS

4.1 Overview

In order to test the validity of the approach, several experiments were performed to evaluate the estimation ability of the algorithm in different environments. All the experiments examine the role of partitioning of the optical flow input descriptor on the performance of the resultant ANNs. The first experiment specifically assesses the role of different illumination modalities on the performance of the ANNs, and compares its performance to that of a magnetic tracker. Since the magnetic tracker is robotically actuated, this represents a similar environment in which a magnetic tracker or the algorithm would possibly be employed.

The second experiment evaluates the algorithm practically within a simulated clinical setting. This is achieved by utilizing a commercial endoscope equipped with WLI and NBI operated by an expert gastroenterologist ($>$ 2,000 lifetime procedures performed) to complete four colonoscopies on a colonoscopy training simulator. We again assess the role of illumination modality and partitioning method; however, we evaluate the efficacy of training the ANNs on magnetic tracker data, which is more easily obtained in a clinical setting. With this work, we additionally investigate the disparity in power of features produced under WLI and NBI, as well as the role of the color channel of the images.

The final experiment adopts an artificial, but accurate setup similar to the first experiment in order to examine the optical flow descriptor produced, and the role that different optical flow algorithms can play in the production of a more accurate neural network. In this experiment, the concentration is solely on the interactions between the partitioning method, descriptor, and the optical flow technique used; thus, the illumination modality is fixed as WLI.

4.2 Benchtop Evaluation of Magnetic Tracker vs. Robotic Encoders vs. ANN Estimates

The benchtop experiment formulated to assess the performance of the four variations of ANNs based on illumination (WLI vs. NBI) and partitioning (grid-based vs. lumen-centered) against a magnetic tracker is shown in Figure 4.1. The setup consists of a tethered endoscopic module (22 mm in length ×27 mm in diameter) rigidly affixed to a 6 DOF industrial robotic manipulator (RV-6SDL; Mitsubishi Corporation, Japan). This rigid connection with the robotic enables the 6 DOF position and orientation of the endoscopic module to be accurately derived from the robot encoders at any given point in time. The data from the robot encoders is considered to be ground truth, and is used to formulate the known output targets of the ANNs in the training set.

The endoscopic module houses a 500 × 582 resolution endoscopic camera (291,000 effective pixels, cross-section 3 mm × 3 mm, and 140° field of view; Introspicio 110, Medigus, Ltd, Israel), 5 white light emitting diodes (LEDs) (NESW007BT; Nichia Corporation, Japan), and 6 blue light (450nm) LEDs (Kingbright Electronic Company, Ltd, Taiwan) for NBI. The two illumination systems were designed and driven to have approximately the same light intensity. The unit also houses a 6 DOF magnetic tracking sensor (1.4 mm positional nominal RMSE, 0.5° rotational nominal RMSE, 240-420 Hz update rate; 3D Guidance trakStar, Mid-range; Ascension Technology Corporation, USA) for comparison to the motion estimation

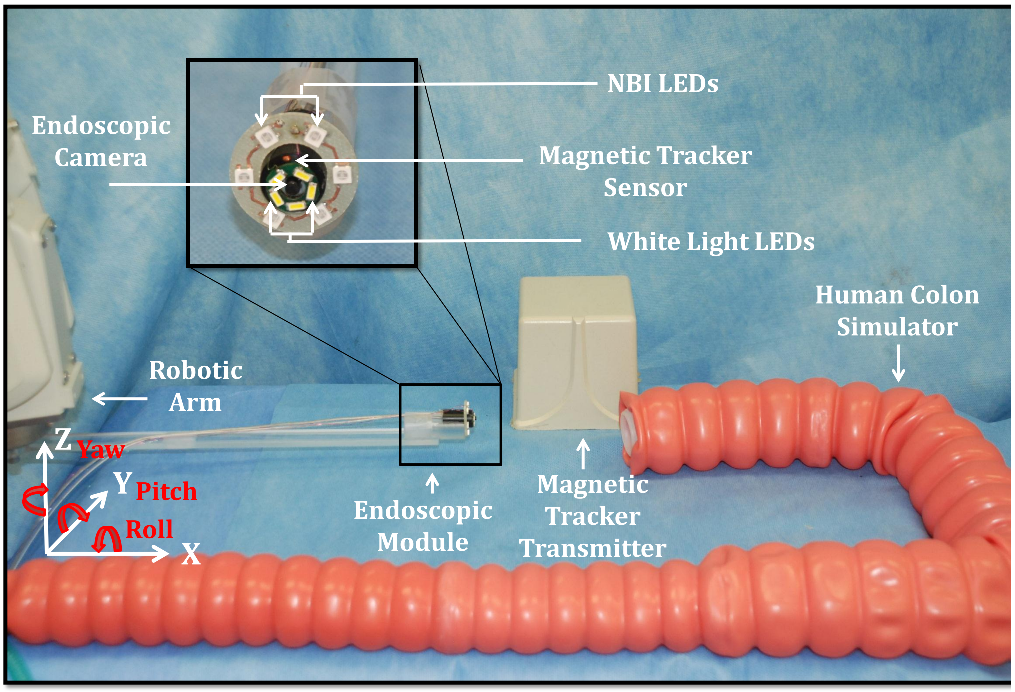


Figure 4.1: Experimental setup for benchtop evaluation of magnetic tracker, robot encoders, and the ANN estiamtes.

outputs of the ANNs. Validation software provided by the manufacturer with the device was used to appropriately position the magnetic tracker transmitter such that interference could be minimized and the highest fidelity readings could be realized.

During training and testing, the endoscopic module is actuated along the straight sections of a plastic human colon simulator (Kyoto Kagaku, Japan). This simulator is very common in the training of medical doctors for performing colonoscopy, and therefore possesses similar characteristics in terms of the gross anatomy of the colon. In order to mimic characteristics of features that are enhanced by NBI, specifically blood vessels and capillaries in the colon, fresh porcine blood was applied to the interior of the simulator. The setup was then covered by an opaque black cloth (not shown in figure) to replicate the gastrointestinal environment.

```
Result: Trained ANN
Select desired illumination modality
Read in image $I_{t-\Delta t}$
Record robot encoders and magnetic tracker sensor position at time $t$
while Not finished with training trajectory do
    Move robot/endoscopic camera to next training pose
    Read in image $I_t$
    Generate and record optical flow-based feature vector
    Record robot and magnetic tracker sensor position
    Set image $I_{t-\Delta t} = I_t$
end
Calculate change in pose ($\Delta P_{target}$) from ground truth to be used as target vectors for ANN
while Validation training error has not increased for six epochs do
    Forward propagate input vector through ANN to get estimated pose $\Delta T_{estimated}$
    Backpropagate error to train network
end
```

Figure 4.2: Algorithm for training set generation and training of ANNs.

The system was controlled using interface software written in C++ to send positional commands via TCP ethernet connection to the robot controller. This moves the arm in real-time into the desired position. Each time the endoscopic module is moved, the movement of the robot, the 5 DOF pose given by the tracker, and the corresponding endoscopic image was recorded. This processed is detailed in Algorithm 4.2. The tracker position was captured using C++ interface software using functions from Ascension's 3D Guidance Application Programming Interface (API). Individual image frames were acquired using a frame grabber connected to the camera video processor, and were read and saved using OpenCV [74] library functions.

Table 4.1: Magnitude, direction, and number of training repetitions for generating ANN training set.

Degree of Freedom Tested	**Magnitude of Training Repetitions**	**Total Number of Training Repetitions**
X only	±0.5 mm to ±5 mm	180
Y only	±0.5 mm to ±3 mm	120
Z only	±0.5 mm to ±3 mm	120
Roll only	±0.5° to ±2°	80
Pitch only	±0.5° to ±2°	80
Yaw only	±0.5° to ±2°	80
Translation only	Variable	120
Rotation only	Variable	80
All degrees	Variable	320

Algorithm 4.2 details the procedure for generating the training and testing sets. With each movement of the robot/endoscopic module, the resultant optical flow feature descriptors are calculated, and the robot and magnetic tracker positions are recorded. Table 4.1 describes the training trajectory, where the coordinates are given with respect to the Cartesian axes

of Figure 4.1. The training trajectory was composed of 1180 steps, which were considered to be representative of the endoscopic movements during colonoscopy [75]. The training set allows 10 repetitions of varying magnitudes for each independent DOF independently tested, and 5 training repetitions for any movement combining multiple DOF. When combinational movement is tested, the magnitudes of movement can range between 0mm or 0° to the maximum shown in Table 4.1. The conclusion of this training trajectory execution marks the end of the acquisition of the inputs and targets of the training set, and the beginning of ANN training.

Training of the ANNs proceeds offline in the manner demonstrated in Figure 3.1, and as described in Section 2.2. The inputs to the ANNs are the optical flow descriptors, compact representations of the evolution of the scene between time t-Δt and time t. The output targets of the algorithm are calculated as a difference between the 6 DOF pose at time t-Δt and time t. Using these data as the training set, the ANN is trained until completion, and approximates a function which relates the relative motion in the image to the 6 DOF metric motion between two time points.

Once the ANNs are trained, testing begins. Testing requires a similar environment and procedure as training, and so the endoscopic module is actuated along a trajectory, and the 6 DOF pose of the tracker and robot are recorded at each time step, along with the endoscopic optical flow descriptors describing the relative motion. The testing set was randomly generated, so as not to provide a testing set too similar to the training set. The bounds of the training set were as follows: 0 mm±5 mm in the Z direction; 0 mm±3 mm in the X and Y directions; and rotations of 0°±2° in roll, pitch, and yaw as defined in Figure 4.1. The main difference between training and testing is that there is no backpropagation of error or modification of system weights; the optical flow descriptors are simply forward propagated to produce the estimated pose. The error measure used to determine the accuracy of the system is RMSE since the error contains both positive and negative magnitudes. RMSE is calculated for the commercial tracker and ANNs against the ground truth given by the robot.

Implementation was achieved via Matlab's Neural Network Toolbox; it was used for both training and testing of the ANNs. The training set was divided so as to allocate 85% of the data to the set on which the ANNs would be trained, 10% to compose the validation set, and 5% for the testing set. These were determined empirically to provide good results. Early stopping was employed to reduce overtraining, and was invoked if the error in the testing set increased for six successive epochs. The ANNs are constructed to follow the $2n+1$ hidden layer architecture suggested in [76], where n is the number of neurons in the input layer (i.e., the number of descriptors in the optical flow descriptor). Thus, the ANNs resulting from grid-based partitioning resulted in a network of architecture 50×101×6, and the lumen-centered partitioning resulted in a network of architecture 10×21×6.

4.3 Evaluation of Proposed Method using Clinical Protocol

The training method proposed in the previous section provides the most accurate and reliable data on which to train the ANN, since it depends on the highly accurate encoders of an industrial manipulator. However, practically, accurate ground truth data may not be feasible or possible to collect. To address this cause, we performed an experiment more similar to the training conditions which would be encountered in practice.

The experimental platform created to replicate these conditions is shown in Figure 4.3. In this experiment, an expert gastroenterologist performed a set of four colonoscopies using the colonoscopy training model (Kyoto Kagaku, Japan). The plastic human colon simulator was placed inside in a basic anatomical configuration (Figure 4.3, upper-right corner). In order to ensure the presence of randomized features, the colon was filled with blood, and then drained. This prevented the ANNs from observing a similar pattern during each colonoscopy.

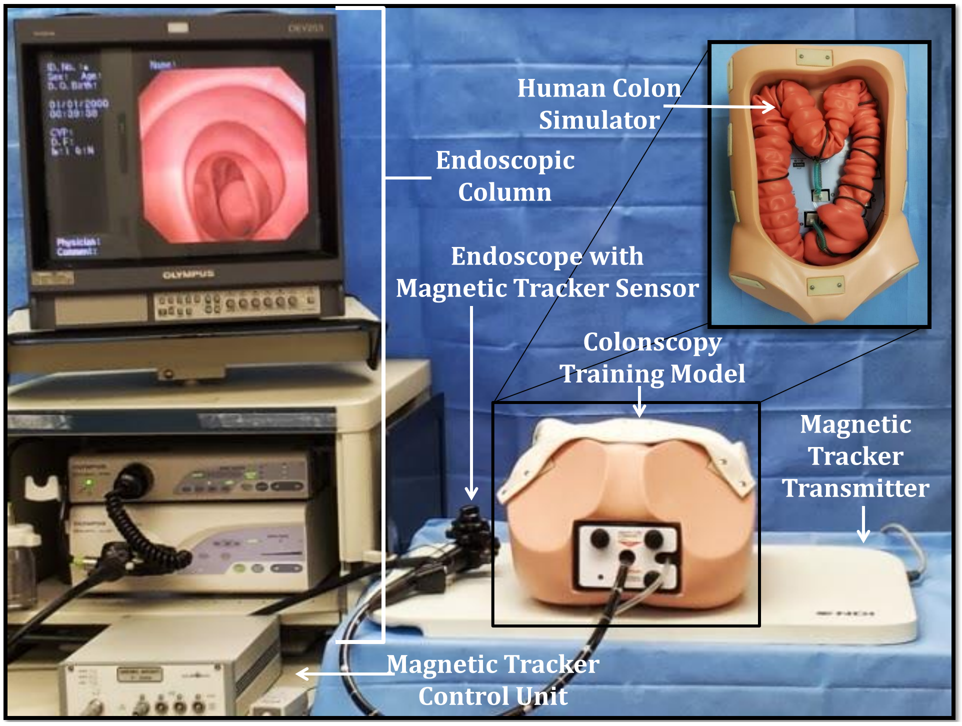


Figure 4.3: Experimental setup for evaluating the proposed method in a clinical setting.

A 5 DOF magnetic tracking system (1.20mm positional nominal RMSE, 0.5° rotational nominal RMSE, 40Hz update rate; Aurora, Tabletop Transmitter; Northern Digital Inc. (NDI); Waterloo, Canada) was inserted into the tool channel of a state-of-the-art flexible endoscope (H180AL/I Colonovideoscope; Olympus, Japan). This endoscope is able to provide both WLI and NBI for colon inspection. Of the four colonoscopies performed by the expert gastroenterologist, 2 were performed under WLI and 2 were performed using NBI. Following each colonscopy, the endoscope was completely removed from the simulator, the interior of the colon agitated so as to again, prevent the training set from being biased due to repeated specific blood patterns.

In order to acquire the data from the magnetic tracking sensor and the data from the endoscopic column, control software was written in C++ using the NDI API and OpenCV. A trial consisted of at least a full traversal from the sigmoid colon to the cecum, and a subsequent return to the sigmoid colon. In concordance with the previous experiment, optical flow descriptors were calculated as described in Section 3.2 for grid-based and lumen-centered partitioning. However, two amendments were made to the processing of the NBI images. It was discovered that histogram equalization was ineffective in producing an image with increased contrast for determining the position of the lumen; thus, this step was ignored. Additionally, the image mask applied to the image slightly differed from the previous experiment due to the shape of the endoscopic image returned and the location of the patient information, which obscured some parts of the image.

```
Result: Trained ANN
Select desired illumination modality
Read in image $I_t$
Record magnetic tracker sensor position at time $t$
while Not finished with colonoscopy do
    Read in new image $I_t + \Delta t$
    Generate and record optical flow-based feature vector
    Record magnetic tracker sensor position
end
Calculate change in pose ($\Delta P_{target}$) from ground truth to be used as target vectors for ANNs
Generate optical flow input descriptors over entire training set
while Validation training error has not increased for six epochs do
    Forward propagate input vector through ANN to get estimated pose$\Delta\mathbf{P}_{predicted}$
    Backpropagate error to train network
end
```

Figure 4.4: Amended algorithm for training set generation and training of ANN in clinical setting.

One additional change was made to the procedure, which is reflected in an updated algorithm detailed in Algorithm 4.4. Instead of calculating the optical flow descriptors during image acquisition, this was performed offline after all the data was collected. This allowed the maximum number of images and corresponding poses of the magnetic tracker to be obtained. The effects of this change do not result in any change that affects the outcome of the results. Specifically, in the previous experiment, the algorithm was able to take as long as necessary to perform the optical flow descriptor before the next actuation movement was made; the offline processing of this experiment therefore results in the same behavior.

The procedure necessary to train the ANNs was maintained; each of the ANNs was trained using the optical flow descriptor inputs and the calculated translational and rotational displacments of the magnetic tracking system. The training data consisted of one trial from the WLI or NBI colonoscopies. The remaining trial was used as testing data to evaluate the performance of the ANNs when trained on more noisy data. The RMSE measure was again used as a metric to evaluate the estimation capabilities of the ANNs, and was calculated as the RMSE between the estimates produced by the ANNs and the known pose displacement given by the magnetic tracker.

4.4 Assessment of Illumination Modality on Feature Strength

Given the dataset acquired in the previous experiment (Section 4.3), a quantitative comparison of the strength of the features extracted from WLI and NBI can be made. This was performed using features calculated in the ST algorithm, which establish a criteria for estimating the strength of trackable corners and edges [46]. This well-established and well-known feature detector determines strong features by calculating the eigenvalues of a pixel of interest using its local neighborhood. In this way, two types of features can be identified – corners and edges. Corners are indicated when the eigenvalues are large (i.e., there is a large variation in both horizontal and vertical direction), and edges are indicated by one large eigenvalue.

To evaluate the strength of features, and also consider the role of each individual component channel, a single image was first divided into its red, green, and blue channels. A grayscale version was also employed as a metric to evaluate the validity of using grayscale instead of full color images. The ST algorithm was run on each of these 4 image components to find the locations of the good features, and using these points, determine the maximum eigenvalue for each of the images. This was repeated for 200 images collected from the dataset acquired from the experiment performed in Section 4.3 for both WLI and NBI. Using these results, the strength of the features based on illumination as well as the contribution of their individual color channels can be assessed.

4.5 Investigation of Feature Descriptor

The performance and estimation accuracy of ANNs is highly dependent on the characteristic of the function being approximated and the consistency of the input/output pairs. For this reason, detection of stable features for consistent optical flow vectors and the formulation of the optical flow input descriptor is essential to accurate ANN estimates. In this experiment, we first compare the impact of defining the optical flow in a scene with the $[\overline{dr}, \overline{d\theta}]$ parameters versus the $[\overline{dx}, \overline{dy}]$ parameters. We then evaluate the effectiveness of the feature descriptors by comparison of the variation within and between certain distance classes. Additionally, we investigate the use of Principal Component Analysis (PCA) to reduce the dimensionality of the input vectors, retaining only the principal components conferring 97% of the variation.

Following this evaluation of the descriptors, we then investigate several different common and state-of-the-art optical flow calculation techniques, and observe their effect on the accuracy of the resultant ANNs. A flow diagram describing the algorithm taken to reach this purpose is shown in Figure 4.5. This is in contrast to Figure 3.1, which shows only one optical flow method being investigated. Additionally, the impact of illumination is no longer considered in this experiment, and WLI is adopted as the sole illumination modality.

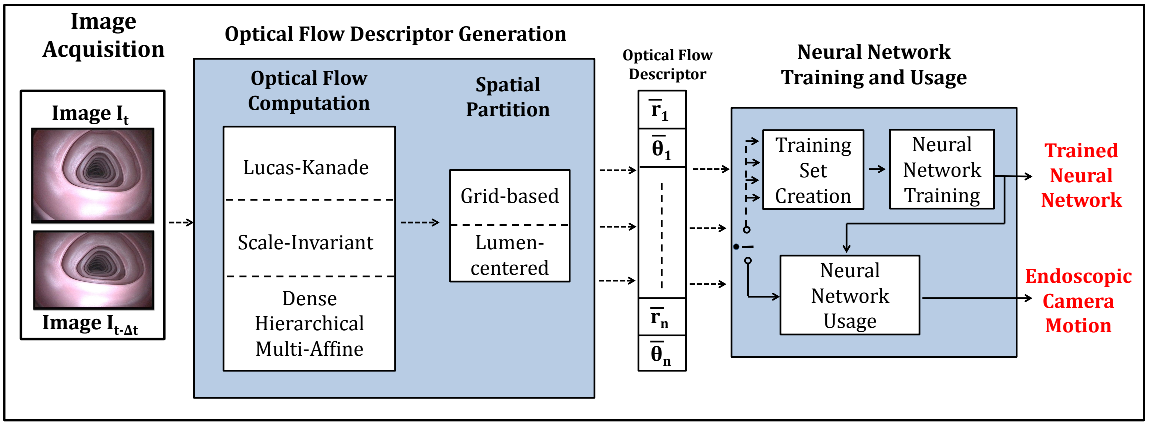


Figure 4.5: Algorithm for comparing ANNs based on state-of-the-art methods of computing optical flow.

A robotically controlled setup, shown in Figure 4.6, was used in order to evaluate the performance of the algorithm in an artificial scenario (a straight endoscope motion in the colon) and by attaching the endoscope to a robot's end effector to precisely record its motion. This was done in order to mimic the most common and likely movement of a teleoperated robotic endoscope, and also obtain a result independent of unwanted complexities. The experimental setup is similar to the setup described in Section 4.2, but utilizes a commercial endoscope (13803PKS; Karl Storz GmbH and Co.; Germany) rather than an endoscopic camera, and does not mount a magnetic tracker onboard the module. The endoscope is rigidly attached to the industrial robotic manipulator.

The mechanism is again actuated using a control software written in C++, at each time step sending positional commands to the robot to increment its position. The robotically

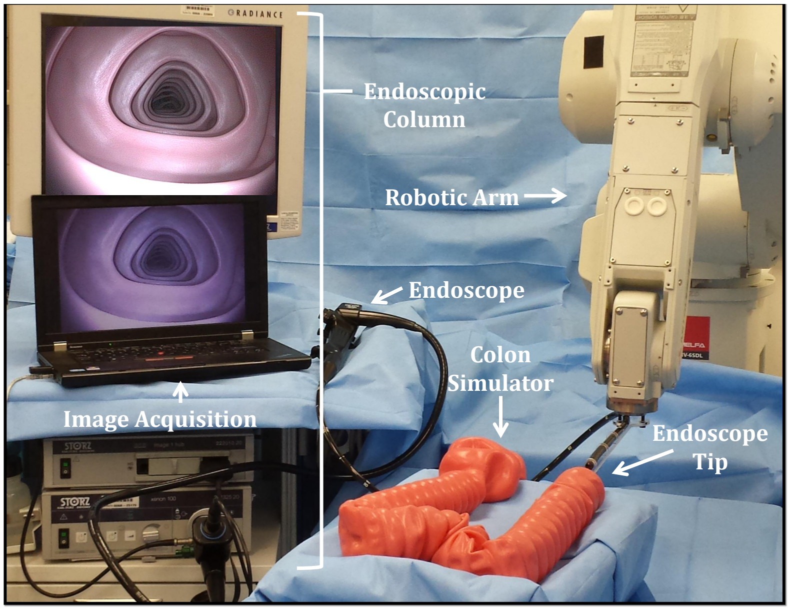


Figure 4.6: Experimental setup replicating movements of a teleoperated endoscope within a human colon simulator for comparison of optical flow generation algorithms.

actuated endoscope moves along the optical axis of the endoscope, which is aligned with the x-axis of the robot, then laterally along the y-axis, and finally vertically, along the z-axis. The algorithm employs a one second delay in order to avoid any possible oscillation of the camera during image acquisition. The data was collected in a similar manner to the experiments described in both Section 4.2 and Section 4.3, and the training set acquisition algorithm is described in Algorithm 4.7. In this experiment, the robot pose is collected at each iteration; however, no magnetic tracker readings are collected, and the composition of the optical flow descriptors is computed offline. As in the previous experiments, the image at the time corresponding to the robot data acquisition was acquired. The endoscope moves within a human plastic colon simulator (shown in Figure 4.6 in increments of $\sim \pm 0.3$mm from 0.3mm to 4mm for 10 iterations each along the x-axis, and in increments of $\sim \pm 0.15$mm from 0.15 to 2mm for 10 iterations each along the y and z axes. This resulted in a data set composed of 260 input/output vector pairs for each DOF.

```
Result: Trained ANN
Read in image $I_t$
Record robotic encoder pose at time $t$
while Not finished with training trajectory do
    Read in new image $I_{t+\Delta t}$
    Record robotic encoder pose
end
Calculate change in pose ($\Delta P_{target}$) from ground truth to be used as target vectors for ANNs
Generate optical flow input descriptors over entire training set
while Validation training error has not increased for six epochs do
    Forward propagate input vector through ANN to get estimated pose $\Delta\mathbf{P}_{predicted}$
    Backpropagate error to train network
end
```

Figure 4.7: Amended algorithm for training set generation and training of ANN for comparison of efficacy of optical flow techniques.

4.5.1 Evaluation of Descriptor Parameters

Section 3.2 describes the summary statistics used for descriptions of the optical flow vector for particular regions of an image scene; however, these different representations may have an effect on the performance of the ANNs. In order to evaluate the magnitude of this difference, 2 additional optical flow descriptors were formed: a combination of lumen-centered partitioning, whose optical flow is calculated using $[\overline{dx}, \overline{dy}]$, and grid-based partitioning, whose optical flow is calculated using $[\overline{dr}, \overline{d\theta}]$. This is combined with the original two descriptors to assess the role of the descriptor representation on the estimation capabilities of the ANNs.

In order to train the ANNs, the datasets were divided into 2 mutually exclusive groups, where 75% of the data was used as training data, and the remaining 25% was used for testing. Each of the groups was populated by randomly selecting examples from the dataset. The same datasets were used for the training and testing of each ANN. In order to reduce the likelihood of converging to a local minima, 100 ANNs were generated for each partitioning/descriptor representation pair. The best ANN for each combination was selected for comparison based on the lowest RMSE obtained.

4.5.2 Analysis of Class Variation

A multivariate analysis of class variation was performed to estimate the separability of the data based on the descriptors. To do this, the data were divided into 72 different classes, based on the DOF in which the endoscope was moving, and the distance travelled. For example, there are 26 different increments moved along the x-axis (i.e., 0.3mm, -0.3mm, 0.6mm, -0.6mm, etc.), and 10 iterations done of each movement. Thus, in the x-direction, there are 26 classes composed of 10 examples each. The length of each example is defined by the partitioning method (i.e., grid-based partitioning has length 50, whereas lumen-centered partitioning has length 10).

In order to evaluate the separability of classes, within-class and between-class variation was calculated. The within-class variation **W** and between-class variation **B** are given as:

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mathbf{y_{ij}} - \overline{\mathbf{y_i}})(\mathbf{y_{ij}} - \overline{\mathbf{y_i}})' \tag{4.5.1}$$

$$\mathbf{B} = n \sum_{i=1}^{k} (\overline{\mathbf{y_i}} - \overline{\mathbf{y}})(\overline{\mathbf{y_i}} - \overline{\mathbf{y_i}})' \tag{4.5.2}$$

where k is the number of classes, and n is the number of examples in the class. The other expressions are calculated as:

$$\mathbf{y_i} = \sum_{j=1}^{n} \mathbf{y_{ij}} \tag{4.5.3}$$

$$\mathbf{y} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mathbf{y_{ij}} \tag{4.5.4}$$

$$\bar{\mathbf{y}}_\mathbf{i} = \frac{\overline{\mathbf{y_i}}}{n} \tag{4.5.5}$$

$$\bar{\mathbf{y}} = \frac{\mathbf{y}}{kn} \tag{4.5.6}$$

where $\mathbf{y_i}$ and $\mathbf{y}$ are the column-wise sum of the examples in the class, and all the examples, respectively; then, $\overline{\mathbf{y_i}}$ and $\bar{\mathbf{y}}$ are the mean of each class, and the overall mean. Given these definitions, $\mathbf{W}$ and $\mathbf{B}$ were calculated for each degree of freedom, with each increment representing a separate class; and over the entire dataset, with each DOF representing a separate class. For each case, Wilks' Λ statistic, was calculated as:

$$\boldsymbol{\Lambda} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}. \tag{4.5.7}$$

The within-class and between-class parameters were then tested by assembling a set of classes with a common feature to create a smaller dataset. To determine the relationship between the variance within the 10 increments tested for a single class to the entire dataset for a DOF, 3 smaller subset datasets were formed, representing each DOF. The Λ obtained from this dataset reflects the ability to distinguish between different increments for a single DOF. Another subset was formed comprising the entire dataset, where each DOF represented a class. This dataset will reflect the ability to distinguish in which DOF a movement occurred.

10 more subsets were formed for the [±0.3mm, 0.6mm, 0.9mm, 1.2mm, and 1.5mm] groups. Each of these subsets contain a total of 30 examples, 10 from each DOF for the

selected increment moved. This information will describe how well the same increment can be distinguished among different DOF. Further details regarding multivariate variance analysis can be found in [77].

4.5.3 Effects of Dimensional Reduction

Dimensional reduction can be advantageous for applications whose samples have a large number of elements. This can be particularly beneficial in the case that ANNs are used; a higher dimensionality generally confers a more complex error surface, which can result in the ANNs converging to local minima instead of the global one. Additionally, ANNs are very sensitive to the training data; the more consistent the data, the better the ANN is able to approximate the function. Lastly, a high dimensionality input vector requires more ANN units and thus a longer time and more computation is necessary for training the ANN.

Dimensional reduction results in a loss of information; in order to make sure the most relevant information is obtained, PCA was used as a method to transform the data into a lesser dimension, while retaining the most variability in terms of variance/covariance. To do this, a new coordinate system is created based on the eigenvalues and eigenvectors of the original data; using the largest eigenvalues and corresponding eigenvectors, a basis is formed which represents the axes of maximum variability in the dataset. For this work, the dimensionality was chosen based on the number of eigenvectors required to explain 97% of the variation in the data.

This was done by first preprocessing the data by mean centering and unit variance. Mean centering removes a bias in the variables and introduces a common point of reference for all dimensions of the data, and creating a common unit variance to account for the range of the dimension. The new scaled version s of an element x in a dimension i is therefore:

$$s = \frac{x - \mu_i}{\sigma} \tag{4.5.8}$$

Following this operation, the covariance matrix $\mathbf{C}$ for the transformed data set $\mathbf{S}$ is:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{S}^T\mathbf{S}, \tag{4.5.9}$$

where n represents the number of examples in the data set. Eigenvalue decomposition is performed on the square covariance matrix to yield a diagonal matrix of eigenvalues $\mathbf{\Lambda}$ and corresponding eigenvectors $\mathbf{Q}$ such that $\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$. The matrices are rearranged such that the diagonalized eigenvalues appear in decreasing order. In order to determine the minimum number of principal components p_{min} required to account for 97% of the data, the ratio r was calculated as:

$$r[j] = \frac{\sum_{k=1}^{j} \lambda_k}{\sum_{m=1}^{n} \lambda_m}, \tag{4.5.10}$$

where j is the index of the eigenvalue. The value j which corresponded to an r ratio closest to but no less than 0.97 was considered to be the number of eigenvector basis vectors required to represent the data set. The value j and eigenvectors $\mathbf{Q}$ were calculated for each of the 4 data sets derived from combinations of partitioning (grid-based vs. lumen-centered) and descriptor representation ($[\overline{dx}, \overline{dy}]$ vs. $[\overline{dr}, \overline{d\theta}]$).

For training and testing of the ANNs, 75% of the data was used for training, and 25% of the data was used for testing; these two sets are mutually exclusive. The training and testing datasets were selected randomly from the overall dataset. The training and testing of the ANNs proceeded in a manner identical to the methods in the previous experiments. In an effort to increase the likelihood of converging to a global minima, 100 different ANNs were trained for each partitioning/descriptor representation combination. The best ANN (i.e., the one with the lowest RMSE) was used for comparison in the results.

4.5.4 Comparison of Optical Flow Algorithms on Performance

In order to reduce the effects of interference between DOFs, a single DOF was selected for usage as the dataset in this experiment. The DOF chosen was the X DOF along the optical axis, as it is arguably the most important DOF necessary for accurate pose estimation. This large dataset is divided into two mutually exclusive groups for training and testing data. Two-thirds of the data was used to train the neural network, and the remaining third was used for testing the performance. Again, the procedure necessary to train the ANNs was maintained; each of the ANNs was trained using the optical flow descriptor inputs and the calculated metric relative motion given the robot's encoders. The mean and standard deviations were used as a metric to evaluate the estimation capabilities of the ANNs, and was calculated as the mean and standard deviation between the estimates produced by the ANNs and the known pose displacement given by the ground truth information given by the robot's encoders. In this experiment, the same approach was taken in order to reduce the effects of Levenberg-Marquardt convergence to local minima; 100 ANNs were generated for each optical flow algorithm/partitioning method. Using the the means and standard deviations as a basis for evaluation, the best ANNs were chosen as representatives for comparison of the ANNs' performance of the motion estimations.

CHAPTER 5

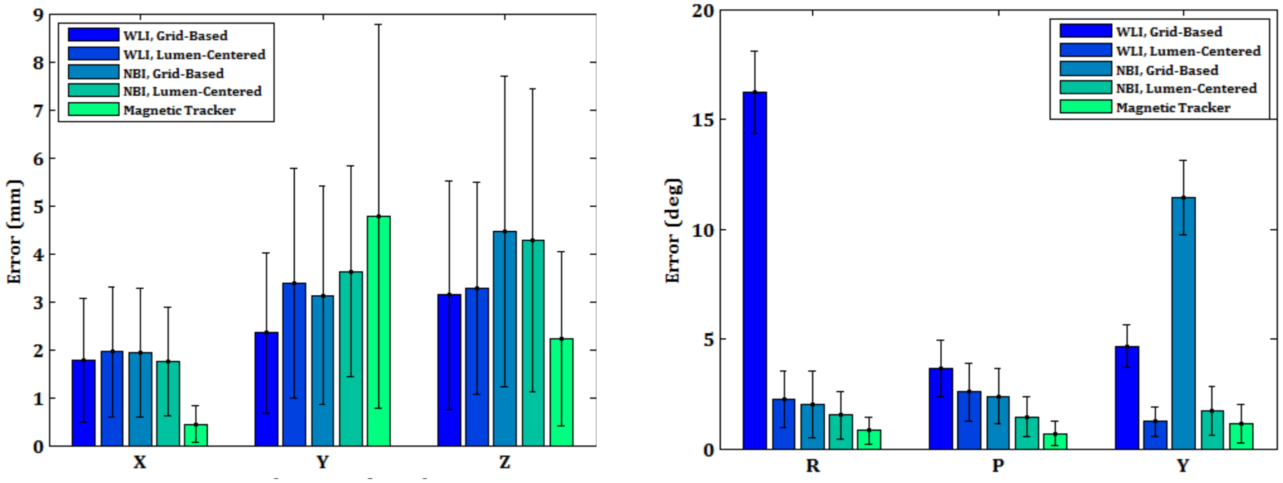
RESULTS AND DISCUSSION

5.1 Benchtop Comparison of ANNs vs. Commercial Magnetic Tracker

Figure 5.1 shows a comparison of the performance of the ANNs and magnetic tracker vs. ground truth over the entire testing set. For all four types of ANNs for all positional DOF, the RMSE is less than 5mm. The ANNs best estimate the motion along the x-axis (i.e., the optical axis of the endoscopic module), which is arguably the axis requiring the most accurate estimates. Although the ANNs perform similarly, the grid-based partitioning method using WLI is the best performing ANN. This is confirmed in the Y and Z DOF. Of particular interest is the Y DOF, in which all of the ANNs outperform the commercial magnetic tracker.

On the other hand, the best performing ANN for the orientation DOF was the ANN combining lumen-centered partitioning with NBI. This resulted in an ANN with RMSE of less than 1.7° for each DOF. In the yaw DOF, the WLI/lumen-centered partitioning ANN slightly outperforms the NBI/lumen-centered partitioning ANN. In the orientation DOF, the commercial tracker performs better than the ANNs; however, the average difference is a somewhat trivial 0.7°.

An important result of this trial is the comparability of the algorithm to data produced by the magnetic tracker, which is used in a setting similar to that of a teleoperable flexible



(a) RMSE for translational DOF

(b) RMSE for rotational DOF

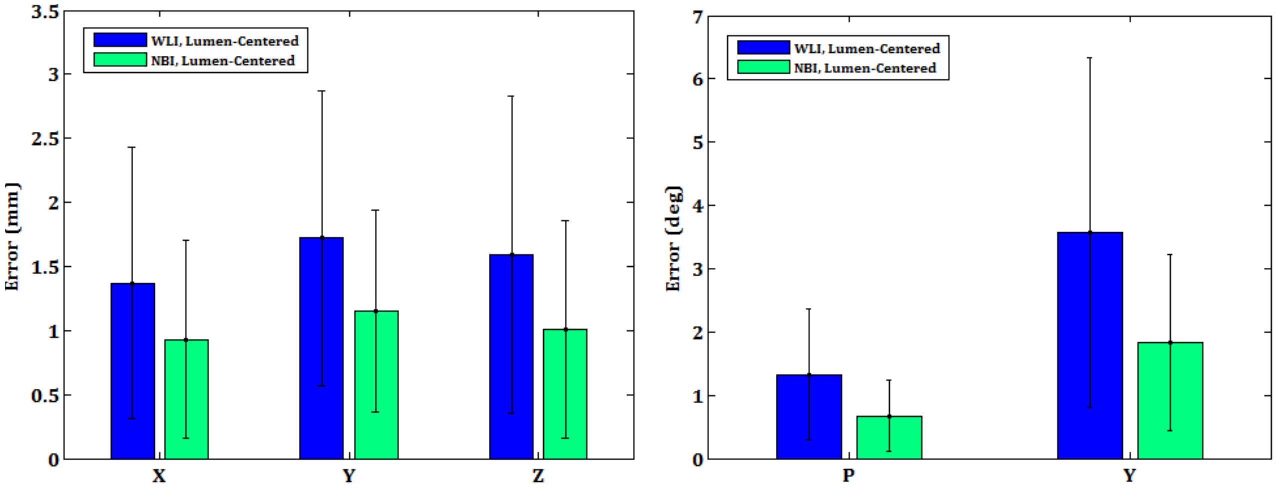
Figure 5.1: RMSE and standard deviation of the ANN variations and state-of-the-art magnetic tracker with respect to ground truth based on robot encoder readings.

endoscope. It is likely that the degraded performance of the tracker is related to its proximity to the actuating motors of the robot; however, the endoscopic module is approximately 300mm from the nearest motors, which are not even used when actuating the translational DOF.

5.2 Clinical Validation Trial with Commercial Endoscope

Figure 5.2 shows a comparison between the performance of the ANN variations for position and rotation DOF. Although the ANNs are able to predict the full 6 DOF given a measurement device that reports 6 DOF, the magnetic tracker used is only able to report 5 DOF. Approximately 1% of the data was removed to account for outliers present in the ANN estimations. Furthermore, the results for grid-based partitioning are not shown in Figure 5.2 since their error is up to an order of magnitude greater than that of the lumen-centered partitioning approach.

A comparison of the illumination modalities for lumen-centered partitioning reveals that ANNs trained under NBI conditions are able to constantly achieve slightly better perfor-



(a) RMSE for translational DOF

(b) RMSE for rotational DOF

Figure 5.2: RMSE and standard deviation of ANN pose estimations against magnetic tracker readings during clinical evaluation of the algorithm.

mance in terms of accuracy and precision than WLI ANNs. Thus, lumen-centered partitioning combined with NBI can be considered to be a superior mechanism to WLI for a vision-based motion estimation in this application. However, this outperformance is minimal; the possibility for convergence to local minima of the ANNs, and importance of the randomly initialized weights should be carefully considered before drawing conclusions about the efficacy of the approach. This is especially important considering that both approaches have comparable RMSE, which is less than 2mm in for positional DOF and 3° in orientation DOF.

An essential result of this trial is evidence that ANNs can be trained on noisy data, and still produce estimates similar to that of when ground truth is known very accurately. This is shown most clearly by the results in the X, Y, and yaw directions. This reflects the noise filtering and generalization capabilities of ANNs. Furthermore, by performing an experiment similar to a clinical setting with a commercial endoscope handled by an expert gastroenterologist, this verifies that this approach is relatively robust. During this trial, the gastroenterologist cleaned the lens using the endoscope's water channel due to blood sometimes obscuring the image, and the endoscope was moved sharply and suddenly. All of these elements produce significant noise and disturbances in the image; obscuring the image with blood and cleaning the lens even violates the static scene assumption. Even further, the effect of the roll movement of the endoscope was essentially filtered out from the data set; the 5 DOF sensor was unable to report this DOF, so it could not be measured or accounted for within the algorithm. Regardless, the algorithm was able to perform satisfactorially despite these challenges.

A major contributor in the pose estimation error is likely due to these types of noise, which obscure the image and violate essential constraints governing the behavior of elements within the algorithm. It is essential to note that this experiment was performed by an expert gastroenterologist, not a robotically controlled endosocope; since a higher degree of control can be obtained using a robotically actuated endoscope, the noise introduced by sudden and fast movement can be eliminated by placing constraints on the types of movements the control system is able to command. This will create smooth, controlled movements.

With a robotically actuated endoscope, algorithms can also be employed to pause motion estimation during periods in which the lens is being cleaned. Again, due to the robotic nature of the device, these events can be sensed from the input device commanded by the user so that motion estimation can be paused while the endoscope is forced to remain motionless.

One final source of error that cannot be overlooked is turning around corners encountered during the endoscopy. This is probably a major contributor to the reduced performance of the grid-based partitioning when performed along the entirety of the colon, especially compared to its accuracy when used in estimations along straight sections of the colon as tested in Section 4.2.

Finally, an important quantity to note when considering real-time motion estimation is the time required for the algorithm to execute and the estimation to be made. Given the frame $I_{t-\Delta t}$, the time required to both acquire the current frame I_t and return a motion estimate is approximately 280 ms during the highest magnitude of movement tested for lumen-centered partitioning, the most computationally expensive out of the two partitioning methods. However, this value is strongly influenced by the number of ST feature correspondences found during each iteration. The maximum number of features allowed in this implementation was capped at 10,000. Additionally, this algorithm was tested using a standard laptop (Lenovo Thinkpad T520, Intel Core i5-2520M CPU at 2.50 GHz, Windows 7 Professional; Lenovo; USA) with unoptimized code. Thus, it is expected that the computational time will be reduced by parallelizing the computation of ST feature correspondences (e.g., by utilizing different processor cores to simultaneously process different regions of the image) and ANN computations, optimizing the code, and using a more capable computer.

5.3 Strength of Features based on Illumination Modality and Color Channel

Figure 5.3 depicts the optical flow vectors created by a 5mm translation along the optical axis of the endoscope during the first experiment (described in Section 4.2). This figure provides a visual comparing the appearance of the colon under the different imaging modalities and spatial partitioning methods. Visual inspection reveals a definitive pattern in the optical flow due to the movement of the camera relative to the static environment.

An analogous image set is shown in Figure 5.4, but the WLI and NBI pictured here are provided by the commercial endoscope. A comparison between Figure 5.3 and this figure demonstrate a visual difference in the appearance of the colon due to utilizing the artificial NBI LEDs versus the NBI provided by the commercial endoscope. Regardless, the effect of the illumination modality is evident; qualitative inspection of the image shows that blood features are more prominent and distinguished using NBI as compared to those of WLI. In the figures shown, the partitioning lines are drawn to illustrate the delineation of regions under the different partitioning methods; however, in practice the division is not created on the image, but only used as boundaries within the processing program.

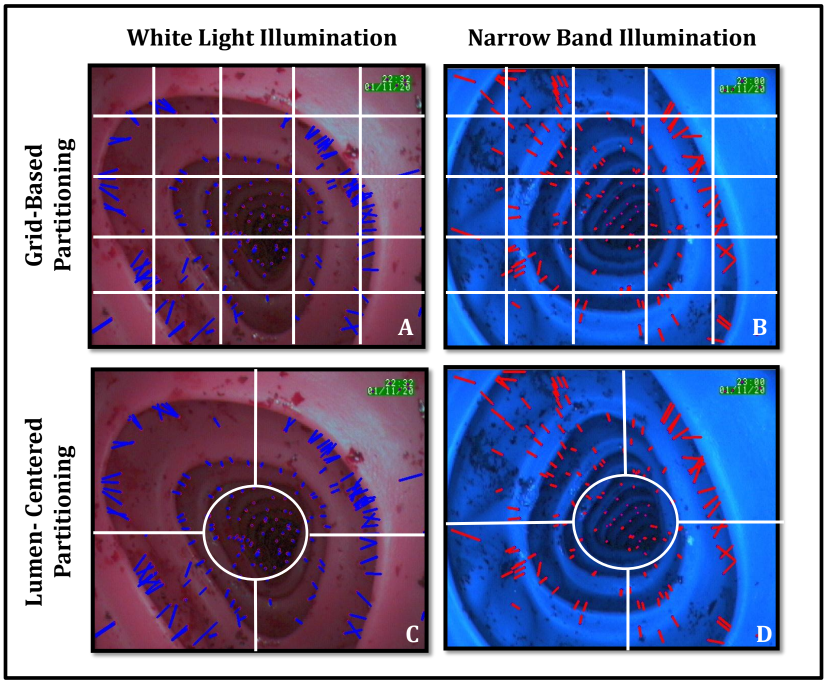


Figure 5.3: Typical optical flow patterns for a 5 mm translation along the Z axis with combinations of illumination modalities and spatial partitions. Tests were performed in a human colon simulator with porcine blood staining.

Figure 5.5 presents a quantitative comparison of the strengths of features garnered from a grayscale representation of the image versus the red (R), green (G), and blue (B) channels for WLI and NBI based on the images obtained from the commercial endoscope. Since corners and edges represent areas of the image with high variation across their local pixel neighborhood, they are indicated by regions with high eigenvalues. As is evident in the figure, NBI creates features with more than twice the strength of WLI features. Although it has a higher variation (i.e., larger standard deviation) in terms of these eigenvalues, even the lowest mean value, given the standard deviation, still far exceeds that of WLI.

The average number of WLI features is approximately 12,600 for all color channels; on average, NBI images have slightly fewer, with approximately 11,700 detected features. This suggested that although WLI is able to produce more features than NBI, they are half the

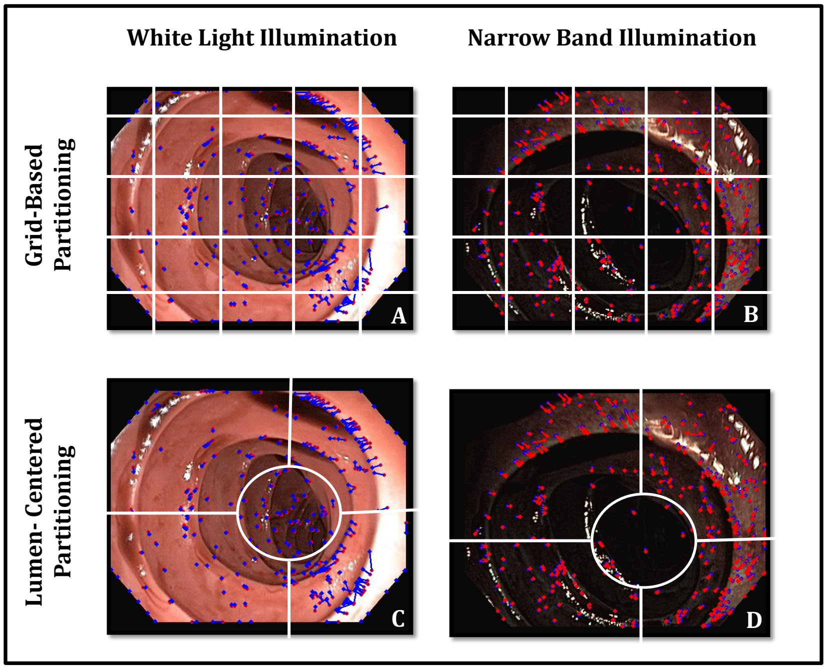


Figure 5.4: A comparison of the image of the lumen of the colon under WLI and NBI using a commercial endoscope.

quality of NBI features. Additionally, this figure demonstrates that utilizing a grayscale image within the algorithm rather than the entire color image is justified, since the feature strengths from the grayscale image have nearly the same mean feature strengths as the other color channels. This allows a faster processing time since a one-channel image can be used instead of a 3-channel image.

These results coincide with those found in Section 5.2, as well as the visual inspection of the images in Figure 5.4, which reflect a higher contrast between the walls of the colon and the blood features under NBI. However, given the improved strength of features under NBI, one would expect a greater disparity between the performance of the WLI and NBI ANNs. This similarity of response is likely due to the ANNs themselves; since ANNs are sensitive to their initial weights, which causes them to sometimes to converge to local minima, it is possible that these ANNs found represent suboptimal function approximations.

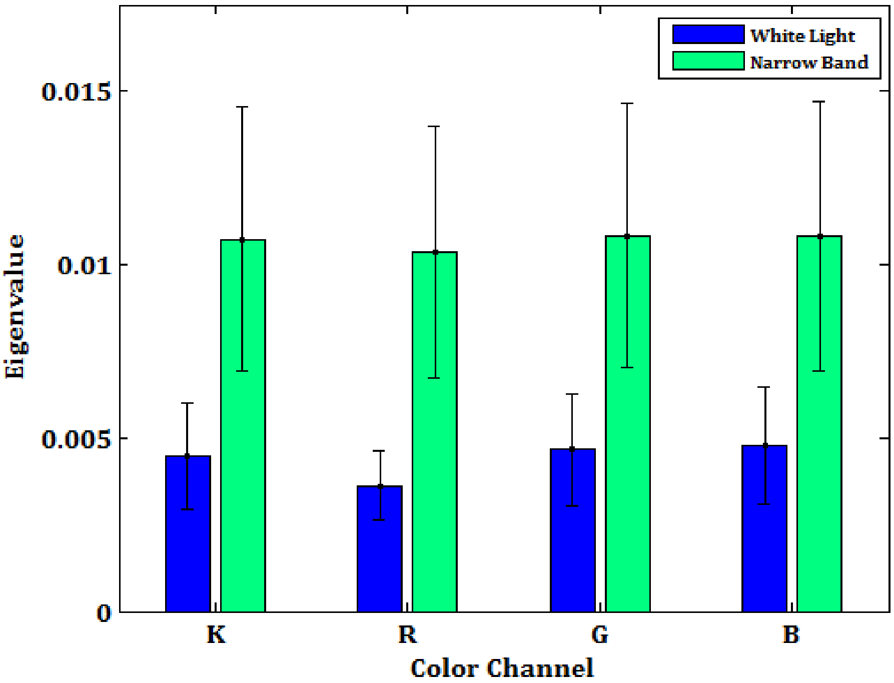
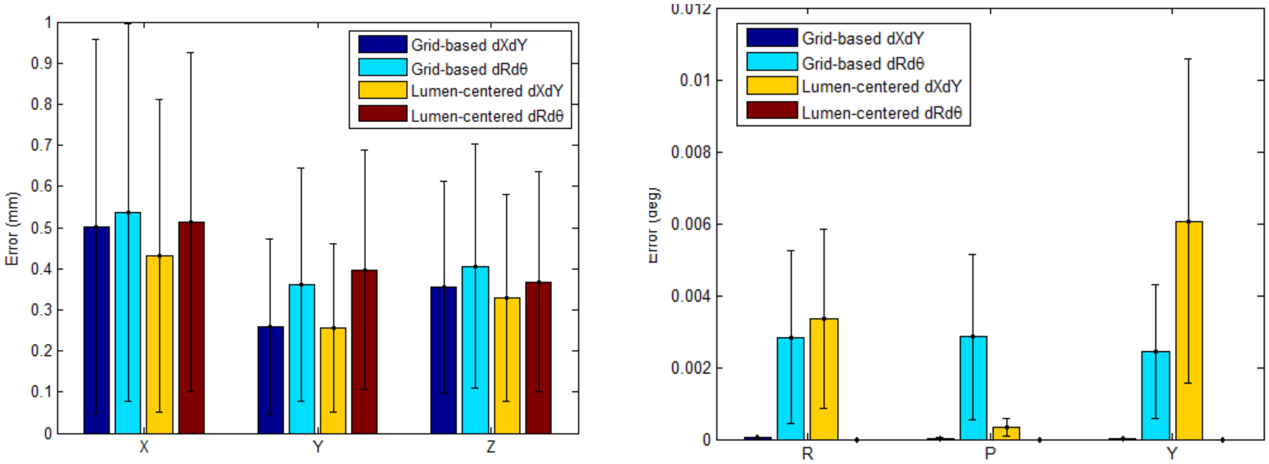


Figure 5.5: A comparison of the strength of features between WLI and NBI per color channel. K represents a grayscale version of the image, R is the red channel, B is the blue channel, and G is the green channel.

5.4 Evaluation of Feature Descriptors

5.4.1 Descriptor Representation

The effects of the parameters used to describe the optical flow vectors are shown in Figure 5.6. For the translation DOF, the best mechanism for partitioning appears to be lumen-centered partitioning, using the $[\overline{dx}, \overline{dy}]$ parameters. In general, the $[\overline{dx}, \overline{dy}]$ parameters appear to produce slightly more accurate ANNs. For the rotational DOF, the best ANN is clearly the ANN resulting from lumen-centered partitioning and $[\overline{dr}, \overline{d\theta}]$ parameters. However, all the errors produced are very low, and should be since there was never rotation in the motion dataset. The reason these errors appear at all is likely due to the initial weights, backpropagation, and early stopping used with the ANNs. Because of these factors, the training likely does not last long enough for the weights to converge to zero.



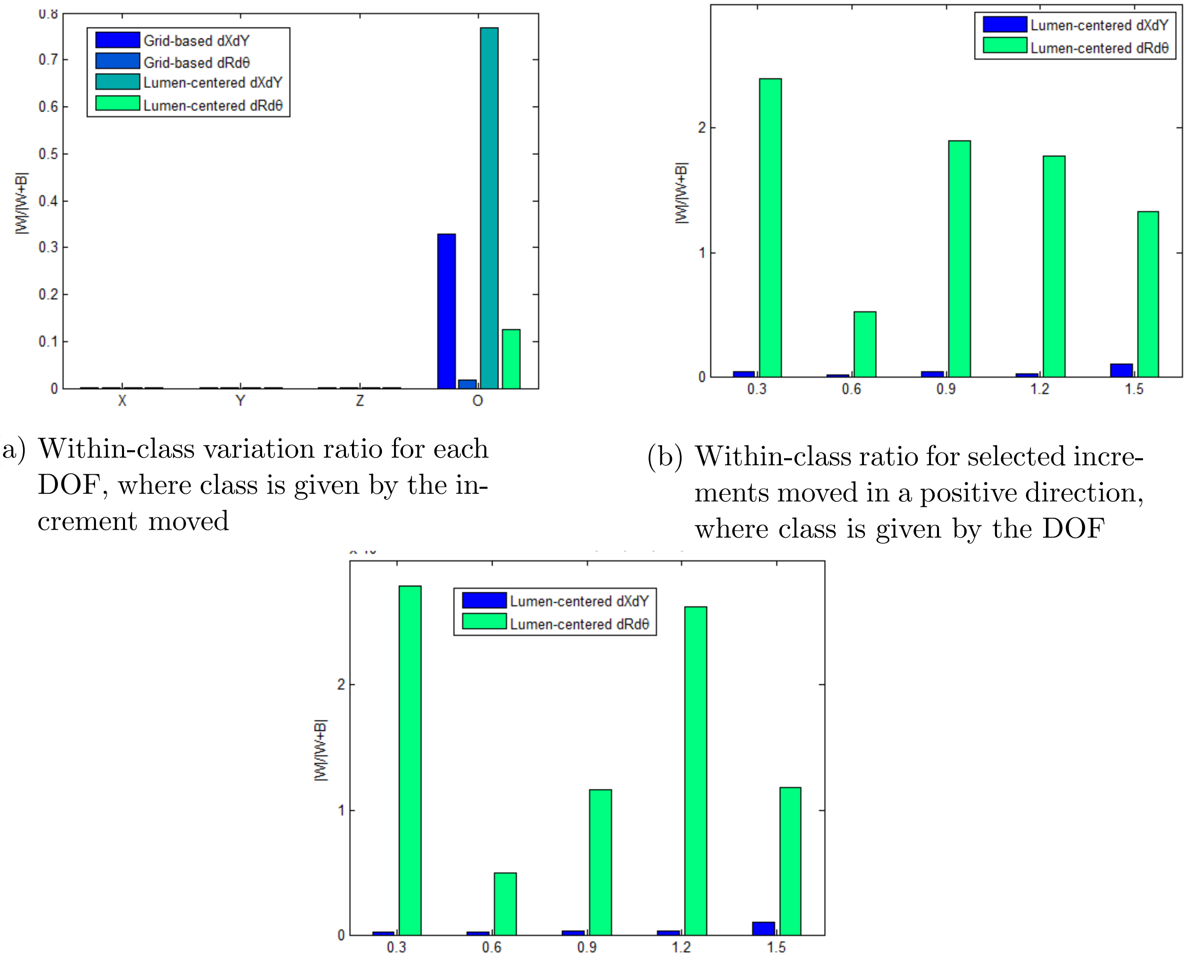
(a) RMSE for translational DOF for ANNs comparing descriptor parameters

(b) RMSE for orientation DOF for ANNs comparing descriptor parameters

Figure 5.6: RMSE for ANNs based on combinations of partitioning and feature descriptor representation

5.4.2 Class Variation

Figure 5.7 shows the Wilks statistic Λ for each of the subsets of the data. As shown in Figure 5.7a, Λ is low for each of the combinations of partitioning and representation. This means that comparatively, each example for a given class is closer to its mean than to the overall subset mean. Consequently, this suggests that given that the DOF being tested is known, it is rather straightforward to distinguish increments of ± 0.3mm. The O subset removes the assumption that the DOF in which the actuation takes place is known; this leads to a much higher Λ value. This suggests that the appearance of movement even within a single DOF varies greatly; thus, it is difficult to distinguish in which DOF a movement occurred. This is particularly pronounced with the $[\overline{dx}, \overline{dy}]$ representation, particularly in the lumen-centered partitioning. This result suggests that the within-class variation (i.e., the variation in appearance for a particular DOF) for this combination accounts for more than 70% of the overall variation in the feature descriptor.



(a) Within-class variation ratio for each DOF, where class is given by the increment moved

(b) Within-class ratio for selected increments moved in a positive direction, where class is given by the DOF

(c) Within-class ratio for selected increments moved in a negative direction, where class is given by the DOF

Figure 5.7: Within-class ratio based on selected subsets of the data.

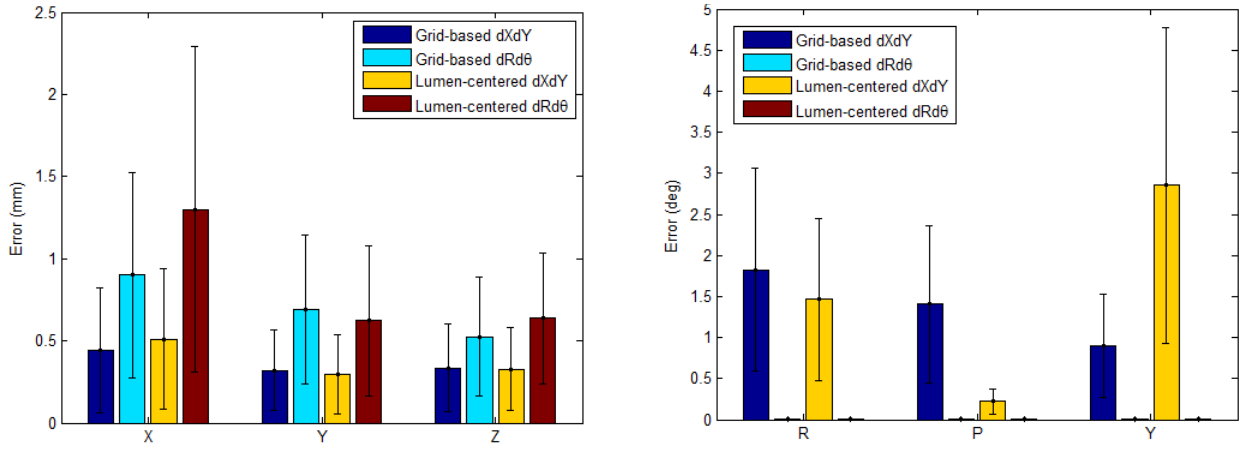
Figures 5.7b and 5.7c show the ability of the lumen-centered partitioning to distinguish to which DOF a given increment belongs. Grid-based partitioning is not shown, since the length of an example (i.e., 50) is greater than kn, or the number of examples in a dataset (30). This leads to a singularity in the within-class matrix. Additionally, these two figures are separated for readability only.

For the $[\overline{dx}, \overline{dy}]$ representation, the within-class variation ratio is generally less than the corresponding data in Figure 5.7a. This suggests that given a certain increment, it is easier to distinguish it from other increments in other DOF of the same distance than if the distance is known. The same is generally true for the $[\overline{dr}, \overline{d\theta}]$ representation. However, as shown, the within-class variation in the $[\overline{dr}, \overline{d\theta}]$ representation appears to be greater. Again, it is important to note that all of these values are on the order of $10^(- 4)$, which is a relatively small variation. The data thus suggests that the grid-based $[\overline{dr}, \overline{d\theta}]$ representation may be the best, since (1) given the DOF, the the distance which was moved can be differentiated from other distances, (2) movement in each DOF can be distinguished from movements in other DOF, and (3) given the distance moved, the DOF in which it occurred can be discerned from other DOF.

5.4.3 Dimensional Reduction

Applying PCA to the datasets determined that the dimension that accounts for 97% of the variation in the data using grid-based partitioning was 27 for the $[\overline{dx}, \overline{dy}]$ representation and 18 for the $[\overline{dr}, \overline{d\theta}]$ representation. For lumen-centered partitioning, the the number of principal components was 6 in both cases.

Figure 5.8 shows the performance of the ANNs with the dimension of the inputs reduced. Compared to the translational DOF in figure 5.6, PCA does not present any significant gains in terms of reduction with RMSE for the $[\overline{dx}, \overline{dy}]$ representation. However, the $[\overline{dr}, \overline{d\theta}]$



(a) RMSE for translational DOF for ANNs comparing descriptor parameters with dimensional reduction

(b) RMSE for orientation DOF for ANNs comparing descriptor parameters with dimensional reduction

Figure 5.8: RMSE for ANNs based on combinations of partitioning and feature descriptor representation with reduced dimensionality

representation resulted in a significant increase in RMSE. For these two cases, dimensional reduction, although retaining 97% of the variation, does not aid in reducing the RMSE.

In terms of the rotational DOF, PCA had a similar effect: exchanging a decrease in RMSE for one combination while increasing the RMSE for another. However, in this case, only the grid-based $[\overline{dx}, \overline{dy}]$ representation was affected, and all the RMSE fell under 3x10^{-3} degrees.

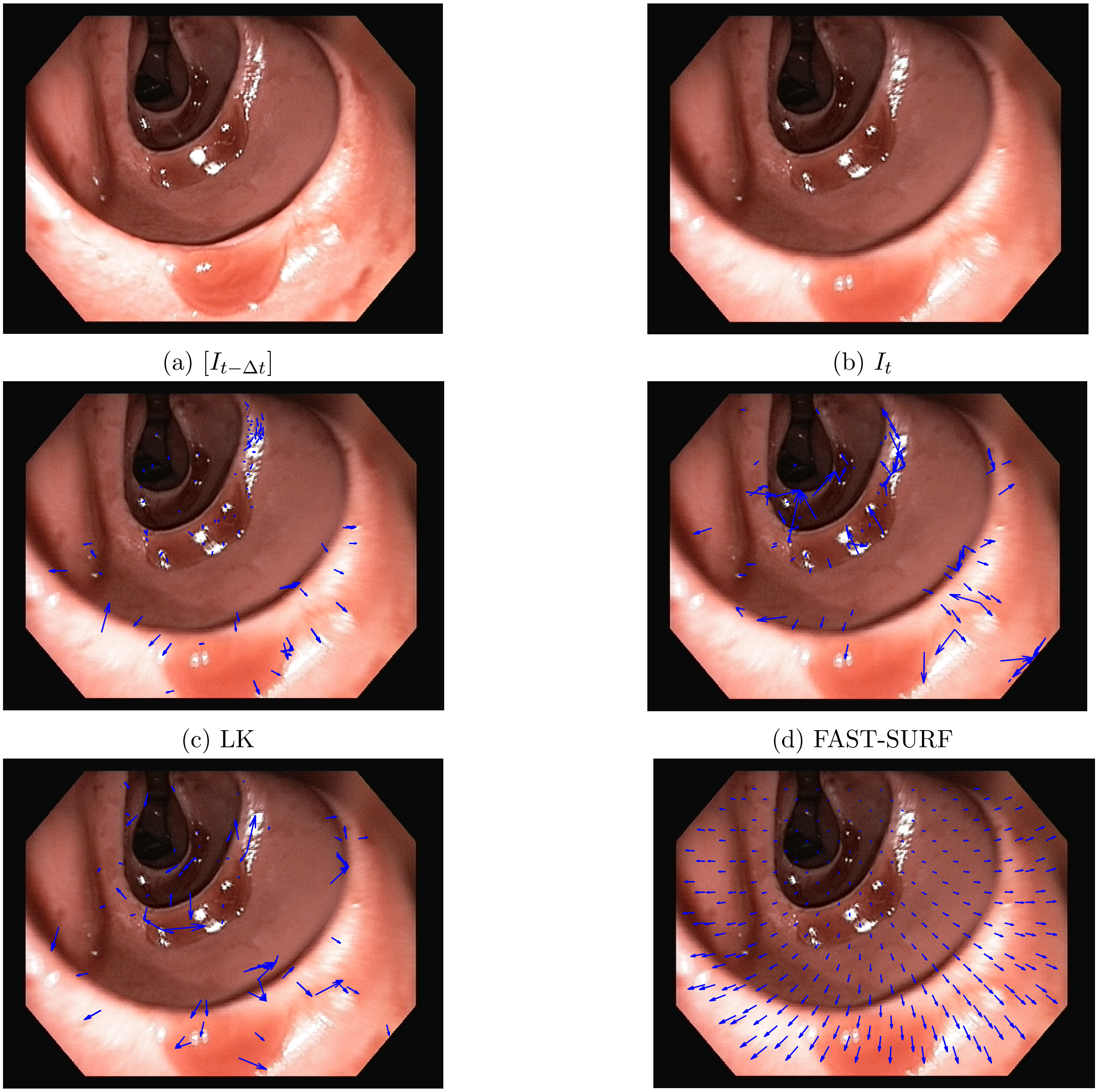
For this particular assignment of the training set, dimensional analysis aided some DOF, while hurting others. The desirable quality is shown by lumen-centered partitioning with $[\overline{dx}, \overline{dy}]$; for each DOF, either the RMSE stayed nearly the same, or it decreased significantly with dimensionality reduction. Because of this, the reduced dimension is equivalent or better than the full dataset; by reducing the number of inputs, the training time of the ANNs can be significantly reduced. For grid-based partitioning with $[\overline{dx}, \overline{dy}]$ representation, it can also be argued that dimensionality reduction may also be useful in practice; the error introduced in the rotational DOF may not be significant enough to pose a problem, although the dimension is cut nearly in half.

It is important to note that these data vary with the training and testing sets randomly populated, and are additionally dependent on the random initial weights, although hopefully the effect of this latter contributor was diminished by the sampling of 100 ANNs. Thus, it will be useful to consider instead of just one training set, several training sets, and training several ANNs on each one. PCA also highlights the importance of a dataset that is representative of the combinations that will be encountered in the testing set, and enough examples such that an accurate representative of the true function can be found.

5.4.4 Optical Flow Algorithm Comparison

Figure 5.9 demonstrates a worst-case scenario in which few feature correspondences can be found within an image. As shown, the sparse methods LK, FAST-SIFT, and FAST-SURF (shown in Figures 5.9c, 5.9d, 5.9e produce very few optical flow vectors, and the ones which are produced have a seemingly random pattern. On the other hand, dHMA (Figure 5.9f) produces a very consistent OF, and can likely provide a better input to the ANNs.

The pose estimation capabilities of each of the optical flow methods combined with grid-based or lumen-centered partitioning is shown in Figure 5.10 for the 1 DOF robotic actuation along the optical axis of the endoscope through the colon simulator. Figure 5.10a reports the mean and standard deviation for the error (i.e., the signed difference between the ground truth provided by the robot encoders and the estimations output by the ANNs) for grid-based partitioning. As shown, the two dense optical flow methods perform similarly, producing an average absolute mean of 1.3mm±1.9mm error along the actuation direction, and definitively outperformed the sparse methods. The maximum mean error produced for the test trajectory was 14.6mm using FAST with SURF descriptors.



(a) $[I_{t-\Delta t}]$

(b) I_t

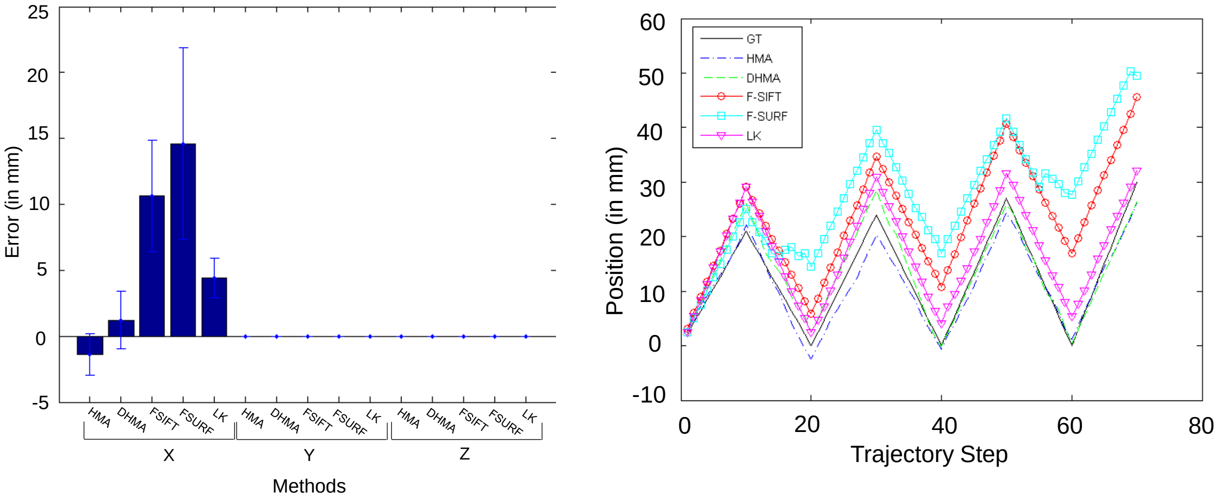
(c) LK

(d) FAST-SURF

(e) FAST-SIFT

(f) dHMA

Figure 5.9: Comparison of optical flow vectors produced in a worse-case scenario.



(a) Mean and standard deviation for translational DOF

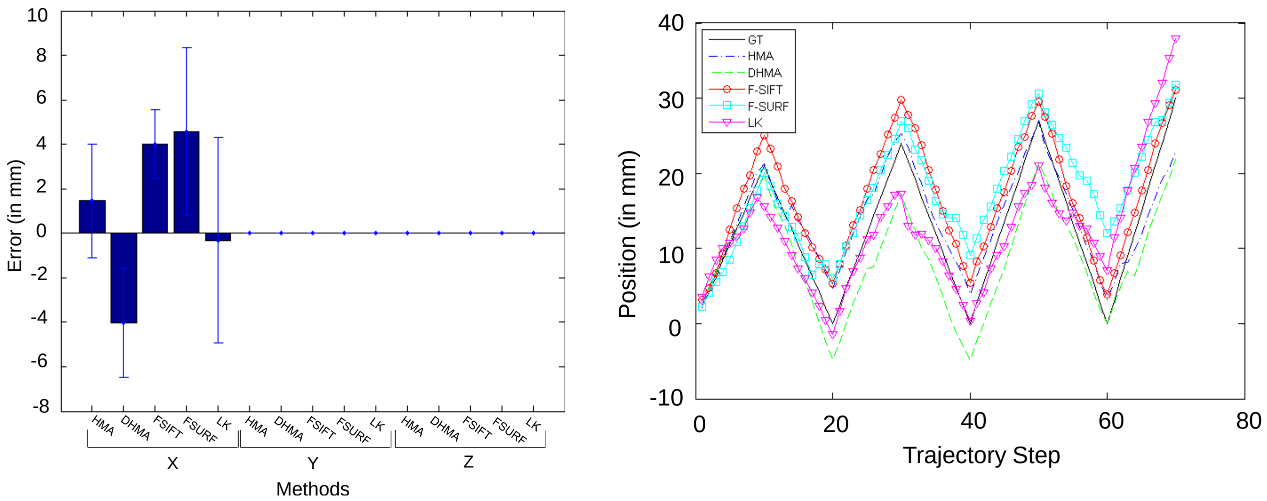
(b) Integrated trajectory of the test trajectory

Figure 5.10: Experimental results with grid-based partitioning combined with sparse and dense optical flow techniques.

Additionally, Figures 5.10a and 5.11a demonstrate that the algorithm is able to accurately able to identify the moving direction since the translation errors along the Y and Z DOF are zero. This means that the ANN weights have been accurately trained to produce a zero output. Not included in the results are graphs for orientation errors, which were also zero in all cases for both grid-based and lumen-centered ANNs.

Figure 5.10b shows the integrated trajectories of the different optical flow methods combined with grid-based partitioning. The details the performance of the algorithm along the entire test trajectory as compared to ground truth. As shown, the dense optical flow techniques perform similarly, with a final difference in endpoint of 3.61mm over a test trajectory 174mm in length. FAST-SURF again performed the worse, with a deviation of 19.51mm from the endpoint given by ground truth.

Figure 5.11 details the performance of the different optical flow methods under lumen-centered partitioning. FAST-SIFT, FAST-SURF, and dHMA all perform similarly, and produce approximately 4-5mm of mean error over the entire test set. However, unlike grid-



(a) Mean and standard deviation for translational DOF

(b) Integrated trajectory of the test trajectory

Figure 5.11: Experimental results with lumen-centered partitioning combined with sparse and dense optical flow techniques.

based partitioning, the LK method achieves the lowest mean error (0.31mm). On the other hand, it possesses the widest distribution of estimates as shown by its large standard deviation (±4.6mm). HMA also provides a low mean error (1.45mm), although it slightly higher than LK, but provides a much more reliable estimate, as shown by its more narrow standard deviation range (±2.55mm).

Although the error statistics for grid-based partitioning were relatively indicative of their performance in the integrated trajectory, this is not the case for lumen-centered partitioning. As shown, the optical flow algorithm which produces the closest endpoint is the FAST-SIFT with a final offset of 1.74mm; however, it has one of the highest mean errors. In contrast, LK, HMA, and DHMA all have endpoints that significantly deviate from the true endpoint, ending at a maximum of 6.9mm away. However, these 3 algorithms correspond to the lowest mean errors.

Overall, the grid-based partitioning method produces the top two ANNs with the lowest error and smallest standard deviation; however, lumen-centered partition appears to have the effect of standardizing the scene for each of the optical flow methods, resulting in a lower overall error for all the optical flow methods. Additionally, this work suggests that given an ANN able to produce estimates with low mean error and standard deviation, the integrated trajectory may also be able to be used as an additional input in a control system, using both together to produce the most accurate estimate.

It is essential to again note the dependence of these results on the training of the ANNs. In order to reduce the impact of the initial neuronal weights, 100 ANNs were tested for each of the optical flow/partitioning combinations. However, it is likely that an even more exhaustive search or alternative training algorithm may be necessary for reliably comparing these methods. Finally, the results of this experiment may be a result of a small training set. As in the previous experiment, a significantly larger dataset may provide the ANNs with a more complete sampling of the function to be approximated.

CHAPTER 6

CONCLUSION

Colorectal cancer affects the lives of millions of men and women worldwide. Although it is nearly always preventable, people avoid the procedure due to fear of potential pain, embarrassment, and the perceived indignity of the procedure. Teleoperable flexible endoscopes introduce the possibility of overcoming these deterrences by imparting better control to the physician and promoting a higher polyp detection rate. These endoscopes can be made more accurate and precise by the introduction of motion estimation algorithms, which are able to provide pose feedback reflecting the positional and rotational movement of the endoscope after actuation. This offers the ability to create a closed-loop control strategy, which can aid in disturbance rejection, noise rejection, and actuation error minimization for flexible devices operating in complex environments. The research presented has the potential to enhance teleoperated and automated endoscopy.

This work first presented an algorithm designed to extract reliable features from an images, estimate the optical flow between the images, and use the resultant patterns of optical flow to train an ANN to reliably compute the metric change in motion. Advantages of the algorithm include that it does not require direct estimation of camera calibration parameters, does not add to the size of the device, and does not create any interference due to the actuation mechanism of the device.

The algorithm was first tested against a state-of-the-art commercial magnetic tracker in a typical robotically actuated operating environment. In this case, the impact of illumination was assessed by creating an endoscopic module attached to an industrial robot, equipped with both white light and narrow band (450nm) LEDs. Methods for best partitioning the image were also explored (grid-based vs. lumen-centered) in order to create the best input vector for the ANNs. All the ANNs achieved positional RMSE of less than 5mm, and in one case, the error in all the ANNs was lower than that of the commercial magnetic tracker. The best combination of illumination and partitioning was WLI with grid-based partitioning (2.42mm RMSE). However, in terms of rotational RMSE, the most accurate ANN was the one using NBI and lumen-centered partitioning (1.69° RMSE). During this trial, the tracker obtained an accuracy of 2.49mm in positional DOF and 0.89° in rotational DOF. With these results, we can conclude that the optical flow-based ANNs have performances comparable to that of a state-of-the-art commercial tracker.

The algorithm was then evaluated based in the manner in which it would be used in a clinical setting. This was achieved by placing a 5 DOF magnetic tracker down the tool channel of a state-of-the-art Olympus endoscope equipped with both WLI and NBI. A colonoscopy training model used for teaching gastroenterologist trainees was then used as a test platform, and the plastic colon simulator within the model was stained with porcine blood. An expert gastroenterologist then performed 4 colonoscopies on the simulator; 2 under WLI and 2 under NBI. For each illumination modality, one trial was used for training the ANNs, while the other was used as a testing set. Again, the 4 combinations utilizing illumination and partitioning were compared. The performance of lumen-centered partitioning with NBI was superior, with 1.03mm ± 0.8mm RMSE in positional DOF, and 1.26°± 0.98° RMSE in rotational DOF, while with WLI, the performance was 1.56mm ± 1.15mm RMSE in positional DOF and 2.45° ± 1.90° RMSE in rotational DOF.

This second experiment additionally investigated the role of the color channels and illumination modality of the strength of the extracted features. Using the images acquired during the 4 colonoscopies, the features extracted via the ST algorithm were compared based on their eigenvalues, a measure of their strengths. A comparison of these eigenvalues revealed that on average, features gathered via NBI were twice as strong as the features extracted under WLI. WLI on average produced slightly more features. Additionally, no significant difference between the feature strengths between the red, blue, and green color channels were observed, and did not vary significantly from a grayscale version of the image. This demonstrates that a grayscale image is sufficient for feature extraction.

Finally, mechanisms for understanding and thus improving the algorithm were then explored. One of the most important parts of the algorithm is establishing the optical flow between two sequential images, since it is the basis of the vector input to the ANNs. Using an experimental platform similar to the first robotically actuated platform, we actuated the robot along each individual DOF, the x-axis, the optical axis of the camera, and the two perpendicular directions - the y-axis and the z-axis. This was performed in a plastic human colon simulator to gather both the training and testing sets.

We first evaluated the role that the optical flow representation had on the results; it was shown that the lumen-centered partitioning method was superior, combined with either of the $[\overline{dx}, \overline{dy}]$ or $[\overline{dr}, \overline{d\theta}]$ representations. Additionally, the within-class variation and between-class variation was evaluated for different subsets of the data. In this case, the grid-based partitioning methods had the most cohesive data classes, and all of the partitioning methods/descriptor representation pairs have low within-class variation, which allows the different DOF classes to be discerned. Also, the multivariate analysis showed that the within-class variance for the increments for lumen-centered partitioning was lower than that of the DOF; this suggests that it is easier for using lumen-centered partitioning to distinguish between the different DOF in which an increment occurs than differentiating between the different increments within a single DOF. Finally, dimensional reduction showed mixed results; although it improved one case and another was arguably comparable, the more than 50% reduction in the dimension of the descriptor input accelerated training of the ANNs.

Additionally, we evaluated 5 state-of-the-art, well-established, and ubiquitous methods for calculating optical flow: LK, FAST-SIFT, FAST-SURF, HMA, and dHMA also for the purpose of optimizing the descriptor representation introduced to the ANNs. For this comparison, only 1 DOF of the dataset acquired in the previous experiment was studied: the X DOF along the optical axis of the camera. Although we did not compare illumination modalities within this experiment, we continued to explore the usage of the two proposed partitioning methods: grid-based and lumen-centered. Using the grid-based partitioning method, the dense optical flow algorithms were the most accurate (1.3mm±1.9mm error; final difference in trajectory 3.61mm over 174mm). The lumen-centered partitioning method resulted in an overall reduction in mean error for all the algorithms. The LK optical flow method produced the lowest mean error, although it had a large standard deviation, which produced large fluctuations and inaccuracies in the integrated trajectories. HMA provides the best performance with lumen-centered partitioning (1.45mm±2.55mm); it has the smallest standard deviation, and is able to closely recreate the trajectory as given by ground truth.

Overall, this experiment demonstrates that the grid-based partition produces the two best performing ANNs based on means and standard deviations; however, the lumen-centered partition appears to have a significant effect on the optical flow descriptors, which produces an overall lower mean error.

It is worth mentioning that the ANNs trained within these 3 experiments are not interchangeable; although estimation of camera calibration parameters and distortion coefficients is not strictly used within the algorithm, the effects of these parameters are inherently present in the optical flow descriptor input to the ANNs. Furthermore, each ANN should be trained on the data which it expects to see during usage; thus, using the ANNs trained on the straight trajectories of the colon would be inappropriate for usage on a training set which includes the corner folds or irregularly spaced or oriented haustral folds. This raises the question: practically, how is the algorithm trained for teleoperable endoscopes?

The training portion of the algorithm is meant to be performed once in the lifetime of the endoscope, provided that the camera optics/illumination do not change significantly. Initial training of the algorithm would proceed by an endoscopist acquiring training data in a method similar to the second trial. However, the numerical training and usage of the ANNs would proceed in an automated manner via software.

6.1 Future Work

Although this work cannot be used for detecting color perforation or looping, it is a novel method which uses components native to the endoscope for egomotion estimation of the endoscopic camera. Future work includes even further exploration of inputs to the ANNs, including different types of segmentation (e.g. watershed, graph cuts) for better description of the optical flow, exploration of RGB color features [78], and other aggregated and custom features. The ANNs themselves will be further explored, including understanding an optimal size for the training set, and better training and selection mechanisms for finding and selecting the "best" ANN. Additionally, alternate methods of regression for performing the motion estimation based on training will be explored. The algorithm will also be implemented as part of a real-time closed-loop feedback control system for a teleoperated flexible endoscope platform to establish feasibility for remote manipulation of a teleoperated platform.

Future work will also include *in vivo* trials, repeating the experiments inside living colons. Porcine model experiments using a commercial NBI endoscope will enable a better and more accurate description of features produced by NBI due to blood vessels. This will also provide us with a better training set for defining optical features and feature descriptors. *In vivo* human trials are also essential, as this will allow for the testing of certain strong assumptions made in this work; particularly, we will be able to quantify the effect of the variance of the appearance and size of the colon among different patients on the algorithm. This will enable us to determine to what extent re-training/calibration of the ANNs are required, and how to perform this automatically. *In vivo* trials will also enable an opportunity to assess the robustness of the proposed method against haustral contractions and insufflation. The expected approach is to freeze endoscope motion during haustral contractions and insufflation, and resume motion estimation after the contraction finishes.

This work demonstrates that an image-based motion estimation algorithm using ANNs which learn the relationship between optical flow and metric pose displacement is comparable that of a commercially available magnetic tracker. The performance of the ANNs is enhanced by the usage of the NBI modality, which produces stronger features and slightly improved motion estimation. The performance of our algorithm demonstrates its feasibility as a feedback mechanism for enabling real-time closed-loop control of teleoperated flexible endoscopes.

REFERENCES

[1] Fact sheet # 297: Cancer. World Health Organization (WHO). Last accessed: 1 February 2012. [Online]. Available: www.who.int/mediacentre/factsheets/fs297/en .

[2] F. Haggar and R. Boushey, "Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors," *Clinics in Colon and Rectal Surgery*, vol. 22, no. 4, pp. 191–197, 2009.

[3] P. Valdastri, M. Simi, and R. J. Webster III, "Advanced technologies for gastrointestinal endoscopy," *Annual Review of Biomedical Engineering*, vol. 14, pp. 397–429, 2012.

[4] M. Hasan and M. Wallace, "Image-enhanced endoscopy," *American Society for Gastrointestinal Endoscopy*, vol. 16, no. 4, pp. 1–4, 2009.

[5] G. Postic, D. Lewin, C. Bickerstaff, and M. Wallace, "Colonoscopic miss rates determined by direct comparison of colonoscopy with colon resection specimens," *The American Journal of Gastroenterology*, vol. 97, no. 12, pp. 3182–3185, 2002.

[6] S. Pasha, J. Leighton, A. Das, M. Harrison, S. Gurudu, F. Ramirez, D. Fleischer, and V. Sharma, "Comparison of the yield and miss rate of narrow band imaging and white light endoscopy in patients undergoing screening or surveillance colonoscopy: A meta-analysis," *The American Journal of Gastroenterology*, vol. 107, no. 3, pp. 363–370, 2011.

[7] Vital Signs Cancer screening, colorectal cancer. Centers for Disease Control and Prevention. Last accessed: 1 February 2012. [Online]. Available: www.cdc.gov/vitalsigns/CancerScreening/indexCC.html .

[8] K. L. Obstein and P. Valdastri, "Advanced endoscopic technologies for colorectal cancer screening." *World J Gastroenterol*, vol. 19, no. 4, pp. 431–9, 2013.

[9] R. Reilink, S. Stramigioli, and S. Misra, "Three-dimensional pose reconstruction of flexible instruments from endoscopic images," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 2076–2082.

[10] ——, "Image-based flexible endoscope steering," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, R. Luo, Ed. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 2010, pp. 2339–2344.

[11] N. van der Stap, R. Reilink, S. Misra, I. Broeders, and R. van der Heijden, “The use of the focus of expansion for automated steering of flexible endoscopes,” in *4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 13–18.

[12] J. Ruiter, E. Rozeboom, M. van der Voort, M. Bonnema, and I. Broeders, “Design and evaluation of robotic steering of a flexible endoscope,” in *4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 761–767.

[13] K. Obstein and P. Valdastri, “Advanced endoscopic technologies for colorectal cancer screening,” *World Journal of Gastroenterology*, vol. 19, no. 4, pp. 431–439, 2013.

[14] J. V. Dam, A. Eickhoff, R. Jakobs, V. Kudis, D. Hartmann, and J. Riemann, “Computer-assisted colonoscopy (the NeoGuide system): Results of the first human clinical trial,” *Gastrointestinal Endoscopy*, vol. 63, pp. 261–266, April 2006.

[15] A. Eickhoff, J. V. Dam, R. Jakobs, V. Kudis, D. Hartmann, U. Damian, U. Weickert, D. Schilling, and J. Reimann, “Computer–assisted colonoscopy (the neoguide endoscopy system): Results of the first human clinical trial (”pace study”),” *American Journal of Gastroenterology*, vol. 102, no. 2, pp. 261–266, 2007.

[16] F. Consentino, E. Tumino, G. Passoni, E. Morandi, and A. Capria, “Functional evaluation of the endotics system, a new disposable self-propelled robotic colonoscope: in vitro tests and clinical trial,” *The International Journal of Artificial Organs*, vol. 32, no. 8, pp. 517–527, 2009.

[17] B. Vucelic, D. Rex, R. Pulanic, J. Pfefer, I. Hrstic, B. Levin, Z. Halpern, and N. Arber, “Aer-o-scope: Proof of concept of a pneumatic, skill-independent, self-propelling, self-navigating colonoscope,” *Gastroenterology*, vol. 130, no. 3, pp. 672–677, 2006.

[18] J. Troccaz, Ed., *Medical Robotics.* ISTE Ltd, London, UK and John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012.

[19] Northern Digital Inc. Last accessed: 1 May 2013. [Online]. Available: www.ndigital.com/medical/aurora.php .

[20] Ascension Technology Corporation. Last accessed: 1 May 2013. [Online]. Available: www.ascension-tech.com/medical .

[21] M. Szura, K. Bucki, A. Matyja, and J. Kulig, “Evaluation of magnetic scope navigation in screening endoscopic examination of colorectal cancer,” *Surgical Endoscopy*, vol. 26, pp. 632–638, 2012.

[22] D. Deguchi, K. Mori, Y. Suenaga, J. Hasegawa, J. Toriwaki, H. Natori, and H. Takabatake, “New calculation method of image similarity for endoscope tracking based on

image registration in endoscope navigation," *International Congress Series*, vol. 1256, no. 0, pp. 460 – 466, 2003.

[23] D. Deguchi, K. Mori, M. Feuerstein, T. Kitasaka, J. C. Maurer, Y. Suenaga, H. Takabatake, M. Mori, and H. Natori, "Selective image similarity measure for bronchoscope tracking based on image registration," *Medical Image Analysis*, vol. 13, no. 4, pp. 621 – 633, 2009.

[24] F. Asano, "Virtual bronchoscopic navigation," *Clinics in Chest Medicine*, vol. 31, no. 1, pp. 75–85, 2010.

[25] K. Mori, D. Deguchi, J. Sugiyama, Y. Suenaga, J. Toriwaki, C. M. Jr., H. Takabatake, and H. Natori, "Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images," *Medical Image Analysis*, vol. 6, no. 3, pp. 321 – 336, 2002.

[26] J. Rey, H. Ogata, N. Hosoe, K. Ohtsuka, N. Ogata, K. Ikeda, H. Aihara, I. Pangtay, T. Hibi, S. Kudo, and H. Tajiri, "Blinded nonrandomized comparative study of gastric examination with a magnetically guided capsule endoscope and standard videoendoscope," *Gastrointestinal Endoscopy*, vol. 75, no. 2, pp. 373–381, 2012.

[27] J. Keller, C. Fibbe, F. Volke, J. Gerber, A. C. Mosse, M. Reimann-Zawadzki, E. Rabinovitz, P. Layer, D. Schmitt, V. Andresen, U. Rosien, and P. Swain, "Inspection of the human stomach using remote-controlled capsule endoscopy:a feasibility study in healthy volunteers (with videos)," *Gastrointestinal Endoscopy*, vol. 73, no. 1, pp. 22–28, 2011.

[28] P. Valdastri, G. Ciuti, A. Verbeni, A. Menciassi, P. Dario, A. Arezzo, and M. Morino, "Magnetic air capsule robotic system: proof of concept of a novel approach for painless colonoscopy," *Surgical Endoscopy*, vol. 26, no. 5, pp. 1238–46, 2011.

[29] H. Keller, A. Juloski, H. Kawano, M. Bechtold, A. Kimura, H. Takizawa, and R. Kuth, "Method for navigation and control of a magnetically guided capsule endoscope in the human stomach," in *Proceedings of the 4th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, J. Desai, L. Phee Soo Jay, and L. Zollo, Eds. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 2012, pp. 859–865.

[30] A. Mahoney and J. Abbott, "Control of untethered magnetically actuated tools with localization uncertainty using a rotating permanent magnet," in *Proceedings of the 4th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, J. Desai, L. Phee Soo Jay, and L. Zollo, Eds. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 2012, pp. 1632–1637.

[31] M. Gao, C. Hu, Z. Chen, H. Zhang, and S. Liu, "Design and fabrication of a magnetic propulsion system for self-propelled capsule endoscope," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2891–2902, 2010.

[32] X. Wang, M. Meng, and X. Chen, “A locomotion mechanism with external magnetic guidance for active capsule endoscope,” in *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 4375–4378.

[33] T. E. Hampshire, H. R. Roth, D. J. Boone, G. Slabaugh, S. Halligan, and D. J. Hawkes, “Prone to supine ct colonography using a landmark and intensity composite method,” in *Abdominal Imaging. Computational and Clinical Applications*, ser. Lecture Notes in Computer Science, vol. 7601, 2012, pp. 1–9.

[34] M. N, F. Nageotte, P. Zanne, and M. D. Mathelin, “In vivo comparison of real-time tracking algorithms for interventional flexible endoscopy,” in *Proceedings of the 6th IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, W. Karl, Ed., 2009, pp. 1350–1353.

[35] N. van der Stap, F. Heijden, and I. Broeders, “Towards automated visual flexible endoscope navigation,” *Surgical Endoscopy*, pp. 1–9, 2013.

[36] G. Khan and D. Gillies, “Vision based navigation system for an endoscope,” *Image and Vision Computing*, vol. 14, no. 10, pp. 763 – 772, 1996.

[37] I. Bricault, G. Ferretti, and P. Cinquin, “Registration of real and CT-derived virtual bronchoscopic images to assist transbronchial biopsy,” *IEEE Transactions on Medical Imaging*, vol. 17, no. 5, pp. 703–714, 1998.

[38] S. Krishnan, C. Tan, and C. Chan, “Closed-boundary extraction of large intestinal lumen,” in *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, 1994, pp. 610–611.

[39] H. Tian, T. Srikanthan, and K. Asari, “Automatic segmentation algorithms for the extraction of lumen region and boundary from endoscopic images,” *Medical and Biological Engineering and Computing*, vol. 39, no. 1, pp. 8–14, 2001.

[40] S. Xia, S. Krishnan, M. Tjoa, and P. Goh, “A novel methodology for extraction colon’s lumen from colonoscopic images,” *Journal of Systemics, Cybernetics and Informatics*, vol. 1, no. 2, pp. 7–12, 2003.

[41] X. Zabulis, A. Argyros, and D. Tsakiris, “Lumen detection for capsule endoscopy,” in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3921–3926.

[42] Z. Zhen, Q. Jinwu, Z. Yanan, and S. Linyong, “An intelligent endoscopic navigation system,” in *Proceedings of the 2006 IEEE International Conference on Mechatronics and Automation*, S. Guo, Ed., 2006, pp. 1653–1657.

[43] G. P. Stein, O. Mano, and A. Shashua, “A robust method for computing vehicle egomotion,” in *IEEE Intelligent Vehicles Symposium (IV2000)*, J. Rillings, Ed. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 2000.

[44] T. Suzuki and T. Kanade, “Measurement of vehicle motion and orientation using optical flow,” in *Proceedings of the 1999 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems.* IEEE Service Center, Piscataway, NJ, USA, 1999, pp. 25–30.

[45] E. Erdemir, D. Wilkes, K. Kawamura, and A. Erdemir, “Learning structural affordances through self-exploration,” in *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, P. Blazevic, Ed. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 2012.

[46] J. Shi and C. Tomasi, “Good features to track,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[47] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *Proceedings of the 9th European conference on Computer Vision*, 2006, pp. 430–443.

[48] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[49] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Computer Vision and Image Processing (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[50] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

[51] G. Puerto-Souza and G. Mariottini, “A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1201–1214, July 2013.

[52] ——, “Wide-baseline dense feature matching for endoscopic images,” in *6th Pacific Rim Symposium in Advances in Image and Video Technology, PSIVT 2013*, November 2013, in press.

[53] A. J. Davison, “Real-time simultaneous localisation and mapping with a single camera,” in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ser. ICCV ’03, B. Werner, Ed. Institute of Electrical and Electronics Engineers (IEEE) Computer Society Press, Washington, DC, USA, 2003, pp. 1403–.

[54] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[55] T. Thormählen, H. Broszio, and P. Meier, “Three-dimensional endoscopy,” in *Falk Symposium No. 124, Medical Imaging in Gastroenterology and Hepatology*, 2002.

[56] J. Liu, T. Yoo, K. Sabramanian, and R. V. Uitert, "A stable optic-flow based method for tracking colonoscopy images," in *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2008, pp. 1–8.

[57] K. Suzuki, "Pixel-based machine learning in medical imaging," *International Journal of Biomedical Imaging*, vol. 2012, pp. 1–18, 2012.

[58] Z. Shi and L. He, "Application of neural networks in medical image processing," in *Proceedings of the Second International Symposium on Networking and Network Security*, F. Yu, G. Yue, M. Leng, and X. Peng, Eds. Academy Publisher, Oulu, Finland, April 2010, pp. 23–26.

[59] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: Modelling, Planning and Control*, 1st ed. Springer-Verlag Limited, London, UK, 2008.

[60] R. Schalkoff, *Artificial Neural Networks.* McGraw-Hill, New York, NY, USA, 1997.

[61] ——, *Pattern Recognition: Statistical, Structural, and Neural Approaches.* John Wiley & Sons, Inc., Hoboken, NJ, USA, 1992.

[62] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 4, pp. 303–314, 1989.

[63] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[64] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly, 2008.

[65] G. A. Puerto-Souza and G.-L. Mariottini, "Hierarchical multi-affine (hma) algorithm for fast and accurate feature matching in minimally-invasive surgical images," in *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[66] M. A. Fischler and R. C. Bollest, "Random sample consensus: A paradigm for model fitting with applications to image analysis," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[67] J. W. Tukey, *Exploratory Data Analysis.* Addison-Wesley, 1977.

[68] J. Bulat, K. Duda, M. Duplaga, R. Fraczek, A. Skalski, M. Socha, P. Turcza, and T. Zielinski, "Data processing tasks in wireless GI endoscopy: Image-based capsule localization & navigation and video compression," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 2815–2818.

[69] H. Korin, R. Ehman, S. Reiderer, J. Felmlee, and R. Grimm, “Respiratory kinematics of the upper abdominal organs: A quantitative study,” *Magnetic Resonance in Medicine*, vol. 23, pp. 172–178, 1992.

[70] L. Sherwood, *Human Physiology: From Cells to Systems*, 6th ed. Thomson Brooks/Cole, Pacific Grove, CA, USA, 2007.

[71] A. Saxena, J. Schulte, and A. Y. Ng, “Depth estimation using monocular and stereo cues,” in *IJCAI'07 Proceedings of the 20th international joint conference on Artificial Intelligence*, M. M. Veloso, Ed. AAAI Press, Menlo Park, CA, USA, 2007.

[72] S. Wilson, B. Lovell, A. Chang, and B. Masters, “Visual odometry for quantitative bronchoscopy using optical flow,” in *WDIC2005 APRS Workshop on Digital Image Computing Proceedings*, B. Lovell and A. Maeder, Eds., vol. 1, no. 1. The University of Queensland, St. Lucia, Queensland, Australia, 2005.

[73] V. Guizilini and F. Ramos, “Semi-parametric learning for visual odometry,” *International Journal of Robotics Research*, vol. 32, no. 5, pp. 526–546, April 2013.

[74] G. Bradski, “The OpenCV Library,” *Dr. Dobb's Journal of Software Tools*, 2000.

[75] K. Obstein, V. Patil, J. Jayender, R. Estépar, I. Spofford, B. Lengyel, K. Vosburgh, and C. Thompson, “Evaluation of colonoscopy technical skill levels by use of an objective kinematic-based system,” *Gastrointestinal Endoscopy*, vol. 73, no. 2, pp. 315–321, 2011.

[76] R. Hecht-Nielsen, “Kolmogorov's mapping neural network existence theorem,” in *Proceedings of the IEEE First Annual International Conference on Neural Networks*, S. Grossberg, Ed., vol. 3, 1987, pp. 11–14.

[77] A. Rencher and W. Christensen, *Methods of Multivariate Analysis*, 3rd ed. John Wiley & Sons, Inc., 2012.

[78] R. Collins, Y. Liu, and M. Leordeanu, “Online selection of discriminative tracking features,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1631–1643, 2005.